Cholesky decomposition method for canonical molecular orbital calculations of proteins

Toshiyuki Hirano and Fumitoshi Sato

Institute of Industrial Science, the University of Tokyo, Japan

1 Introduction

The canonical molecular orbital calculations of proteins provide significant theoretical views, which are difficult to observe experimentally, and are promising technology in various fields such as drug discovery and nanoelectronics devices [1]. Even with modern high-performance computers and parallel technologies in quantum chemistry, however, electronic structure calculations of proteins still remain a time-consuming and laborious task. We aimed at the development of computational methods to efficiently calculate electronic structure of proteins in distributed parallel computing, which becomes mainstream now. In particular, efficiency of two-electron repulsion integral (ERI) computations based on Cholesky decomposition (CD) method is focused in this study.

The rest of paper is constructed as follows. In the section 2, we discuss the theoretical foundation based on the CD method and the use of the pivoted Cholesky decomposition to calculate low-rank approximations of matrices. The section 3 is devoted to numerical results using ProteinDF on distributed memory parallel computer. Finally, we discuss capabilities related to the analyses of electronic structure calculations of proteins in the section 4.

2 Methodology

2.1 Cholesky decomposition method

In self-consistent field (SCF) calculations, the construction of the Fock matrix is one of the most time-consuming steps. In particular, much computer time is spent on the task of computing the Coulomb part $\boldsymbol{J}$ and the Fock exchange part $\boldsymbol{K}$, which include ERI calculations. The matrix $\boldsymbol{J}$ and $\boldsymbol{K}$ are formally computed as follows:

$$J_{pq} = \sum_{rs} P_{rs}(pq|rs) \tag{1}$$

$$K_{pq} = \sum_{qs} P_{qs}(pq|rs) \tag{2}$$

Here, $\boldsymbol{P}$ denotes the density matrix, and the ERI $(pq|rs)$ defined in Eq. (3).

$$(pq|rs) = \iint g_p^*(\boldsymbol{r}) g_q(\boldsymbol{r}) \frac{1}{|\boldsymbol{r}-\boldsymbol{r}'|} g_r^*(\boldsymbol{r}') g_s(\boldsymbol{r}') \mathrm{d}\boldsymbol{r}\mathrm{d}\boldsymbol{r}' \tag{3}$$

The super matrix $V_{ij,kl}$, which is defined by $V_{ij,kl} = (pq|rs)$, is symmetric and positive semi-definite, and can be decomposed by means of the Cholesky procedure into a product of a lower triangular matrix $\boldsymbol{L}$ as follows:

$$V_{pq,rs} = (pq|rs) \approx \sum_{K=1}^{M} L_{K,pq} L_{K,rs} \tag{4}$$

The accuracy of the Cholesky representation of the ERI, Eq. (4), is measured by the residual. Consequently, the accuracy can be controlled by decomposition threshold, δ.

$$\left| V_{pq,rs} - \sum_{K=1}^{M} L_{K,pq} L_{K,rs} \right| \leq \delta \tag{5}$$

According to some computational results, the Cholesky decomposition (CD) method gives reliable, and balanced results even with rather high decomposition thresholds ($\delta = 1.0\times10^{-4}$) [2].

The number of dimensions of the original super matrix V formally becomes $N(N+1)/2$ with N basis functions, so such huge memory for the matrix may not be assigned in each computer node. Therefore, the size of the matrix is reduced in advance of the CDAM method [3]. The diagonal elements of the $\boldsymbol{V}$ matrix, $V_{pq,pq} = (pq|pq)$, are preliminarily estimated and screened by threshold τ.

$$(pq|pq) \approx (I'|I') \geq \tau \tag{6}$$

Here, the AO products, $I = pq$, can be regarded as candidates of the Cholesky basis. Note that the screening technique of the Eq. (6) is associated with the Cauchy-Schwarz screening, which is very useful for providing rigorous ERI bound. Thus, the number of non-negligible basis function pairs will scale as $\mathcal{O}(N)$, thanks to locality of basis functions.

