Large-Scale Electronic-State Calculations of Protein-Ligand Systems for Drug Design with Fragment Molecular Orbital Method

Project Representative

Shigenori Tanaka Graduate School of System Informatics, Kobe University

Authors

Shigenori Tanaka Graduate School of System Informatics, Kobe University

Yuji Mochizuki Faculty of Science, Rikkyo University

Katsumi Yamashita NEC Soft Ltd.

Chiduru Watanabe Institute of Industrial Science, The University of Tokyo

Yoshio Okiyama Institute of Industrial Science, The University of Tokyo

Kaori Fukuzawa Mizuho Information and Research Institute Inc.

Tatsuya Nakano National Institute of Health Sciences

Large-scale fragment molecular orbital (FMO) calculations for protein systems were performed on the Earth Simulator. The four-body corrected fragment molecular orbital (FMO4) method was recently developed and implemented at the second-order Møller-Plesset perturbation (MP2) level. A series of accuracy tests relative to the previous two-body (FMO2) and three-body (FMO3) treatments were carried out. As expected, FMO4 provided better results of total energies in comparison with the reference values by conventional molecular orbital calculations. A nonconventional fragmentation by discriminating the main and side chains in amino acid residues was then examined for a large complex of HIV-1 protease (total 198 residues) with lopinavir, where the four-body corrections were shown to be substantial.

1. Introduction

The fragment molecular orbital (FMO) scheme [1-3] enables one to perform fully quantum-mechanical calculations for large-scale molecular systems like proteins at affordable cost of computations with parallelism. It is recognized that the FMO2 scheme [1-3] in which the fragments up to the dimers are taken into account provides reasonable accuracy in energy calculations such as that for interaction energy analyses to describe the details of protein-ligand docking in the pharmacophore [2, 3]. However, the inclusion of three-body terms (FMO3) is desirable to ensure the total reliability in some cases, e.g., hydrogen-bonded water clusters [2-4]. Fedorov and Kitaura thus developed the three-body corrected FMO scheme at the levels of Hartree-Fock (HF) approximation (FMO3-HF) [5, 6] and second-order Møller-Plesset perturbation (MP2) theory (FMO3-MP2) [7] on the GAMESS-US package. Recently, Katouda [8] provided an alternative FMO3-MP2 scheme with the technique of resolution-of-identity approximation of integrals in the GAMESS-US. On the other hand, with the ABINIT-MPX package, Fujita et al. [9] examined the importance of three-body contributions and

also the matching with several approximations in the FMO3-HF energy for the hydration of a sodium ion. Both FMO3-MP2 energy and gradient were then implemented with an efficient integral-direct parallelism [10].

Here, we report the development of four-body FMO (FMO4) calculations of HF and MP2, that is, FMO4-HF and FMO4-MP2, on the ABINIT-MPX package [11]. The four-body corrections were already proposed and tested in the literature [12-15]. These references showed certain improvements by the four-body treatment over the three-body one, in particular for the calculations of solids [12, 14]. The present motivation to develop the FMO4 method arises from the interest in more detailed modeling for the fragment-based drug discovery or design (FBDD) [16, 17]. In FBDD, it is highly desirable that various functional groups of ligands are divided as the respective fragments and also that the main and side chains of amino acid residues in proteins are segmented correspondingly. Such a way of fragmentation is of nonconventional type in earlier FMO calculations [1-3], while its importance has been recognized [18]. A large complex of HIV-1 protease and lopinavir (previously

employed in Ref. [19]) is calculated with 1024 processors of the Earth Simulator (ES2) as a currently available platform of massively parallelized computation.

