Study on Raffenetti's P File Format in Conventional *Ab Initio* Self-Consistent-Field Molecular Orbital Calculations in Parallel Computational Environment

Hiroyuki TERAMAE[a]* and Kazushige OHTAWARA[b,c]

[a]Department of Chemistry, Faculty of Science, Josai University
1-1 Keyakidai, Sakado, Saitama 350-0295, Japan
[b]ATR Adaptive Communication Research Laboratories
2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288, Japan
[c] Present address: Technology Development Division, Victor Company of Japan, Ltd.
58-7 Shinmei-cho, Yokosuka, Kanagawa 239-8550, Japan
**e-mail: teramae@josai.ac.jp*

(Received: August 7, 2008; Accepted for publication: October 17, 2008; Advance publication: November 22, 2008)

We compare the CPU time and the wall clock time of the Raffenetti's P file algorithm with the usual algorithm on the two electron integrals storing with four suffixes of the *ab initio* Hartree-Fock calculations. The calculations are performed with the flutoprazepam, triazolam, clotiazepam, etizolam, and flutazolam molecules. These molecules are all minor-tranquilizers with the benzodiazepine or thienodiazepine backbone. The 3-21G basis sets are employed. Almost in all cases, P file algorithm gave slower speed than the usual algorithm. The number of two electron integrals increases almost two times larger than the usual algorithms. In a large molecule, the matrix of the two electron integrals becomes very sparse and the recombination of the integrals just increases the total number of the integrals. It is concluded that the P method sometimes calculates faster but sometimes does not. In a large scale calculation, it should be suggested to perform a test calculation to confirm which method is faster prior to the real calculations.

Keywords: Raffenetti's PK file, Parallel computation, Molecular orbital

1 Introduction

A Hartree-Fock molecular orbital calculation (*ab initio* molecular orbital calculations) became very popular and has been establishing a new field of chemistry as a computer experiment, since it has been applied on the electronic structure calculations of various molecules for the last three decades. In the applications of the calculations on more realistic molecules, it has been an important problem to handle the two electron integrals, which increases very rapidly as the molecular size increases.

If each molecular orbital is expanded by the basis functions and expressed as,

$$\Psi_i = \sum_{r=1}^{N} c_{ri}\phi_r$$

The Hartree-Fock equations are expressed as,

$$\sum_{s=1}^{N}(F_{rs} - \varepsilon_i S_{rs})c_{ri} = 0$$

where

$$S_{rs} = \int \phi_r(1)\phi_s(1)dr$$

$$F_{rs} = H_{rs} + \sum_{i=1}^{N}\sum_{u=1}^{N} P_m\left[2\langle rs|tu\rangle - \langle rt|su\rangle\right]$$

$$H_{rs} = \int \phi_r^*(1)\left[-\frac{1}{2}\left(\frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial y_1^2} + \frac{\partial^2}{\partial z_1^2}\right) + \sum_{A=1}^{atom}\frac{Z_A}{r_{1A}}\right]\phi_s(1)dr$$

DOI: 10.2477/jccj.H2020

H_{rs} is the core Hamiltonian and is a one electron integral.

$$\langle rs|tu\rangle = \iint \phi_r^*(1)\phi_s(1)\frac{1}{r_{12}}\phi_t^*(2)\phi_u(2)dr_1dr_2$$

is a two electron integral. The suffixes r, s, t, and u run from 1 to the total number of basis function N, and therefore the total number of the two electrons integral is of the order of N^4.

The density matrix element is expressed as

$$P_m = \sum_{i=1}^{occ} c_{ti}c_{ui}$$

and contains the variations c_{ti} to be resolved by the Hartree-Fock equation. The Hartree-Fock equations, therefore, should be solved iteratively (self consistent field, SCF). The two electron integrals are required several times to calculate the Fock matrix elements F_{rs}.

