Acceleration of Protein Simulations on the Earth Simulator by Vectorization and Parallelization

Group Representative
Minoru Saito Hirosaki University

Author
Minoru Saito Hirosaki University

The purpose of subgroup 2 is to computationally demonstrate and 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. The purpose of our first stage (in 2003) was the acceleration of COSMOS90 on the Earth Simulator by vectorization and parallelization. The performance speed was measured for a biologically important protein under the same conditions as biological studies. We successfully obtained the same performance speed (0.005 sec/step for 16034 atoms using 8 nodes) as expected in April in 2003. In this measurement, long-range Coulomb interactions were explicitly calculated without truncating. This performance speed is faster than other typical software widely used in the world, CHARMM, AMBER7, and NAMD2.4.

Keywords: 1. Molecular dynamics, 2. Protein, 3. Benchmark, 4. Hemoglobin, 5. Structural change.

1. Introduction

Life science has entered a new era by completing the exploration of all human genes. In this era, nations and companies are competing with each other to extract useful information about human health from genes. Genes have genetic codes for all proteins with own functions to initiate and maintain life. Various modern technologies (from computers to accelerators) are utilized in the life science to study genes and proteins. The accelerators become a powerful tool to explore 3-dimensional structures of proteins. High-speed computers become a necessary tool to simulate proteins and reveal their dynamical features.

Computer simulations of proteins require enormous computation time, because of the following difficulties. Proteins are large molecules consisting of thousands of atoms and have complicated structures. Furthermore, they largely fluctuate and easily change the whole structure even at the room temperature. Proteins maintain their structures through delicate balances among various interactions between atoms, such as, bond, angle, torsion, Lennard-Jones, Coulomb, and hydrogen bonds with water molecules surrounding proteins. Easy approximations for the interactions or surrounding waters cause an unbalance between interactions and make simulation results unreliable.

Our purpose is to demonstrate large conformational changes of proteins by performing realistic simulations without simplifications. To perform the realistic simulations of proteins, we must include all atoms of proteins in water and their all interactions from chemical bonds to long-range Coulomb interactions. There are two ways to speed up realistic simulations; one is to use a faster hardware such as a super computer and another is to use a software adopting faster algorithm. The fastest simulation becomes possible by installing such software on a super computer. We have installed the software COSMOS90 on the Earth Simulator to demonstrate large conformational changes of proteins by performing long-time simulations.

COSMOS90 was developed by the author (M.Saito) in 1990 and made it possible to simulate a protein in water with all degrees of freedom and with long-range Coulomb interactions using the Particle-Particle and Particle-Cell (PPPC) method(1). The PPPC method was proposed also by the author to efficiently calculate long-range Coulomb interactions between atomic charges in the order *NlogN* instead of N^2 by dividing a system into hierarchical cubic cells based on the Barnes & Hut tree code (Fig. 1). COSMOS90 was developed on a vector super computer (VP400) in 1990 and several years ago installed on vector-parallel super computers (VPP500 and VPP5000) using VPP Fortran and recently on the Earth Simulator using MPI (March in 2003). In 2003, the author has tuned up COSMOS90 on the Earth Simulator by vectorizing and parallelizing it.

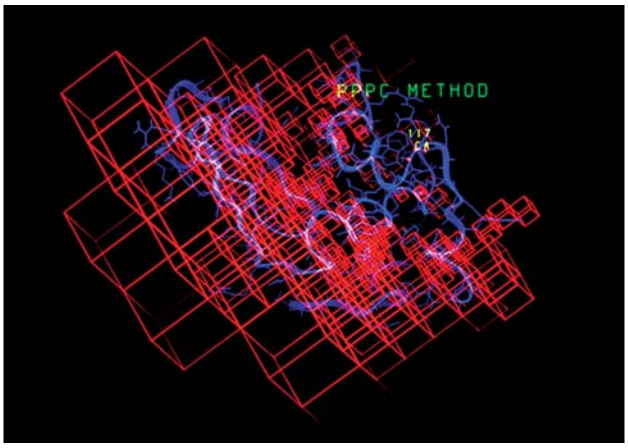

Fig. 1 Particle-Particle and Particle-Cell (PPPC) method. Long-range Coulomb interactions are calculated in the order *NlogN* using the PPPC method. This figure shows a space subdivision based on PPPC to calculate Coulomb forces acting to the 117th α-carbon atom of human lysozyme (130 amino acids).

