Large-Scale Electronic Structure Calculations of Biomolecular and Related Systems by the 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
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

We have performed large-scale fragment molecular orbital (FMO) calculations for biomolecular and related systems on the Earth Simulator. The four-body corrected fragment molecular orbital (FMO4) method was developed and implemented at the second-order Møller-Plesset perturbation (MP2) level with the use of the Cholesky decomposition with adaptive metric (CDAM) technique. A series of feasibility studies 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 obatained by conventional molecular orbital calculations. The usefulness of novel, refined fragmentation scheme was then examined for a number of molecular and condensed systems including inorganic interfacial systems, where the four-body corrections were shown to be substantial.

Keywords: Fragment molecular orbital (FMO) method, Møller-Plesset (MP) perturbation theory, Four-body correction, Cholesky decomposition with adaptive metric (CDAM)

1. Introduction

The fragment molecular orbital (FMO) method [1-3] has enabled one to perform fully quantum-mechanical, electronic-state calculations for large molecular systems like proteins at affordable cost of computation in a parallelized way. 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. On the other hand, with the ABINIT-MPX package, Fujita et al. [8] 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 [9].

Here, we report the recent development of four-body FMO (FMO4) calculations at HF and MP2 levels, that is, FMO4-HF and FMO4-MP2, on the ABINIT-MPX package [10]. The four-body corrections were already proposed and tested in the literature [11-14]. These studies showed certain improvements by the four-body treatment over the three-body one, in particular for the calculations of solids [11,13]. The present motivation to develop the FMO4 method, on the other hand, arises from the interest in more detailed modeling for the fragment-based drug discovery or design (FBDD) [15,16]. 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 [17]. We also apply the FMO4 method to the investigation of the interaction between adsorbed molecules and silica surface modeled by a large cluster containing many silicon atoms. The MP2 calculation is then accelerated by the Cholesky decomposition with adaptive metric (CDAM) technique. Systematic analyses are made for inter-fragment interaction energies (IFIEs) with and without a statistical correction for screening. The 1024 or 512 processors of the Earth Simulator (ES2) as a currently available platform of massively parallelized computation are used for a number of benchmark calculations in the present research.

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:

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

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

where N is the number of fragments in a given system and IJ are the fragment 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. [18] 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 [5,6,8] as

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

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

may be regarded as a next-order form of many-body expansion [19] 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 [8]. Care should be taken for the application of ESP approximations [18] to the FMO3-HF calculations, as addressed in Refs. [6,8]. The FMO3-MP2 correction [7,9] 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. [13] for the modeling of solid systems and also in Ref. [14] for proteins. The form of FMO4-HF energy is essentially the same as that of many-body expansion series:

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

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

where the actual tetramer HF calculation is performed similarly to the trimer case. 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 [15,16], 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 save the computation time in large-scale applications of FMO4-MP2.

In addition to protein-ligand systems, silica clusters were fragmented with an orbital-projection technique of bond-detachment atom (BDA) for silicon [20]. The fundamental unit of fragmentation was Si_3O_6, with or without hydrogen terminations. The assignment of formal charges in the respective fragments having 3D networks of Si-O bonds was crucial, where four BDAs and complementary bond-attachment atoms (BAAs) were set in Si_3O_6 in the charge-neutral fragmentation [21]. Further details of the fragmentation scheme for silicon systems will be published elsewhere.

For highly polarized molecular systems such as proteins (with or without hydration) and solids with bond polarity, mutual screening among involved components might be substantial. If such an effect is considered for the inter-fragment interactions, the potential overestimation from ionic or polar contributions could be remedied. For hydrated proteins, Ref. [22] reported a correction scheme for IFIEs based on the polarizable continuum model. Alternatively, we devised a posteriori recipe in which only a given set of IFIE values are required and then an entropic effect on screening among distributed fragments is taken into account in a statistical manner as follows [23]. This statistically corrected IFIE (abbreviated as SCIFIE) w_{ij} is related to the pair correlation function h_{ij} between the fragments as

