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

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

Authors
Fumitoshi Sato*1, Tamotsu Yoshihiro*1 and 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, called ProteinDF which can treat a whole protein as a molecule and calculate more than 100 residues which contains about 10,000 canonical orbitals (100 million elements) by workstation cluster. ProteinDF is coded by C++. 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, in order to perform an all-electron calculation on a large-sized protein of 30,000 canonical orbitals (1 billion elements) by ES, we mainly carried out the vectorization tune, but could not satisfy the condition of limited release. At present, the maximum vectorization ratio of rate-limiting routines is 92% and the parallelization ratio in 31 residues protein is 92% with 6 nodes. During this process, it turned out that it is difficult to make the effective vectorization tune of ProteinDF by ES any further. The project of our group will be ended with current year.

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 was the largest system to be calculated by the DF method.

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. MOs of proteins are delocalized over the whole molecule, which indicates that almost all the elements of the coefficient matrix expressing MO are not 0. In fact, matrices are not sparse. 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 and Results of the Last Year

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 1(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 (70,000 statements). The latter consists of four time-consuming routines; molecular integrals, exchange correlation (XC) fitting, diagonalization and the matrix multiplication. Their tasks are depending on the 2.3^{rd}, 1.8^{th}, 3.3^{rd} and 2.9^{th} 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 1(b)). However, MPI-2/ES is implemented to the inconvenient way for calling dynamic processes repeatedly (spawn function). In the last 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.

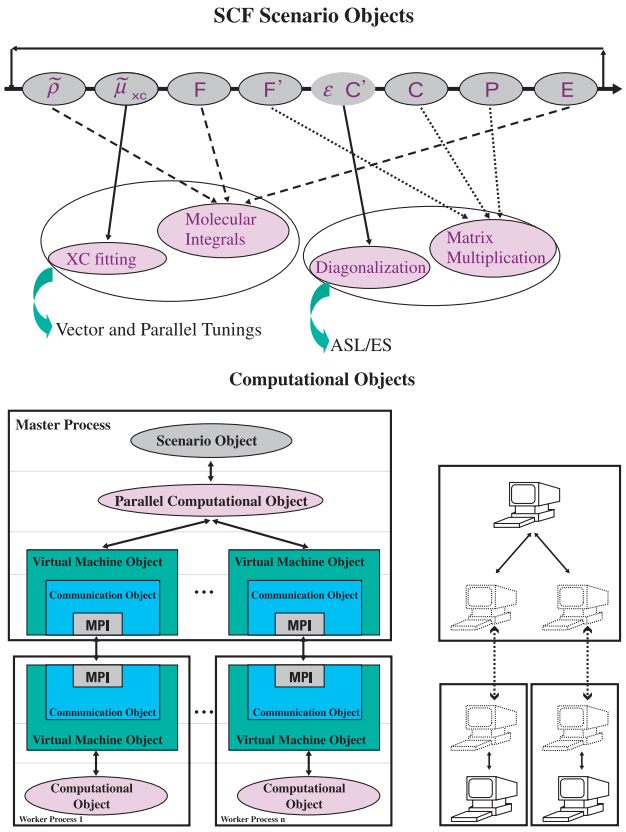

Fig. 1 Structure of ProteinDF; (a) SCF scenario and computational objects, (b) hierarchical object structure for parallelization.

In the four time-consuming routines of ProteinDF, diagonalization and matrix multiplication were completely transposed to those in ASL/ES library.

3. Results

In this year, in order to achieve an all-electron calculation of the large-scale protein in 30,000 orbitals (one billion elements), 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 3 and 31 residues proteins, respectively.

The object programs are DfGrid, DfEri, DfOverlap. DfGrid is the exchange-correlation (XC) fitting program, and DfEri and DfOverlap are the molecular integral programs. These contain 20,000 statements for serial and 50,000 statements for parallel control. The executed tuning work was as follows;

A. The compulsion vector statement lines were inserted in all possible loops.
B. We limited the computational methods and the “if” statements are excluded as much as possible in the rate-limiting calculation routines.
C. According to the simplicity of B, the caused tedious calculations were excluded.
D. The array discontinuousness accesses were corrected as much as possible.
E. The loops were exchanged so that the loop length might become long.
F. The loop unrollings and loop unitings were carried out.
G. In-line codes were generated in all routines with high use frequency by hand.

ProteinDF adopts the Resolution of Identity method, and then the molecular integrals are three-centered calculation. Especially in G, in-line codes of molecular integrals were generated in SSS, SSP, SSD, PSS, PSP, PSD, PPS, PPP, and PPD type.

As the results, the best vectorization ratio of the rate-limiting routines was 92%, while the parallelization ratio of the total self-consistent field SCF calculation on 31 residues protein by six nodes was 92% in present (Table 1). Then, it was not able to meet the computer resource limitation release requirement.