2. Methods

In the original scheme of FMO method [1], the FMO2-HF energy ("HF" is omitted here for simplicity) is given by the energies of fragment monomers and dimers:

$$E^{FMO2} = \sum_{I>J} E_{IJ} - (N-2)\sum_I E_I = \sum_{I>J} \Delta E_{IJ} + \sum_I E_I \quad (1)$$

$$\Delta E_{IJ} \equiv E_{IJ} - E_I - E_J \quad (2)$$

where N is the number of fragments in a given system and IJ are the fragment dimer indices. The HF calculation for each monomer is carried out under the presence of environmental electrostatic potential (ESP) which is a key point of the FMO scheme [2-4]. The fragment indices are distributed over the groups of processors (upper level), and the Fock matrix constructions are then parallelized with respect to the indices of atomic orbital (AO) within an assigned group (lower level). This dual parallelization accelerates the computations significantly [2-4]. For further acceleration, Nakano et al. [20] devised a couple of approximations to evaluate the ESP matrix elements based on the Mulliken AO charge (ESP-AOC) and the Mulliken point charge (ESP-PTC).

The FMO3-HF energy formula [5, 6, 9] as

$$E^{FMO3} = \sum_{I>J>K} E_{IJK} - (N-3)\sum_{I>J} E_{IJ} + \frac{(N-2)(N-3)}{2}\sum_I E_I$$
$$= \sum_{I>J>K} \left[\Delta E_{IJK} - \Delta E_{IJ} - \Delta E_{IK} - \Delta E_{JK} \right] \quad (3)$$
$$+ \sum_{I>J} \Delta E_{IJ} + \sum_I E_I$$

$$\Delta E_{IJK} \equiv E_{IJK} - E_I - E_J - E_K \quad (4)$$

may be regarded as a next-order form of many-body expansion [21] and has been used widely to improve the numerical accuracy of total energies [4]. In FMO3, a trimer is specified by IJK, and the parallelized HF calculations are carried out when three composite monomers are adjacent within a threshold of van der Waals contact [9]. Care should be taken for the application of ESP approximations [20] to the FMO3-HF calculations, as addressed in Refs. [6, 9]. The FMO3-MP2 correction [7, 8, 10] may then be considered for the HF-calculated trimers. Fedorov et al. [5, 6] found that the accuracy of FMO3-HF with single-residue fragmentation is better than that of FMO2-HF with double-residue fragmentation for model Ala-polymers, illuminating the importance of explicit three-body corrections. The following literature [7] reported the corresponding MP2 results.

The formulas for the four-body corrections were presented in Ref. [14] for the modeling of solid systems and also in Ref. [15] for proteins. The form of FMO4-HF energy is essentially the same as that of many-body expansion series:

$$E^{FMO4} = \sum_{I>J>K>L} E_{IJKL} - (N-4)\sum_{I>J>K} E_{IJK} + \frac{(N-3)(N-4)}{2}\sum_{I>J} E_{IJ}$$
$$- \frac{(N-2)(N-3)(N-4)}{6}\sum_I E_I$$
$$= \sum_{I>J>K>L} \{\Delta E_{IJKL} - \Delta E_{IJ} - \Delta E_{IK} - \Delta E_{IL} - \Delta E_{JK} - \Delta E_{JL} - \Delta E_{KL}$$
$$- [\Delta E_{IJK} - \Delta E_{IJ} - \Delta E_{IK} - \Delta E_{JK}] - [\Delta E_{IJL} - \Delta E_{IJ} - \Delta E_{IL} - \Delta E_{JL}]$$
$$- [\Delta E_{IKL} - \Delta E_{IK} - \Delta E_{IL} - \Delta E_{KL}] - [\Delta E_{JKL} - \Delta E_{JK} - \Delta E_{JL} - \Delta E_{KL}]\}$$
$$+ \sum_{I>J>K} [\Delta E_{IJK} - \Delta E_{IJ} - \Delta E_{IK} - \Delta E_{JK}]$$
$$+ \sum_{I>J} \Delta E_{IJ} + \sum_I E_I \quad (5)$$

$$\Delta E_{IJKL} \equiv E_{IJKL} - E_I - E_J - E_K - E_L \quad (6)$$

where the actual tetramer HF calculation is performed similarly to the trimer case. Ref. [15] showed that the tetramer corrections improve the total energy of Leu1-zervamicin at the HF/STO-3G level. The minimal set of STO-3G basis should, however, be too small to conclusively assess the accuracy, and further testing with at least double-zeta (DZ) quality basis (e.g., 6-31G) would be necessary at the correlated level. The four-body MP2 corrections are then taken into account straightforwardly.