In the classical programs such as Gaussian 70 used the two electron integral file with the packed four suffixes we utilize the file iteratively during the SCF calculations [1, 2]. From the symmetries of the suffixes, there are integrals of the same value. These are eliminated from the calculations and therefore there are six types of contributions to the Fock matrix element from a two electron integral as following:

$$\begin{aligned}
&F_{rs}\cdots P_{tu}\langle rs|tu\rangle\\
&F_{tu}\cdots P_{rs}\langle rs|tu\rangle\\
&F_{rt}\cdots -\frac{1}{2}P_{su}\langle rs|tu\rangle\\
&F_{su}\cdots -\frac{1}{2}P_{rt}\langle rs|tu\rangle\\
&F_{ru}\cdots -\frac{1}{2}P_{st}\langle rs|tu\rangle\\
&F_{st}\cdots -\frac{1}{2}P_{ru}\langle rs|tu\rangle
\end{aligned}$$

Raffenetti has proposed a more efficient procedure to calculate the two-electron contribution to the Fock matrix, nowadays widely known as P super matrix algorithm [3]. Hereafter we call it P method, and another traditional four suffixes method as NOP method. The basis of the P method is to make a recombination of two electron integrals like,

$$I_{rstu} = \langle rs|tu\rangle - \frac{1}{4}\langle rt|su\rangle - \frac{1}{4}\langle ru|st\rangle$$

The contribution to the Fock matrix element using P method becomes very simple.

$$\begin{aligned}
&F_{rs}\cdots P_{tu}I_{rstu}\\
&F_{tu}\cdots P_{rs}I_{rstu}
\end{aligned}$$

If the number of two electron integrals does not change before and after the recombination and if the overhead for the recombination is small enough, the computational time will be much faster. The total number of multiply/add instructions would have been greatly reduced because the instructions decrease from six to two. The P method did work well, and both Hondo [2] and Gaussian [4] series of programs incorporate P method. Nowadays, the direct SCF method [5], which does not store the two electron integrals but calculates them repeatedly, is usually used. The P method, therefore, does not study well if the method works in any case.

On the other hand, due to the recent development of the microprocessor, it becomes possible to utilize the personal computer cluster to make the Hartree-Fock molecular orbital calculations with the parallel processing of the two electron integrals and the Fock matrix mentioned above. Because we now can use many CPUs and large size of memories that could not be supposed previously, it sometimes happens to break a previous common sense, that is, a *paradigm shift*. For example, in our previous work [7], we reported that the efficiency of the files system becomes good because the files for the two electron integrals are divided and stored in each local disk system when applying the parallel processing. The input/output (I/O) processing is applied on the divided files and it naturally becomes the parallel I/O. The operating system sometimes uses the memory buffer for I/O operation and in the extreme case all the two electron integrals are processed on memory. In this case, the processing time is very fast, because the I/O operation will be done just one time and the remaining read operation will be processing on memory. We can achieve the faster processing without re-writing programs. Within our molecular orbital calculations, therefore, the conventional SCF method that stores the integrals on files is faster than the direct SCF method. In the conventional SCF method, the treatment of the two electron integrals is very important as mentioned above, and it becomes important to study the P method which is really faster than the NOP method.

We are recently developing the molecular dynamics calculations based on the *ab initio* Hartree-Fock molecular orbital calculations, which requires the iterative calculations of 1000-3000 points. The total performance becomes large enough if the reduction of the single point calculation is so small. In the present article, therefore, we perform the moderated sized parallel processing of the *ab initio* Hartree-Fock calculations from 217 basis functions to 274 basis functions, and compare the CPU and wall clock times of P and NOP methods.

2 Method of Calculations

Table 1 shows our computational environment. The calculations are performed with the use of an 8 CPU/8 chassis PC cluster of Intel Pentium 4 CPU 2.4GHz with Intel 845 chipset. The network is 1000BaseT gigabit Ethernet.

RedHat Linux version 8 is used for the operating system of the cluster system. The general molecular orbital program package GAMESS [5] is used throughout this study. The original code of GAMESS is used, because the code for two electron integrals was already written and suitable for parallel processing. The socket communication library within GAMESS is used.