$$h_{ij} = e^{-\beta w_{ij}} - 1.$$

where β is an inverse temperature parameter to be optimized, which is not directly correlated with room temperature. The Ornstein–Zernike relation associates the h_{ij} with the direct correlation function c_{ij} as

$$h_{ij} = c_{ij} + \sum_{k \neq i,j} c_{ik} h_{kj}.$$

and the Percus–Yevick (PY) approximation for classical many-body problem,

$$c_{ij} = e^{-\beta w_{ij}} - e^{-\beta(w_{ij} - u_{ij})},$$

is employed, providing a closure equation to determine w_{ij} for a given set of u_{ij}. Full descriptions of SCIFIE were given in Ref. [23]. The PY-based SCIFIEs are thus computed from the list of u_{ij}^{FMO4} [24], which can be compared with the regular ones.

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. [25]. 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. 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 [18] (in which the two-electron integrals were computed, unlike the classical approximation of ESP-PTC with Mulliken charges) was adopted for this protease complex. Further, in addition to HIV-1 protease complex, ER (estrogen receptor)-estradiol and NA (neuraminidase)-oseltamivir complex systems were also employed for benchmark calculations.

For the HF and MP2-corrected energies of the HIV-1 protease complex, the FMO4 results were used as a tentative reference

Table 1 Timing and performance data of FMO calculations for some protein-ligand complex systems. HIV: Human Immunodeficiency Virus; ER: estrogen receptor; NA: neuraminidase.

System (PDB-ID)	Method	Node	Time (sec)	Vectorization (%)	GFLOPS	Efficiency (%)
HIV-1 (1MUI) Main-chain division; Ligand division	FMO4-HF/6-31G	128	2576.562	95.800	2201.350	2.10
	FMO4-MP2/6-31G	128	4392.643	97.781	6902.978	6.58
HIV-1 (1MUI) Main-chain & side-chain division; Ligand division	FMO2-HF/6-31G	128	649.144	94.611	1394.082	1.33
	FMO2-MP2/6-31G	128	674.473	94.868	1537.004	1.47
	FMO3-HF/6-31G	128	1425.098	95.632	2114.378	2.02
	FMO3-MP2/6-31G	128	1720.652	96.632	2981.944	2.84
	FMO4-HF/6-31G	128	5254.855	89.718	1726.794	1.65
	FMO4-MP2/6-31G	128	6579.866	93.855	3257.717	3.11
ER (1ERE) Main-chain division; No ligand division	FMO4-MP2/6-31G	128	6371.446	97.606	6514.472	6.21
ER (1ERE) Main-chain division; Ligand division	FMO4-MP2/6-31G	128	5907.218	97.533	6595.924	6.29
ER (1ERE) Main-chain & side-chain division; No ligand division	FMO4-MP2/6-31G	128	12748.282	91.099	2625.183	2.50
ER (1ERE) Main-chain & side-chain division; Ligand division	FMO4-MP2/6-31G	128	11919.590	90.447	2550.045	2.43
NA (2HU4) Main-chain division; No ligand division	FMO3-MP2/6-31G	64	9300.139	97.419	2708.761	5.17
NA (2HU4) Main-chain & side-chain division; Ligand division	FMO3-MP2/6-31G	128	7101.818	94.675	2351.043	2.24

since the regular MO calculations of this sized molecule were impossible with the ABINIT-MPX program; we regarded the energy with conventional main-chain fragmentation as the best effort value. An unacceptable difference was then 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) [24,26]).

The timing and performance data for the benchmark calculations are shown in Table 1. It is notable that the incremental cost of MP2 over HF is maintained quite small for the calculations of monomers and dimers [25]. 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. 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 [18]. In addition, the computational time can be reduced with the use of the Cholesky decomposition technique [27], as seen in Table 2, in which the results for SiO_2-NaCl system are shown as well.