Employing the DZ or DZ-plus-polarization (DZP) basis sets, the size of fragment tetramer would be demanding for the FMO calculations of real proteins potentially containing twenty variations of amino acid residues from the smallest Gly to the largest Trp, unless an alternative protocol to the conventional single-residue fragmentation is taken. As addressed above, the segmentation of main and side chains in amino acid residues (or the bond cutting at both C_α and C_β atoms) is rather essential for FBDD [16, 17], and it may be beneficial to make the FMO4 calculations tractable. Nonetheless, the increased number of fragments with this new fragmentation is an alternative factor to enlarge the gross computational cost covering up to tetramers. The use of massively parallel computers is thus encouraged to reduce the computation time in large-scale applications of FMO4-MP2.

3. Results

The FMO4 method was implemented in a recent version of ABINIT-MPX with the vectorizable HF and MP2 modules (under MPI control) of Ref. [22]. The frozen-core restriction was imposed at the MP2 stage throughout. As a test case, the HIV-1 protease-lopinavir complex was calculated at the FMO4-MP2/6-31G level, by using 1024 processors of ES2. This model was the same one as that employed in Ref. [19]. The number of amino acid residues of HIV-1 protease was 198 (99 of each subunit). The number of fragments by the main/side chain fragmentation was 358, where no Cys-Cys bridge was contained. The lopinavir ligand was divided into 4 fragments,

and a water molecule crucial in the hydrogen-bond network was also included in the pharmacophore. The numbers of atoms, fragments and 6-31G basis AOs were thus 3225, 363 (203 in the conventional fragmentation) and 17423, respectively. The number of used nodes of ES2 was 128, and each node consisted of 8 vector processors (102.4 GFLOPS per processor) with 128 GB shared memory. The fragments from monomers to tetramers were processed in a single node throughout. The ESP-AOC approximation [20] (in which the two-electron integrals were computed, unlike the classical approximation of ESP-PTC with Mulliken charges) was adopted for this protease complex.

The HF and MP2-corrected energies of the HIV-1 protease complex are given in Table 1 (upper part). The FMO4 results are used as a tentative reference since the regular MO calculations of this sized molecule were impossible with the ABINIT-MPX program; we regard the energy with conventional main-chain fragmentation as the best effort value. An unacceptable difference is found for the FMO2 results with the main/side chain fragmentation, which implies that at least FMO3 expansion is required for reliable analyses (even for the inter-fragment interaction energy (IFIE) [23, 24]).

The breakdown timings are tabulated in the lower part of Table 1. Let us recall here that the monomer-stage calculation requires the achievement of SCC (self-consistent charge) condition [2-4]. This is the reason why its computational time is longer than that of the dimer stage for both fragmentations. It is notable that the incremental cost of MP2 over HF is maintained quite small for the calculations of monomers and dimers [22]. The MP2 calculations for trimers and tetramers show sizable increases in the computational time over the HF

calculations, leading to the incremental cost factor relative to FMO2 (about ten times for FMO4). Comparison in timings between two fragmentations indicates that the tetramer part governs the slightly increased cost of FMO4 calculation with the nonconventional fragmentation of main/side chains. The FMO4-MP2/6-31G job was actually completed in 1.4 hours (to be compared with 1.2 hours of the conventional case). If massively parallel computing resources such as the current ES2 or the K-computer are available, the FMO4 calculations (with much long task list of up to fragment tetramers) can be carried out for real proteins, in short time without the ESP-PTC approximation which has a vulnerability of the Mulliken partitioning of charges [20]. In addition, the computational time can be reduced with the use of the Cholesky decomposition technique [25], as seen in Fig. 1.