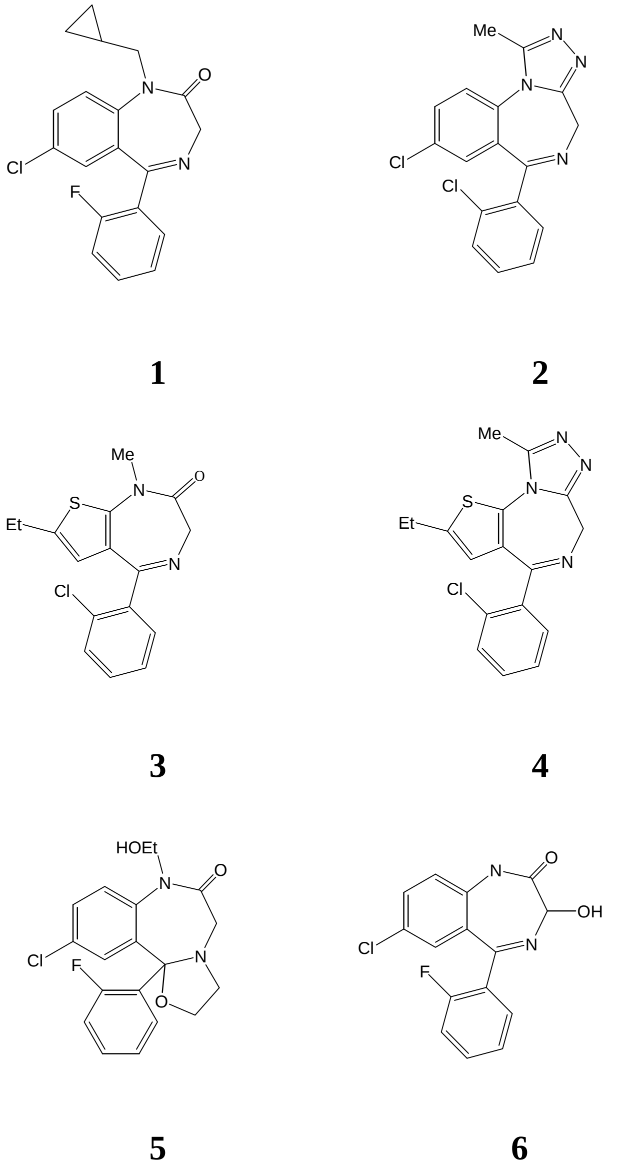
The computational speed is measured with a series of minor-tranquilizers with the benzodiazepin and thienodiazepin frameworks; flutoprazepam (**1**, $C_{19}H_{16}ClFN_2O$), triazolam (**2**, $C_{17}H_{12}Cl_2N_4$), clothiazepam (**3**, $C_{16}H_{15}ClN_2OS$), etizolam (**4**, $C_{17}H_{16}ClN_4S$), flutazolam (**5**, $C_{19}H_{18}ClFN_2O_3$), and lorazepam (**6**, $C_{15}H_{10}Cl_2N_2O_2$) molecules (Scheme 1). The 3-21G basis set [8] is used throughout this study. We repeat the calculations ten times of single SCF and gradient of each molecule and take the fastest time among them.

3 Results and Discussion

Table 2 shows the molecular formula, number of atoms, number of basis functions, and the amount of two electron integral files of each molecule. When computing with single CPU, these are all severe calculations. In the case of flutazolam, which is the largest calculation among the present study, the amount of the file exceeds 2GB. In the parallel environment, however, all files are buffered on the main memory when 4-8 CPUs are used. We would like to note the following: even if the operating system cannot handle the files larger than 2GB, the calculation is still possible if you can divide files under the parallel environment.

Table 1. The configuration of PC cluster.

CPU	8CPU Pentium4 2.4GHz, 512K Cache
Chipset	Intel 845
Memory	1 GB DDR266/ board
Network	1000 BaseT Ethernet
Hard Disk	60GB / 5400rpm
Operating System	Linux kernel 2.4/RedHat 8.0
Fortran Compiler	Gnu Fortran77
Parallel Library	MPICH ver.1.2.4



Scheme 1.

Table 2. Molecular formula, number of atoms, number of basis functions, and the amount of the two electron integrals of each molecules.

Molecule	Flutoprazepam	Toriazolan	Clotiazepam	Etizolam	Fultazolam	Lorazepam
formula	$C_{19}H_{16}ClFN_2O$	$C_{17}H_{12}Cl_2N_4$	$C_{16}H_{15}ClN_2OS$	$C_{17}H_{16}ClN_4S$	$C_{19}H_{18}ClFN_2O_3$	$C_{15}H_{10}Cl_2N_2O_2$
atoms[a]	40	35	36	38	44	31
Basis[b]	248	239	227	245	274	217
TEI[c]	1.5GB	1.3GB	1.0GB	1.2GB	2.9GB	0.9GB

[a] Number of atoms. [b] Number of basis functions. [c] Amount of two electron integrals in giga byte unit.

Table 3. Lists of CPU, system and wall clock time of P and NOP methods. N is number of CPUs.

