All-electron Calculation on Very Large-Sized Proteins by Density Functional Method

Group Representative
Fumitoshi Sato Institute of Industrial Science, University of Tokyo

Authors
Fumitoshi Sato *[1] ・ Tamotsu Yoshihiro *[1] ・ Tetsuya Ueno *[1]
*1 Institute of Industrial Science, University of Tokyo

In order to understand the electronic properties of proteins, we are developing a gaussian-based density functional method program for proteins, ProteinDF. It can treat a whole protein as a molecule and calculated more than 100 residues metallo-protein which contains about 10,000 canonical orbitals (100 million elements) by workstation cluster. The purpose of this project is to attain all-electron calculation on 1,000 residues complex protein which has 100,000 orbitals (10 billion elements) on the Earth Simulator (ES) by the optimal codes of ProteinDF.

In this year, the original MPMD program of ProteinDF was successfully reconstructed to achieve MPMD without generating plural dynamic processes for MPI-2/ES. We carried out several all-electron calculations up to 100 residues proteins on ES and especially concentrated on parallel tuning. At present, the vectorization ratio of 92% in one residue and the parallelization ratio of 96% in 31 residues protein have been reached with 8 nodes.

In the next year, we are going to develop and tune ProteinDF further, and perform an all-electron calculation on a large-sized protein of 30,000 canonical orbitals (1 billion elements).

Keywords: Protein, All-Electron Calculation, Density Functional Method, Canonical Orbital, Quantum Chemistry

1. Introduction

To elucidate the electronic properties of proteins, we are developing a gaussian-based density functional (DF) molecular orbital (MO) program, ProteinDF. Using ProteinDF, we performed the first all-electron calculation on cytochrome *c*, which contains 104 residues and a *c*-type heme. The number of atoms, electrons, orbitals, and auxiliary functions are 1,738, 6,586, 9,600, and 17,578. To our knowledge, this is the largest system to be calculated by the DF method.

We show the graphics of the 3293-th highest occupied MO (HOMO) with cytochrome *c* structure in Figures 1(a) and (b), where the isosurface values are (a) ±0.05 and (b) ±0.00005, respectively. The main components of HOMO are 3d orbitals of heme Fe, and HOMO is delocalized over the whole molecule. Here, we point out the following fact. At first, the MOs are calculated to solve the Kohn-Sham-Roothaan equation. This is the nonlinear matrix equation of which dimension is the number of orbitals. Spreading MO seen in Figure 1(b) shows that almost all the elements of the coefficient matrix expressing MO are not 0. In fact, matrices are not sparse. Second, cytochrome *c* functions in a protein matrix as a reversible donor/acceptor of an electron, and this MO is considered as the carrier bag in electron and hole transfer. Figure 1 suggests that the direct coupling between cytochrome *c* HOMO and acceptor MO cannot be ignored. As a conclusion, the all-electron canonical wavefunctions should be calculated to reveal the characters of proteins; that is, any approximation can not be introduced for solving the equation. The purpose of this project is to achieve an all-electron calculation on 1,000 residues complex protein which functions by delivering and receiving an electron among several active centers. The canonical orbital calculations are indispensable to treat the accurate overlap between the MOs of active centers. There is only ES in performing such calculation.

In this report, we explain our ProteinDF program and show the progress in this year.

2. Overview of ProteinDF

ProteinDF is the program to solve the Kohn-Sham-Roothaan equation based on the gaussian-type orbitals by the resolution of identity method. It is coded by the object-oriented language C++ to relax the complexity of large software systems. We particularly notice to keep the independence among programming units. The self-consistent field (SCF) structure of ProteinDF is illustrated in Figure 2(a). To support the various kinds of computations, ProteinDF is successfully divided into two types of objects, i.e. scenario objects (85,000 statements) and computational objects (65,000 statements). The latter consists of four time-consuming routines; molecular integrals, exchange correlation (XC) 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. We only tune these routines. Because these routines are called repeatedly, original ProteinDF is parallelized by MPMD method using the hierarchical object structure (Figure 2(b)).

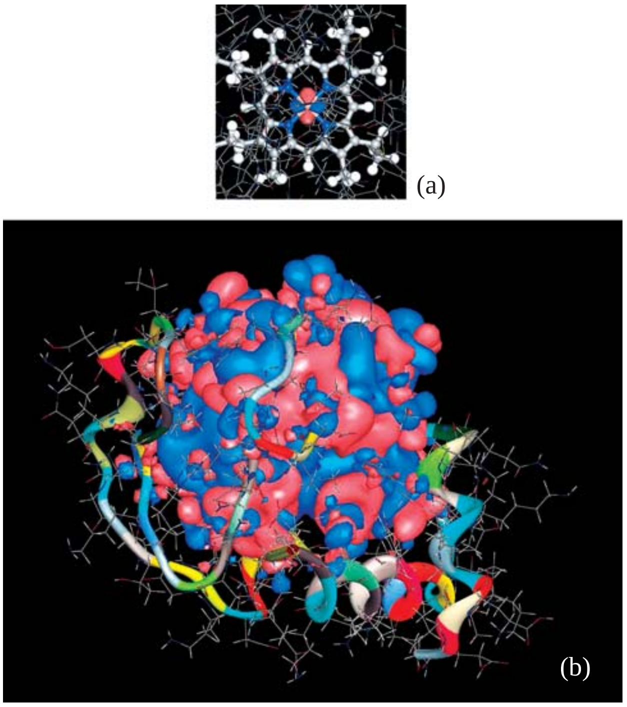

Fig. 1 HOMO of cytochrome *c*, where the isosurface values are (a) ±0.05 and (b) ±0.00005.