Although we here refrain from the presentation of IFIE results of this complex [24], the enhanced resolution of analyses matches with the FBDD scheme including the lead search and optimization [16, 17]. We hope that the FMO4 method will become a useful tool to accelerate drug discovery and design. Manifestly, several efforts are necessary to improve the speed and reliability of FMO calculations. The introduction of the fast multipole method to evaluate the ESP elements is a plausible option in this regard.

In this report, we have addressed the development of the four-body FMO (FMO4) scheme [11]. Test calculations were systematically carried out at the HF and MP2 levels in comparison with the reference energies of regular MO calculations. It was confirmed that the FMO4 method is better in the accuracy of energy than the FMO3 method by one-order

Table 1　Total energies and timings of FMO calculations for HIV-1 protease - lopinavir complex.

Total energy	Main/side frg.	(total 363)	Main frg.	(total 203)
(au)	HF	MP2	HF	MP2
FMO2	-77554.2779	-77709.0268	-77589.3751	-77744.7417
Diff.(4)[a]	35.4754	36.1544	0.3059	0.3737
FMO3	-77590.0557	-77745.5971	-77589.6024	-77745.0359
Diff.(4)[a]	-0.3024	-0.4159	0.0786	0.0796
FMO4	-77589.7533	-77745.1812	-77589.6810	-77745.1155
Diff.(M)[b]	-0.0723	-0.0657		
Timing[c]				
(sec)				
Monomer	356.0	359.2	252.0	253.1
Dimer	120.8	141.2	131.7	175.3
Trimer	765.3	1040.1	648.2	1110.0
Tetramer	2352.3	3374.4	1358.2	2663.2
Total	3594.4	4914.9	2390.1	4201.6

[a] Relative to the corresponding FMO4 energy.

[b] Relative to the FMO4 energy with conventional fragmentation in main chains (as the best effort value).

[c] Breakdown timing analysis of the FMO4 jobs with 1024 processors of ES2.
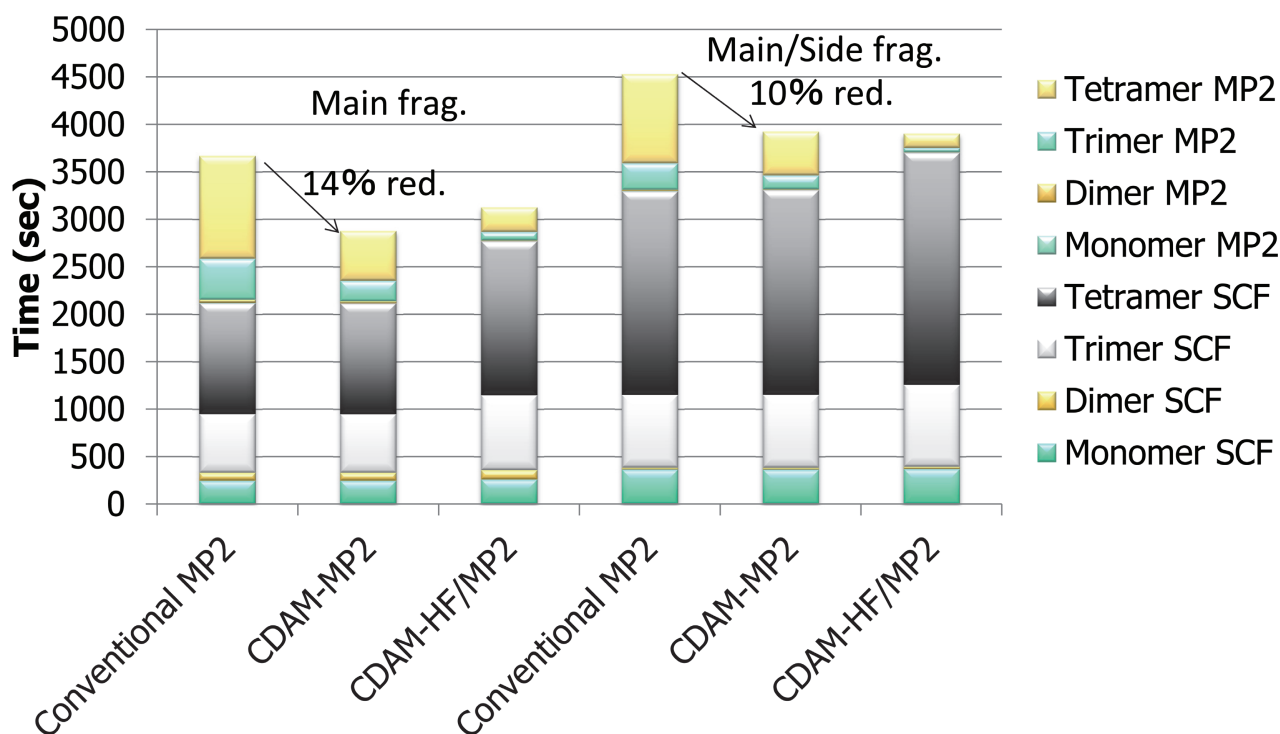
Fig. 1 Breakdown timing data for FMO4-MP2/6-31G calculations of HIV-1 protease complex. The computational times on ES2 with and without the CDAM (Cholesky decomposition with adaptive metric) [25] are compared in both the cases of main chain and main/side chain fragmentations.