Molecule	NOP				P			NOP/P Ratio	
	N	CPU	SYS	Wall	CPU	SYS	Wall	CPU+SYS	Wall
clotiazepam	1	124.98	21.23	259.14	113.34	51.18	750.53	0.89	0.35
	2	64.69	8.63	94.58	59.71	26.06	320.55	0.85	0.30
	4	34.38	4.62	50.58	32.04	10.64	100.82	0.91	0.50
	8	19.20	2.72	32.50	18.10	4.84	33.74	0.96	0.96
etizolam	1	157.29	28.89	397.69	142.25	72.07	994.16	0.87	0.40
	2	81.44	11.91	119.48	76.82	36.00	444.19	0.83	0.27
	4	43.16	5.98	62.57	41.40	16.23	163.19	0.85	0.38
	8	23.92	3.63	38.68	23.22	6.59	42.18	0.92	0.92
flutazolam	1	265.37	88.77	1078.06	233.41	116.05	1660.53	1.01	0.65
	2	145.96	39.85	451.09	124.32	56.85	795.94	1.03	0.57
	4	73.22	15.66	113.64	66.61	28.44	340.35	0.94	0.33
	8	40.35	8.98	66.96	37.30	12.13	113.36	1.00	0.59
flutoprazepam	1	179.72	38.16	482.96	163.63	82.20	1134.93	0.89	0.43
	2	92.95	12.83	136.09	87.73	41.13	521.62	0.82	0.26
	4	49.53	6.92	71.23	46.18	17.12	197.99	0.89	0.36
	8	27.65	4.15	41.48	26.30	7.10	46.97	0.95	0.88
lorazepam	1	109.40	18.92	220.66	96.37	43.55	564.10	0.92	0.39
	2	56.41	8.37	84.96	50.68	18.63	198.83	0.93	0.43
	4	31.59	4.68	46.70	26.62	7.92	45.65	1.05	1.02
	8	17.32	2.56	28.92	14.70	4.57	29.17	1.03	0.99
triazolam	1	152.96	29.02	401.02	138.81	67.93	963.88	0.88	0.42
	2	78.96	11.62	114.63	71.59	33.13	431.96	0.86	0.27
	4	44.16	5.98	63.46	39.22	15.03	152.92	0.92	0.41
	8	24.09	3.68	37.05	21.51	6.38	39.88	1.00	0.93

Table 3 shows the CPU and wall clock time for the P and NOP methods, respectively. In all cases for N=1, it is clearly shown that the wall clock time by the NOP method is shorter and 0.35-0.65 times smaller than that of the P method. Furthermore, when comparing the sum of CPU and system time, the NOP method shows shorter time, except for the results of fultazolam that are almost the same.

Concerning the wall clock time, in all molecules, the difference becomes smaller when the number of CPUs increases. After all two electron integrals are buffered on the main memory, the wall clock time by the P method decreases more quickly than that by the NOP method showing the difference between the two methods. In the case of lorazepam, that is the smallest calculation of the present work, the difference between two methods disappears when 4 CPUs are used. In other molecules, the difference between the two methods also disappears when 8 CPUs are used again except for the fultazolam case. In the case of fultazolam, the NOP method result is still 0.59 times shorter than that of the P method even in the 8 CPU case, and the difference does not disappear in the present study. However, we consider from Table 3 that the difference between the two methods will vanish as the number of CPUs increases.

Table 4. The numbers of two electron integrals for NOP and P method.

Molecule	NOP	P	NOP/P Ratio
clotiazepam	91696888	176425145	0.52
etizolam	114762339	227059480	0.51
flutazolam	192258830	366232404	0.52
flutoprazepam	130469433	255888062	0.51
lorazepam	82657615	154668492	0.53
triazolam	114937886	219850389	0.52

The difference in the wall clock times between the two methods is brought by the difference of the amount of the files of the two electron integrals. We usually handle just the two electron integral that is larger than a certain threshold value (10^{-8} in the present study). Table 4 shows the number of two electron integrals by the P and NOP methods used in the present calculations. It is worthwhile to note that the number by the P method is almost two times larger than that by NOP method. From the definition of the P method, I_{rstu} has a certain value if the integral $\langle rs|ru\rangle$ is smaller than the cutoff value but either $\frac{1}{4}\langle rt|su\rangle$ or $\frac{1}{4}\langle ru|st\rangle$ is larger than the cutoff value. As a result, the number of two electron integrals increases

Table 5. The difference of computational time and number of two electron integrals of C_2F_6 molecule. N is number of CPUs.