3. Results

3.1. Program Conversion

Original ProteinDF adopted the PVM as a communication library to achieve MPMD. Since ProteinDF is parallelized by using the hierarchical object structure in Figure 2(b), it was easily transformed to MPI-2 library on ES. However, MPI-2/ES is implemented to the inconvenient way for calling dynamic processes repeatedly (spawn function).

In this year, we added the conversion from the original MPMD program of ProteinDF to the new MPMD without generating two or more dynamic processes for MPI-2/ES. It was successfully reconstructed to introduce the job controller object as shown in Figure 3.

3.2. Test Calculations

There is a simple relation between the number of residues and orbitals. If the Double-Zeta type basis set is adopted (this is the default of ProteinDF), the number of orbitals is estimated to be 100 times larger than that of residues. Since this is the dimension of matrix, the number of computational elements is proportional to the square.

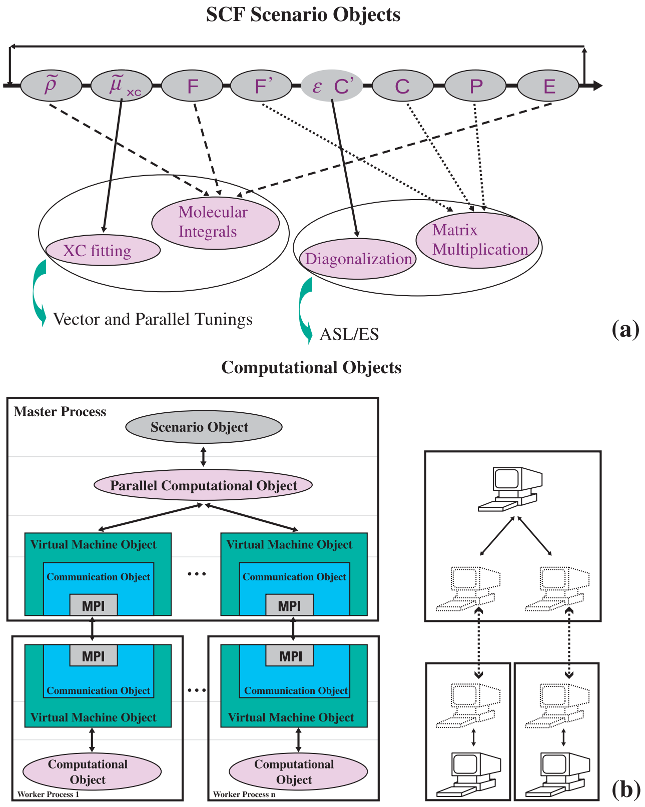

Fig. 2 Structure of ProteinDF; (a) SCF scenario and computational objects, (b) hierarchical object structure for parallelization.

Toward the final goal, we plan the test calculations of 1, 3, 11, 31, 104 (plus one heme; cytochrome *c*), 306, and about 1,000 residues proteins. According to this plan, the system size becomes linearly large step by step (from 10 thousands to 10 billion elements). In this year, we carried out all-electron calculations up to cytochrome *c* on ES.

3.3. Tunings

ProteinDF has four time-consuming routines; molecular integrals, XC fitting, diagonalization and matrix multiplication in SCF. As shown in Figure 2(a), diagonalization and matrix multiplication were completely transposed to those in ASL/ES library. In addition, matrix inversion of which dimension is the number of auxiliary functions is carried out before SCF. This calculation was also substituted. Since matrices in ProteinDF are not sparse, we adopted dense matrix routines.

On the other hand, vector and parallel tunings were performed for molecular integrals and XC fitting routines. Owing to the size limitation for profiling, the vector and parallel tunings were performed by using 1 and 31 residues protein, respectively. In this year, we concentrated on parallel tuning. At present, the vectorization ratio was about 92% (Table 1), and the parallelization efficiency and ratio was 71% and 96% with 8 nodes (Table 2).