or more. Particularly, FMO4 worked well for a nonconventional fragmentation procedure of peptides in which the main and side chains of amino acid residues were segmented. The HIV-1 protease - lopinavir complex as a practical example was calculated at the FMO4-MP2/6-31G level, by using ES2. The incremental cost of FMO4 relative to FMO2 was observed to be about ten times for this archetypical example of protein-ligand complex, while it would be justified by considering the utility of FMO4 in fragment-based drug discovery and design (FBDD) [16, 17]. The use of massively parallel computers is recommended for FMO4 calculations. Work to perform extensive IFIE analyses [2-4, 23] is underway for the HIV-1 protease complex as well as the estrogen receptor complex in the FBDD context [24]. Considering the recent developments of linear-scaling methods, the FMO scheme should be improved to provide better total energies for future benchmark comparisons. The FMO4 method is a promising approach in this direction.

References

[1] K. Kitaura, E. Ikeo, T. Asada, T. Nakano, and M. Uebayasi, Chem. Phys. Lett. 313 (1999) 701.

[2] D. G. Fedorov and K. Kitaura, J. Phys. Chem. A 111 (2007) 6904.

[3] D. G. Fedorov and K. Kitaura, ed., *The Fragment Molecular Orbital Method: Practical Applications to Large Molecular Systems*, CRC Press, Boca Raton, 2009.

[4] M. S. Gordon, D. G. Fedorov, S. R. Pruitt, and L. V. Slipchenko, Chem. Rev. 112 (2012) 632.

[5] D. G. Fedorov and K. Kitaura, J. Chem. Phys. 120 (2004) 6832.

[6] D. G. Fedorov and K. Kitaura, Chem. Phys. Lett. 433 (2006) 182.

[7] D. G. Fedorov, K. Ishimura, K. Ishida, K. Kitaura, P. Pulay, and S. Nagase, J. Comp. Chem. 28 (2007) 1476.

[8] M. Katouda, Theor. Chem. Acc. 130 (2011) 449.

[9] T. Fujita, K. Fukuzawa, Y. Mochizuki, T. Nakano, and S. Tanaka, Chem. Phys. Lett. 478 (2009) 295.

[10] Y. Mochizuki, T. Nakano, Y. Komeiji, K. Yamashita, Y. Okiyama, H. Yoshikawa, and H. Yamataka, Chem. Phys. Lett. 504 (2011) 95.

[11] T. Nakano, Y. Mochizuki, K. Yamashita, C. Watanabe, K. Fukuzawa, K. Segawa, Y. Okiyama, T. Tsukamoto, and S. Tanaka, Chem. Phys. Lett. 523 (2012) 128.

[12] B. Paulus, P. Fulde, and H. Stoll, Phys. Rev. B 51 (1995) 10572.

[13] J. Friedrich, M. Hanrath, and M. Dolg, J. Chem. Phys. 126 (2007) 154110.