		NOP			P			NOP/P ratio	
Basis Set	N	CPU	SYS	Wall	CPU	SYS	Wall	CPU+SYS	Wall
6-31G*	1	17.68	3.15	26.39	11.90	3.76	19.90	1.33	1.33
	2	9.12	1.70	16.10	6.32	1.98	11.20	1.30	1.44
	4	4.83	0.98	9.08	3.46	1.01	8.68	1.30	1.05
	8	2.86	0.66	6.66	2.02	0.65	5.32	1.32	1.25
	Number of Integrals			20868299			23940759	Ratio	0.87
3-21G	1	1.52	0.36	3.14	1.07	0.46	2.75	1.23	1.14
	2	0.85	0.28	2.53	0.64	0.27	2.33	1.24	1.09
	4	0.52	0.19	2.48	0.42	0.23	2.37	1.09	1.05
	8	0.39	0.17	2.91	0.33	0.18	2.89	1.10	1.01
	Number of Integrals			2124399			277210	Ratio	0.76

in the case of P method. In the calculation of the relatively larger molecule like the present calculations, almost all of the two electron integrals are under cutoff value and the effect of increasing the number of two electron integrals denoted above becomes significant. In Table 3, the system time of the P method is always larger than that of the NOP method, which indicates the overhead for the file I/O operation is larger in the case of P method. In a smaller molecule, this is not true because a large part of the two electron integrals have values larger than the cutoff threshold. The time required for the calculation of P method, is therefore, smaller than that of NOP method. Table 5 shows the results of C_2F_6 molecule case as an example of a small molecule. The number of two electron integrals by NOP method is 0.87 times when the 6-31G** basis set is used, and 0.76 times when the 3-21G basis set is used. In both basis sets, the calculations finish faster in P method. It should also be noted that the degree of acceleration is larger in the 6-31G** basis set case than in the 3-21G basis set case, which is easily seen from NOP/P factor of the number of two electron integrals.

In the present paper, we have studied the CPU time and the wall clock time required for the *ab initio* Hartree-Fock molecular orbital calculations with and without the Raffenetti's P super matrix algorithm under the parallel environment using the PC cluster. As realistic examples, the six different drug molecules of the minor-tranquilizer and the 3-21G basis set are used. In almost all of the cases, the P method cannot calculate faster than the NOP method in such a large calculation. It should be concluded that the P method sometimes calculates faster but sometimes does not. In large scale of calculations, it should be suggested to perform a test calculation to confirm which method is faster prior to the real calculations.

We are grateful to Dr. Kazumasa Shinjo and Dr. Shinsuke Shimogawa of ATR Adaptive Communication Laboratories for stimulating discussions and suggestions. This work was partially supported by Telecommunication Advancement Organization of Japan (TAO).

References

[1] For example, see,
W. J, Hehre, L. Radom, P. v. R, Schleyer, and J. A. Pople, *Ab Initio Molecular Orbital Theory*, Wiley, New York (1986), and references cited therein.

[2] M. Dupuis, J. Rys, and H. F. King, *J. Chem. Phys.*, **65**, 111 (1976).

[3] R. C. Raffenetti, *Chem. Phys. Lett.*, **20**, 335 (1973), see also page 54 of reference [1].

[4] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, V. G. Zakrzewski, J. A. Montgomery, Jr., R. E. Stratmann, J. C. Burant, S. Dapprich, J. M. Millam, A. D. Daniels, K. N. Kudin, M. C. Strain, O. Farkas, J. Tomasi, V. Barone, M. Cossi, R. Cammi, B. Mennucci, C. Pomelli, C. Adamo, S. Clifford, J. Ochterski, G. A. Petersson, P. Y. Ayala, Q. Cui, K. Morokuma, P. Salvador, J. J. Dannenberg, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. Cioslowski, J. V. Ortiz, A. G. Baboul, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. Gomperts, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, J. L. Andres, C. Gonzalez, M. Head-Gordon, E. S. Replogle, and J. A. Pople, *Gaussian98*, Gaussian, Inc., Pittsburgh PA (2001).

[5] J. Almlf, J. K. Faegri, and K. Korsell, *J. Comput. Chem.*, **3**, 385-399 (1982).

[6] M. W. Schmidt, K. K. Baldridge, J. A. Boatz, S. T. Elbert, M. S. Gordon, J. J. Jensen, S. Koseki, N. Matsunaga, K. A. Nguyen, S. Su, T. L. Windus, M. Dupuis, J. A. Montgomery, *J. Comput. Chem.*, **14**, 1347-1363 (1993).

[7] H. Teramae and K. Ohtawara, *J. Chem. Software*, **8**, 55-61 (2002).

[8] J. S. Binkley, J. A. Pople, W. J. Hehre, *J. Am. Chem. Soc.*, **102**, 939-947 (1980).