Once the Cholesky vectors $\boldsymbol{L}$ are obtained, the Coulomb part $\boldsymbol{J}$ and exchange part $\boldsymbol{K}$ of the Fock matrix are approximately gained as:

$$J_{pq} = \sum_{rs}(pq,rs)P_{rs} \approx \sum_{rs}\sum_{I} L_{I,pq} L_{I,rs} P_{rs} \tag{7}$$

$$K_{pq} = \frac{1}{2}\sum_{rs}(pr,qs)P_{rs} \approx \sum_{i}\sum_{I} X_{I,pi} X_{I,qi} \tag{8}$$

$$X_{I,pi} = \sum_{r} L_{I,pr} Q_{ri} \tag{9}$$

$$P_{pq} = \textstyle\sum_J Q_{p,J} Q_{q,J} \tag{10}$$

In this study, the matrix $\boldsymbol{X}$ is generated by $\boldsymbol{Q}$, which is given by the CD of the density matrix $\boldsymbol{P}$, instead of MO coefficient matrix. Then, we can use the density matrix as initial guess for the SCF calculation.

Once $\boldsymbol{L}$ is estimated, the procedures in the $\boldsymbol{J}$ and $\boldsymbol{K}$ matrices generation can be given by simple matrix-matrix operations without the ERI estimations. Since the well-optimized linear algebra library, such as BLAS, for various computer systems is available, the $\boldsymbol{J}$ and $\boldsymbol{K}$ matrices are constructed by simple and effective method.

2.2 Low-rank approximate Cholesky decomposition

To get smaller super matrix without losing calculation precision is an advantage in terms of memory capacity and number of floating-point arithmetic operations. The CD of n-dimensional matrix usually gives n-dimensional triangular matrix. In this study, the pivoted CD to compute low-rank approximations [4] is adopted. The resulting truncation error is rigorously controlled in terms of the trace norm.

In general, accessing all elements of the full matrix is needed in CD. Consequently, the huge full matrix should be temporarily stored in memory. The elements that are left unused in the pivoted CD are also included in the full matrix. In order to produce Cholesky vectors on limited memory, it is desired that the full matrix is not kept in memory, and that the ERIs, which are elements of the full matrix, are directly calculated. Putting all the above components together, we arrive at the algorithm (Fig. 1). Since there is no necessary to obtain the huge super matrix temporarily, the Cholesky vectors are gained by minimal computer resources.

Although the Cholesky vectors are directly obtained and the huge super matrix is not required, the Cholesky vectors remains large size, and may not be stored in memory. Hence the Cholesky vectors $\boldsymbol{L}$ should be distributed to all processesor elements (PEs). In this study, the entries of the Cholesky vectors are stored as set of row vectors for efficiency. On this algorithm, the access direction of the matrix is only the row direction. With keeping diagonal elements, ℓ_{m,π_i}, in all PEs, the operations for row-vectors elements of the Cholesky vectors held in each PE are independent (line: 11), so no communications are required. And, the molecular integrals may be carried out only for the row-vectors of Cholesky vectors (line: 9)。 This method can be directly parallel processing the Cholesky vectors.

3 Results and discussion

We implemented this method to ProteinDF program [5]. The canonical molecular orbital calculations of Insulin (51 residues) and Interleukin (133 residues) are carried out by using ProteinDF. The number of AO pairs in Insulin and Interleukin, which were 9,952,491 and 70,918,095, were dramatically reduced to 169,842(1.7%) and 460,785 (0.6%) by CDAM pre-screening ($\tau = 10^{-4}$), respectively. So, the combination of CDAM and pivoted CD were considered to be valid and feasible for canonical molecular orbital computations. In order to clarify effective threshold value, further investigation is required.