2. Parallelizaion of COSMOS90

A computation flow in COSMOS90 was shown in Fig. 2. COSMOS90 has a large loop for MD time step, as like the other software widely used in the world (CHARMM, AMBER, GROMOS, and DISCOVER). In the MD time step loop, bonded (bond, angle, and torsion) and nonbonded (Lennard-Jornes and Coulomb) forces acting to atoms are computed. The nonbonded forces are efficiently computed in COSMOS90 by preparing an interaction table based on the PPPC method. After the force calculations, new positions of atoms are computed by solving the equations of motion. The vectorization tuning was performed for all subprocesses in the time step loop. The parallelization tuning was performed for the all subprocesses except for the Barnes & Hut tree code mainly by distributing array data cyclically on all processors. The calculations of bonded forces were parallelized by dividing array data to blocks. The atomic positions updated by the equations of motion were sent to all processors. The above strategy for parallelization is common for the VPP5000 and the Earth Simulator, as follows. The loop division by the VPP Fortran [!XOCL SPREAD DO] was replaced by [DO I=myrank+1,N_{atom},nnode]. For the communications between all processors, [!XOCL UNIFY, !XOCL MOVEWAIT] of the VPP Fortran were replaced by the MPI functions, allreduce and allgather. To avoid the communication latency and utilized the large band width of the Earth Simulator, we packed several array data into a large work array and sent to all processors by MPI_allgather. This improvement for the communication sped up COSMOS90 on the Earth Simulator, although it produced new overheads, copies of array data. For the parallelization based on MPI, processors inside a node were treated with the same manner as those between nodes (flat programming).

3. Performance speed of COSMOS90

The performance speed was measured for a DNA-protein complex in water (Fig. 3). This system consists of 16034 atoms. Recently, the author performed MD simulations for this system and calculated relative binding free energy between DNA and the protein (λ-repressor). The binding free energy change calculated for Thymine→Uracil substitution in DNA (–1.5±0.4 kcal/mol) was in good agreement with an experimental value (–1.8 kcal/mol)(2). To measure the performance speed, we adopted the same approximation parameters and calculation conditions as these simulations and avoid the special conditions increasing vectorization and parallelization ratios. Therefore, the present study guarantees us to carry out productive simulations on the Earth Simulator with the same performance speed as in this study. The wall-clock time consumed to compute a step of MD was meas-

Loop for MD step

(1) Calculating bonded forces (bond, angle, torsion)

(2) Calculating nonbonded forces based on PPPC

(a) Subdividing a system into hierarchical cubic cells

(b) Making an interaction table for P-P and P-C

(c) Calculating Coulomb and L-J forces from P and C

(3) Updating positions according to eq. of motion

(4) Communicating new positions

Next step

Fig. 2 Computational flow in COSMOS90. COSMOS90 efficiently calculates long-range Coulomb interactions by preparing a nonbonded interaction table based on the PPPC method.

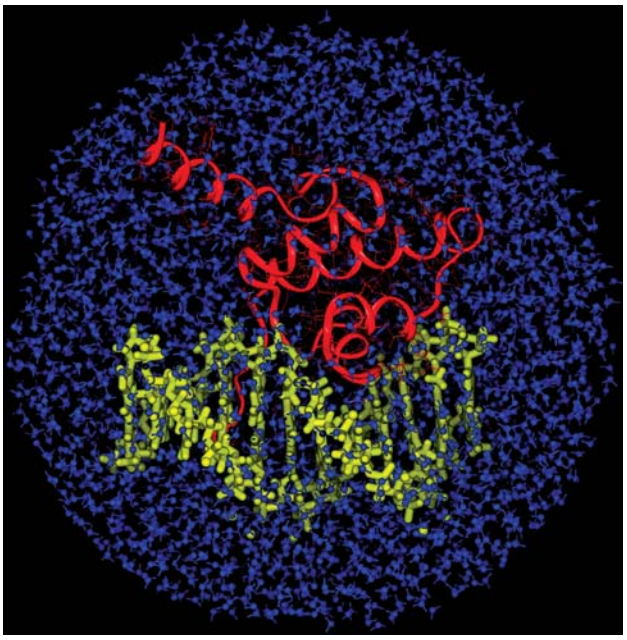

Fig. 3 Protein-DNA complex in water (16034 atoms). Radius of water sphere is 34 angstrom.

ured by the clock function of VPP Fortran for VPP and by MPI_Wtime of the MPI libraries for the Earth Simulator. The performance speed was measured for two load modules obtained by the vector and scalar compiles of a source code to clarify the acceleration ratio by the vectorization.

The performance speed of COSMOS90 on the Earth Simulator was shown in Table 1, together with the speed on VPP500 and VPP5000[(3)]. The performance speed of a single processor was 0.185 sec/step for the vector module and 1.607 sec/step for the scalar module. Thus, the vectorization accelerated the performance speed for a single processor by 8.7 times faster than the scalar performance. This result means that COSMOS90 adequately brought out the vector ability of the single processor. The performance speed on a single node (consisting of 8 processors) was 0.026 sec/step which was almost the same as the speed 0.029 sec/step estimated from the difference between VPP (9.6 Gflops for a processor) and the Earth Simulator (64 Gflops for a node). The performance speed on 8 nodes (64 processors) was 0.005 sec/step, which was the same as that estimated from the difference between VPP and the Earth Simulator and was our objective speed at the last year (in 2003). To compare our performance speed with those of the other software widely used in the world, the performance speeds of CHARMM, AMBER7, and NAMD2.4 were cited in Table 2, where the speeds were measured on an alpha cluster (Lemieux at Pittsburgh Supercomputing Center)[(4)]. The maximum speed is 0.023 sec/step marked by NAMD2.4 on 128 processors. Their measurements were performed for 1.5 times larger system than the present study. We should scale down the values in Table 2. by 30 % to quantitatively compare their results with our results. Then, our speed on the Earth Simulator (0.005 sec/step) is about 3 times faster than the

Table 1 Performance speed of COSMOS90. Time for computing a step of MD simulation for a DNA-protein complex in water (Fig. 3).