Although we here refrain from the presentation of IFIE results of these complex systems [24], the enhanced resolution of analyses matches with the FBDD scheme including the lead search and optimization [15,16]. 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. Further, we have found that the use of SCIFIE makes the calculated values of effective inter-fragment interactions more amenable to experimental observations [21,23].

In this report, we have addressed the development of the four-body FMO (FMO4) scheme [10,21,24]. 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 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. Some benchmark calculations on ES2 were performed at the FMO4-MP2/6-31G level with the aid of CDAM method [27]. The incremental cost of FMO4 relative to FMO2 was observed to be about ten times for archetypical examples of protein-ligand complexes, while it would be justified by considering the utility of FMO4 in the FBDD [15,16]. The use of massively parallel computers is recommended for FMO4 calculations. The FMO4 method is thus a promising approach in this direction. Considering the recent developments of linear-scaling methods, the FMO scheme would be improved to provide better total energies for future benchmark comparisons. In addition, more relevant information on effective inter-fragment interactions could be obtained in terms of SCIFIEs [21,23].

Table 2 Timing and performance data of FMO calculations with CDAM for some complex systems.

System	Method	Node	Time (hour)	Vectorization (%)	GFLOPS	Efficiency (%)
HIV-1 (1MUI) Main-chain & side-chain division; Ligand division	FMO4-CDAM-MP2/6-31G	128	1.6	97.639	5482.573	5.23
HIV-1 (1MUI) Main-chain & side-chain divsion; Ligand division	FMO4-CDAM-MP2/6-31G*	128	4.8	98.532	10097.681	9.63
NA (2HU4) Main-chain & side-chain division; Ligand division	FMO4-CDAM-MP2/6-31G	64	11.6	97.226	1998.252	3.81
NA (3CL0) Main-chain & side-chain division; Ligand division	FMO4-CDAM-MP2/6-31G	64	11.6	97.117	2084.437	3.98
SiO_2-Na^+	FMO4-CDAM-MP2/6-31G	64	6.1	99.050	9581.024	18.27
SiO_2-Cl^-	FMO4-CDAM-MP2/6-31G	64	5.9	99.059	9766.691	18.63
SiO_2-Cl^-	FMO4-CDAM-MP2/6-31G	128	3.2	99.041	17866.554	17.04
SiO_2-Na^+-Cl^-	FMO4-CDAM-MP2/6-31G	128	3.5	99.007	16654.045	15.88

Acknowledgements

This work was supported by the RISS project at the Institute of Industrial Science (IIS) of the University of Tokyo. YM and KF acknowledge the SFR-aid by Rikkyo University.

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

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

[10] 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.

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

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

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

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

[15] 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.

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

[17] 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.

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

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

[20] T. Ishikawa, Y. Mochizuki, K. Imamura, T. Nakano, H. Mori, H. Tokiwa, K. Tanaka, E. Miyoshi, and S. Tanaka, Chem. Phys. Lett. 430 (2006) 361.

[21] Y. Okiyama, T. Tsukamoto, C. Watanabe, K. Fukuzawa, S. Tanaka, and Y. Mochizuki, Chem. Phys. Lett. 566 (2013) 25.

[22] D. G. Fedorov and K. Kitaura, J. Phys. Chem. A 116 (2012) 704.

[23] S. Tanaka, C. Watanabe, and Y. Okiyama, Chem. Phys. Lett. 556 (2013) 272.

[24] C. Watanabe, K. Fukuzawa, Y. Okiyama, T. Tsukamoto, A. Kato, S. Tanaka, Y. Mochizuki, and T. Nakano, J. Mol. Graph. Model. 41 (2013) 31.

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

[26] 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.

[27] 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生体分子系・ナノ界面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
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ならびにナノ界面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法に基づく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、HIV-1プロテアーゼ、エストロゲン受容体、ノイラミニダーゼとリガンド分子の複合体を用い、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。また、シリカ表面に吸着したペプチドや小分子、イオンの系に対しても、その有効性が示された。

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, コレスキー分解