4 Conclusion

A new ERI parallel computation method based on modified CDAM method and the pivoted low-approximate CD was developed and implemented to ProteinDF. The method was able to control the efficiency of molecular integral estimations with improving the accuracy. As the way of achievement on the ERI computation efficiency in this method resembles the well-known cutoff method, negligible ERI values are strictly estimated based on CD. Additionally, the method gives smaller and optimized Cholesky vectors. We emphasis that the Cholesky vectors $\boldsymbol{L}$ are calculated and stored before the SCF loop, the Coulomb and exchange part of Fock matrix are given by the multiplication of matrices, without expensive ERI computations. Though required memory is larger than the direct method, in which the ERIs are calculated at every SCF iteration in return, we can get high efficiency of computation on the distributed parallel computer systems. The memory size per computer node tends to increase, and this problem will be overcome. We suppose that this method is useful for electronic structure calculation of the large molecule system in parallel and distributed memory system.

Acknowledgements

A part of this research was done in "Revolutionary Simulation Software for the 21st century (RSS21)" project supported by Research and Development for Next-generation Information Technology of Ministry of Education, Culture, Sports, Science and Technology (MEXT). And this research was partly supported by "Research and Development of the Next-Generation Integrated Simulation of Living Matter", a part of the Development and Use of the Next-Generation Supercomputer Project of the Ministry of Education, Culture, Sports, Science and Technology (MEXT). We thank the Institute of Statistical Mathematics for the facilities and the use of Fujitsu PRIMERGY RX200S5.

References

[1] Y. Tokita, S. Yamada, W. Luo, Y. Goto, N. Bouley-Ford, H. Nakajima, Y. Watanabe, Protein Photoconductors and Photodiodes, Angew. Chem. Int. Ed., 50, 11663 –11666, 2011.

[2] F. Aquilante, R. Lindh, T.B. Pedersen, Unbiased auxiliary basis sets for accurate two-electron integral approximations, J. Chem. Phys., 127, 114107, 2007.

[3] Y. Okiyam, T. Nakano, K. Yamashita, Y. Mochizuki, N. Taguchi, S. Tanaka, Acceleration of fragment molecular orbital calculations with Cholesky decomposition approach, Chemical Physics Letters, 490, 84-89, 2010.

[4] H. Harbrechta, M. Petersa, R. Schneider, On the low-rank approximation by the pivoted Cholesky decomposition, Appl. Numer. Math., 62, 428-440, 2012.

[5] T. Inaba, F. Sato, Development of parallel density functional program using distributed matrix to calculate all-electron canonical wavefunction of large molecules, J. Comput. Chem., 28, 984-995, 2007.

line	pseudo code
1	**begin**
2	**set** $m \coloneqq 0$;
3	**set** $\boldsymbol{d} \coloneqq diag(\boldsymbol{A})$ and $error = \|\boldsymbol{d}\|_1$;
4	**initialize** $\boldsymbol{\pi} \coloneqq (1, 2, \dots, n)$;
5	**while** $error > \varepsilon$ **do**
6	**set** $i \coloneqq argmax\left\{d_{\pi_j} : j = m, m+1, \dots, n\right\}$;
7	**swap** π_m and π_i;
8	**set** $\ell_{m,\pi_i} \coloneqq \sqrt{d_{\pi_m}}$;
9	compute molecular integral $a_{\pi_m,\pi_i} = (\pi_m \| \pi_i)\ (i; m+1 : n)$
10	**for** $m+1 \le i \le n$ **do**
11	**compute** $\ell_{m,\pi_i} \coloneqq \left(a_{\pi_m,\pi_i} - \sum_{j=0}^{m-1} \ell_{j,\pi_M} \ell_{J,\pi_I}\right) / \ell_{m,\pi_m}$;
12	**update** $d_{\pi_i} \coloneqq d_{\pi_i} - \ell_{m,\pi_i}^2$;
13	**compute** $error \coloneqq \sum_{i=m+1}^{n} d_{\pi_i}$;
14	**increase** $m \coloneqq m + 1$;
15	**end**

Fig. 1. Pseudo code for pivoted Cholesky decomposition of ERIs