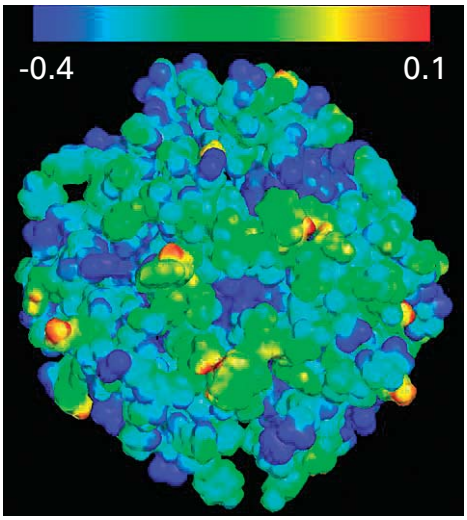

Fig. 2 The electrostatic potential of insulin hexamer (306 residues) calculated by ProteinDF

4. Conclusion

In order to perform all-electron calculation on 1,000 residues protein which has about 100,000 orbitals (10 billion elements) on ES we tried to make the optimal codes of ProteinDF. In this year, we had a plan to achieve further vector tuning and calculate the all-electron canonical wavefunction of 306 residues protein. In fact, recently it was proven that ProteinDF has an ability to attain the calculation (Fig. 2) by using another computer (Altix3700). Then, if the tuning on ES would be successful, the all-electron calculation on 1,000 residues protein has been able to be achieved by ES in the next year.

However, during the tuning process, it turned out that it is difficult to improve the efficiency of vectorization ratio of ProteinDF by ES any further. It is very regrettable that the project of our group will be ended with current year.

Table 1 Parallelization ratio of the molecular integrals and exchange correlation calculation in 31 residues protein calculation. Here, the base time was estimated to that of two-node calculation. The listed values were the average of three trials.

Number of CPUs	15	23	31	39	47
overall (REAL-TIME)	20989.38	18000.03	16392.37	15881.13	15097.50
Efficiency	100	91.23	91.71	90.75	91.32
1SCF (elapse time)	818.00	698.50	630.00	613.00	581.50
Efficiency	100	91.57	92.33	91.15	91.72
XC fitting	42.67	33.33	28.67	26.00	24.00
Efficiency	100	96.22	96.32	96.31	96.43
Main Molecular Integrals	416.67	342.33	308.67	294.33	275.33
Efficiency	100	94.05	93.80	93.19	93.71
Molecular Integrals for Density Fitting	202.67	174.33	150.67	150.33	139.67
Efficiency	100	90.98	93.68	91.56	92.65

5. Acknowledgement

The authors acknowledge to the Earth Simulator Center for providing computer resources for 2 and quarter years. We are also grateful to NEC Informatec Systems, Ltd for tuning of ProteinDF. This work was done in "Frontier Simulation Software for Industrial Science (FSIS)" project supported by IT program of MEXT.

Bibliographies

1) T. Inaba, S. Tahara, N. Nisikawa, H. Kashiwagi, and F. Sato, “All-Electron Density Functional Calculation on Insulin with Quasi-Cannonical Localized Orbitals”, J. Comp. Chem. (2005) in press.
2) T. Inaba, H. Kashiwagi, and F. Sato, “An All-Electron Calculation on Insulin Hexamer by the Density Functional Program ProteinDF”, Proc. of WATOC05, ES-P30, Cape Town, South Africa, Jan. 2005.
3) T. Inaba, and F. Sato, “All-electron calculation on insulin hexamer by the density functional method”, Proc. of SC2004, P16, Pittsburgh, USA, Nov. 2004.
4) N. Ihara, T. Inaba, N. Tsunekawa, F. Sato, and H. Kashiwagi, “Density Functional Calculations on Proteins with ProteinDF Program”, Proc. of the 7th Asian Workshop on First-Principles Electronic Structure Calculations, pp.6, Taipei, Taiwan, Nov. 2004.
5) T. Ueno, T. Inaba, and F. Sato, “The large scale calculation of all-electron on proteins by ProteinDF and Quasi-canonical localized orbital”, Proc. of the 2nd inter. COE Symp. on Large-Scale Computing Methods for Materials Chemistry and Bioscience, pp.10, Sendai, Japan, Nov. 2004.

密度汎関数法による超大型タンパク質の全電子計算

プロジェクト責任者
佐藤　文俊　　東京大学　生産技術研究所

著者
佐藤　文俊*1，吉廣　　保*1，上野　哲哉*1
*1　東京大学　生産技術研究所

タンパク質の電子状態を研究するため、当サブグループはガウス型関数を用いたタンパク質のための密度汎関数法プログラムProteinDFを開発している。ProteinDFはタンパク質全体を分子として扱い、100残基10,000軌道（1億要素）以上のカノニカル軌道計算をワークステーションクラスタで実行することができる。ProteinDFはC++でコーディングされている。本プロジェクトの目的は、ProteinDFを地球シミュレータ（ES）上でチューニングし、1,000残基100,000軌道（100億要素）の全電子計算を達成することである。

本年度は、30,000軌道（10億要素）の大規模タンパク質の全電子計算を達成するため、主にベクトルチューニングに力を注いだが、計算機資源限定解除条件を満たすことができなかった。具体的には計算律速ルーチン（スカラ約20,000行、並列約50,000行）において、可能なループ全てに強制ベクトル化行を挿入、計算機能を限定してif文を可能な限り排除、それに伴う冗長計算部分の排除、可能な限り配列連続アクセスに修正、ループ長が長くなるようループの入れ替え、ループ展開とループ融合、使用頻度の高いルーチンを全て手動でインライン展開、などを実施した。これらの結果により3倍程度の高速化されたが、ベクトル化率はそれほど上昇せず、現行では、計算律速ルーチンの最良ベクトル化率は92%であった。また、31残基タンパク質の6ノードでのSCFの並列化率は92%であった。

この過程で、ES上ではProteinDFのベクトル化率をこれ以上良くすることが困難であることが判明したため、当サブグループは当年を以ってプロジェクトを終了することになった。

キーワード：タンパク質，全電子計算，密度汎関数法，カノニカル軌道，量子化学