Machine No. of CPU	Scalar	Vector
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	1.607	0.185
1 node	0.208	0.026
2 nodes	0.107	0.014
4 nodes	0.057	0.008
8 nodes	0.031	0.005

Table 2 Performance speed of other MD software for protein simulations (sec/step). The speed was measured for a DHFR protein in water (23558 atoms).

processors	CHARMM c29b1	AMBER7	NAMD2.4
1	1.332	1.020	1.385
2	0.685	0.460	0.750
4	0.356	0.250	0.390
8	0.217	0.150	0.198
16	0.142	0.100	0.105
32	0.108	0.070	0.061
64	0.098	0.060	0.040
128	0.104	-	0.023

maximum speed of NAMD2.4, even though the system size difference is corrected.

To clarify a possibility of the acceleration by more number of processors, we measured the wall-clock time for each subprocess in the MD time step loop (Fig. 2). The each subprocess in Fig. 2 consumed the following percentage of time per step. (1) 3% for computing nonbonded forces, (2)a and (2)b 12% for subdividing a space and making a interaction table for nonbonded forces, (2)c 55% for computing nonbonded forces, (3) 7% for solving the equations of motion, (4) 17% for communications, (5) 6% for others. The most time consuming subprocess is (2)c, the computations of nonbonded forces. The accelerations of this subprocess is 7.3 times by the vectorization and 55.5 times by the parallelization on 64 processors. Therefore, more number of processors may shorten computation time for this subprocess and thus whole time per step. The subprocess that we have not parallelized yet is (2)a dividing a space into hierarchical cubic cells based on the Barnes & Hut tree code. We are parallelizing this subprocess, although this subprocess consumes a small percentage of time per step.

4. Purpose at the next state

The purpose of our project at the next stage is to computationally demonstrate large structural changes of proteins on the Earth Simulator using COSMOS90 tuned up in this study. As a target protein, we chose a hemoglobin molecule (Fig. 4). A hemoglobin molecule can efficiently transfer oxygen molecules from the lungs to the muscles. The binding of an oxygen molecule enhances additional oxygen bindings on 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 computationally demonstrate and visualize such structural changes of proteins using the Earth Simulator and software COSMOS90.

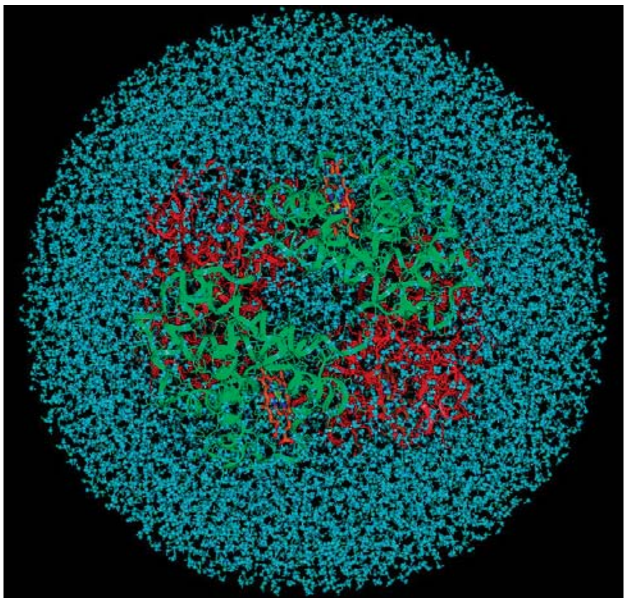

Fig. 4 Hemoglobin in water. This system contains 52416 atoms and consists of 754 amino acids and 14450 water molecules.

References

(1) M. Saito: Molecular dynamics simulations of proteins in water without the truncation of long-range Coulomb interactions, *Molecular Simulation*, vol.8, pp.321–333 (1992).

(2) M. Saito and A. Sarai: Free energy calculations for the relative binding affinity between DNA and λ-repressor, *Proteins*, vol.52, pp.129–136 (2003).

(3) M. Saito and K. Sayano: Acceleration of protein simulations by vectorization and parallelization on the Earth Simulator, *High Performance Computing System 2004*, pp.81–86 (2004).

(4) http://www.scripps.edu/brooks/Benchmarks/
The performance speeds were measured on Lemieux (Hewlett-Packard alpha-server SC ES45/667MHz at Pittsburgh Supercomputing Center) for a protein, dihydrofolate reductase (DHFR), in a cubic box (62.23 Å). This system contains 23,558 atoms, 159 amino acids, and 7023 water molecules. Long-range Coulomb interactions were calculated by PME method.