[14] H. M. Netzloff and M. A. Collins, J. Chem. Phys. 127 (2007) 134113.

[15] L. Huang, L. Massa, and J. Karle, Proc. Nat. Acad. Sc. 105 (2008) 1849.

[16] R. Law, O. Barker, J. J. Barker, T. Hesterkamp, R. Godemann, O. Andersen, T. Fryatt, S. Courtney, D. Hallett, and M. Whittaker, J Comp. Aided Mol. Des. 23 (2009) 459.

[17] M. Whittaker, R. J. Law, O. Ichihara, T. Hesterkamp, and D. Hallett, Drug Discov. Today 7 (2010) e163.

[18] A. Yoshioka, K. Takematsu, I. Kurisaki, K. Fukuzawa, Y. Mochizuki, T. Nakano, E. Nobusawa, K. Nakajima, and S. Tanaka, Theor. Chem. Acc. 130 (2011) 1197.

[19] Y. Mochizuki, S. Koikegami, T. Nakano, S. Amari, and K. Kitaura, Chem. Phys. Lett. 396 (2004) 473.

[20] T. Nakano, T. Kaminuma, T. Sato, K. Fukuzawa, Y. Akiyama, M. Uebayasi, and K. Kitaura, Chem. Phys. Lett. 351 (2002) 475.

[21] H. Stoll, Chem. Phys. Lett. 191 (1992) 548.

[22] Y. Mochizuki, K. Yamashita, T. Murase, T. Nakano, K. Fukuzawa, K. Takematsu, H. Watanabe, and S. Tanaka, Chem. Phys. Lett. 457 (2008) 396.

[23] S. Amari, M. Aizawa, J. Zhang, K. Fukuzawa, Y. Mochizuki, I. Iwasawa, K. Nakata, H. Chuman, and T. Nakano, J. Chem. Inf. Model. 46 (2006) 221.

[24] C. Watanabe, K. Fukuzawa, K. Yamashita, Y. Okiyama, T. Tsukamoto, T. Murase, Y. Mochizuki, S. Tanaka, and T. Nakano, in preparation.

[25] Y. Okiyama, T. Nakano, K. Yamashita, Y. Mochizuki, N. Taguchi, and S. Tanaka, Chem. Phys. Lett. 490 (2010) 84.

フラグメント分子軌道法による薬剤設計のためのタンパク質－リガンド系に対する大規模電子状態計算

プロジェクト責任者

田中　成典　　神戸大学　大学院システム情報学研究科

著者

田中　成典　　神戸大学　大学院システム情報学研究科

望月　祐志　　立教大学　理学部化学科

山下　勝美　　NEC ソフト　ニューソリューション事業部

渡邉　千鶴　　東京大学　生産技術研究所

沖山　佳生　　東京大学　生産技術研究所

福澤　　薫　　みずほ情報総研　サイエンスソリューション部

中野　達也　　国立医薬品食品衛生研究所　医薬安全科学部

　フラグメント分子軌道（Fragment Molecular Orbital; FMO）法に基づき、タンパク質とリガンド分子結合系の大規模電子状態計算を地球シミュレータ（ES2）を用いて行った。フラグメントの 4 体項までを考慮する FMO4 法が最近開発され、Møller-Plesset の 2 次摂動（MP2）レベルでの計算が ABINIT-MPX プログラムに実装されている。従来までの FMO2 および FMO3 法による計算と比較したところ、エネルギー精度の顕著な改善が見られた。例として、198 残基の HIV-1 プロテアーゼとロピナビルの複合体を用い、従来のアミノ酸主鎖分割に加えて、主鎖・側鎖分割、さらにはリガンド分子の分割を試みたところ、FMO4 法を用いることで、計算精度を落とすことなく以前より細かいフラグメント分割が可能となることが判明した。今後、FMO4 法を用いた超並列計算が Fragment Based Drug Design などによる合理的薬剤設計の重要なツールとなることが期待される。

キーワード：フラグメント分子軌道法, メラー・プレセット摂動法, 4 体フラグメント補正, HIV プロテアーゼ