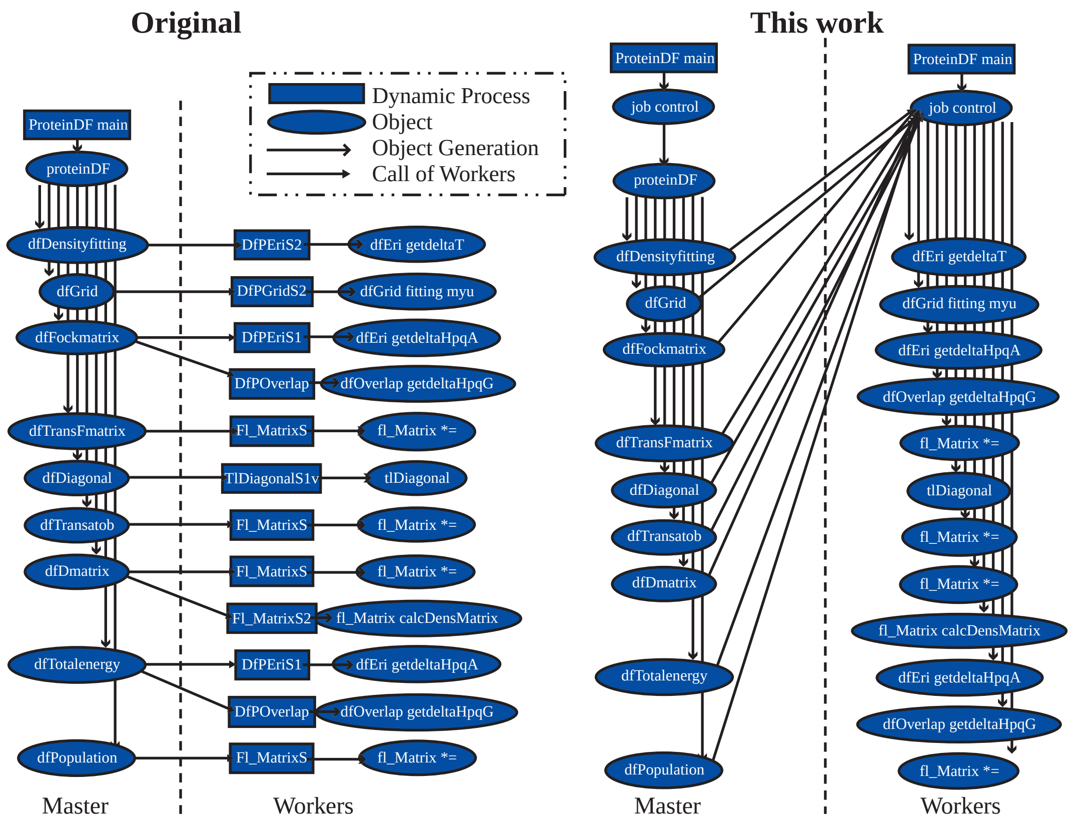

Fig. 3 Relation of Dynamic Process Generation and an Object Generation by the original and new MPMD in ProteinDF.

Table 1 The vectorization ratio of molecular integrals in one residue calculation (a part is extracted).

Computational Routine	Vectorization ratio/%	Performance/MFLOPS
DfEri::auxSet	98.60	2613.3
DfEri::ericalc	91.48	104.6
DfEri::fmtRecursiv	99.17	2978.9
DfOverlap::ovpqgcalc	92.05	109.7
DfDensityfitting::calcnSRou	96.60	606.1
DfDensityfitting::calcRamda	95.40	146.6
DfDensityfitting::calcRouAlpha	96.87	646.3
DfDensityfitting::calctplusRn	96.96	396.8

Table 2 Parallelization efficiency of the molecular integrals and exchange correlation calculation in 31 residues protein calculation.

number of CPUs	serial	2	4	8	32
XC fitting/sec	5776	2967	1512	775	235
efficiency		0.97	0.96	0.93	0.77
Molecular Integrals	21935	11839	6450	3607	964
(DfEri+DfOverlap)		0.93	0.85	0.76	0.71
Molecular Integrals	13607	5699	3007	1722	442
(DfDensityfitting)		1.19	1.13	0.99	0.96

4. Work for the FY2004

In the very large-sized protein, the computational time is proportional to the cube of the number of orbitals, while the communication time is proportional to the square. Then, the larger the size is, the higher the parallelization efficiency becomes. It was thought that the parallel tuning was almost done in this year. However, intra-node library is not prepared for dense matrix diagonalization routine to give all eigenvalues and eigenvectors in ES now. Development of this library is eagerly wanted to progress till next year. In the next year, we will achieve further vector tuning and calculate the all-electron canonical wavefunction of 306 residues protein.

5. Acknowledgement

The authors acknowledge to the Earth Simulator Center for providing computer resources. This work was done in "Frontier Simulation Software for Industrial Science (FSIS)" project supported by IT program of MEXT.

References

1) H. Kashiwagi, H. Iwai, K. Tokieda, M. Era, T. Sumita, T. Yoshihiro, and F. Sato, "Convergence process with quasi-canonical localized orbital in all-electron SCF calculation on proteins", Mol. Phys. vol.101, no.1–2, pp. 81–86, January 2003.