Application of order-N first-principles DFT calculations with temperature controlled molecular dynamics to biomolecular system

Takao Otsuka1 and Makoto Taiji1

1 Laboratory for Computational Molecular Design, Quantitative Biology Center (QBic), RIKEN, QBic Building B, 6-2-4 Furuedai, Suita, Osaka 565-0874, Japan

e-mail: totsuka@riken.jp

Abstract. We have performed large-scale first-principle molecular dynamics (FPMD) calculations of the large-scale hydrated DNA system as an example of large-scale biomolecular system, using order-N density functional theory (DFT) code CONQUEST, which is based on the combined method with an order-N DFT method using the density matrix minimization scheme and the extended Lagrangian Born-Oppenheimer molecular dynamics (XL-BOMD) scheme. Our large-scale FPMD calculations by CONQUEST show stable FPMD simulation and good energy conservation in NVE ensemble. We also show temperature controlled MD simulations such as velocity scaling, stochastic velocity scaling, and Nosé-Hoover chain methods. Our protocols using velocity scaling, stochastic velocity rescaling, and Nosé-Hoover chain methods are very effective for temperature control in early stage of order-N FPMD simulations. Our order-N FPMD simulation would be able to realize the large-scale biomolecular simulations.

1. Introduction

Molecular simulation technology is now commonly used to explore biological functions of biomolecular systems. Especially, molecular simulation methodologies are expected to help us understand the mechanism of biological phenomena with dynamical molecular motion, for example, such as enzyme reaction, photoexcitation, or molecular interaction and so on. In these biological phenomena, electron plays an important role so that it needs molecular dynamics (MD) simulation based on quantum mechanics. Recently, quantum mechanics and molecular mechanics (QM/MM) hybrid methods or its combined molecular dynamics (QM/MM-MD) method are also used well in this field in order to elucidate the biological functions and mechanism [1]. However, there are still some problems in those QMMM hybrid method; for example, the artificial boundary problem, which is usually constrained with the artificial boundary between QM and MM regions, and the limitation of the size of QM region, which is also restricted due to the computational costs of QM calculation. Therefore, there are increasing demands for all-atom QM simulations on complex biological systems.

In recent progress for the QM calculations, an order-N (or linear-scaling) methodology for first-principle density functional theory (DFT) have been developed [2]. We have also developed order-N DFT code, CONQUEST, which has high parallel efficiency and enable us to perform DFT studies on very large-scale systems [2,3]. Recently, CONQUEST have been introduced very stable and efficient order-N first-principle MD (FPMD) method combined with the extended Lagrangian Born-
Oppenheimer MD scheme proposed by Niklasson [4,5] and have shown the reliable and efficient FPMD on large-scale systems [6,7].

In practical application of large-scale FPMD to large-scale biomolecular systems, in many case, the initial structure is still prepared by using classical MD simulations. The equilibrium structure obtained by classical MD simulation is set as ones of large-scale FPMD simulation. In such case, especially huge-scale biomolecular systems, it is very important to achieve the setting temperature rapidly, without taking too much real computational time in large-scale FPMD simulations. In this study, we report further progress and preliminary results for the temperature controlled FPMD simulations on the order-N FPMD code in biomolecular system. As the temperature controlled method, we used velocity scaling method, stochastic velocity rescaling method [25] and Nosé-Hoover chain method [26-28] in NVT ensemble. Recently, two method of temperature control such as the velocity scaling and the Nosé-Hoover chain method were implemented in CONQUEST code [7,8]. In this study, we implemented the stochastic velocity rescaling method in CONQUEST code. Although we have already shown the effective results of the velocity scaling method [7] in biomolecular system, here we applied the stochastic velocity rescaling method and the Nosé-Hoover chain method in CONQUEST code to biomolecular system. Especially, in this study, we focused on early stage of large-scale FPMD steps to investigate the initial behavior of temperature in order-N FPMD simulations. We also investigated the combined protocol of velocity scaling and Nosé-Hoover chain method. We used a test DNA model, which contains a decamer DNA base pairs, Mg counter ion, and a large number of water molecules, consisting of about 3400 atoms. We demonstrate that the FPMD simulations on hydrated DNA system are robust and accurate even in NVT ensemble.

2. Theoretical background
Here we briefly describe some key points of the order-N DFT method and the combined XL-BOMD with temperature controlled method. The details of the order-N methodology used in the CONQUEST code are explained in previous works [8-10].

In CONQUEST, we use the Kohn-Sham density matrix defined as

\[ \rho(r, r') = \sum_n f_n \psi_n(r) \psi_n^*(r'), \]

where \( \psi_n(r) \) is the Kohn-Sham orbital for band index \( n \), and \( f_n \) is its occupation number between 0 and 1. The DFT total energy of system can be calculated from density matrix, with the use of the pseudopotential technique and the standard exchange-correlation functionals such as the local density approximation (LDA) or the generalized gradient approximation (GGA). The density matrix is represented by localized orbitals called as “support functions”

\[ \rho(r, r') = \sum_{iaj\beta} \phi_{ia}(r) K_{ia,j\beta} \phi_{j\beta}(r'), \]

where \( \phi_{ia}(r) \) are the support function that are non-zero only inside “support region” centered on the atoms \( i \), and \( a \) runs over the support functions on a given atom. The matrix \( K_{ia,j\beta} \) is the density matrix expressed by non-orthogonal basis of support functions, that is,

\[ K = 3LSL - 2LSLSL, \]

where the matrix \( L \) and \( S \) are an auxiliary density matrix and the support function overlap matrix, respectively. In CONQUEST, we can use two types of basis function; one is B-splines on regular grids [11] and another is numerical pseudo atomic orbitals (PAOs) [12-14]. For the PAO treatment, recently,
an efficient and accurate scheme for large size basis set have become available in CONQUEST, called the multisite support function (MSSF) method [15,16]. This scheme can perform accurate calculations without increasing the CPU time significantly. The density matrix $K_{ia, jb}$ using the support functions is obtained by the conventional diagonalization of Hamiltonian matrix or the order-N method based on the combination of McWeeny’s purification transformation [17] and the density matrix minimization (DMM) scheme by Li, Nunes, and Vanderbilt [18]. In the practical linear-scaling computations, we use DMM method introducing the region cut-off for the matrix elements [19,20].

In the MD simulation using CONQUEST, the XL-BOMD scheme proposed by Niklasson [4,5] is introduced into CONQUEST code [6]. The XL-BOMD scheme can maintain the time-reversal symmetry while keeping the efficiency of density matrix reuse. The Born-Oppenheimer Lagrangian in the XL-BOMD scheme is expressed by

$$L^{XBO}(X, X, R, \tilde{R}) = L^{BO}(R, \tilde{R}) + \frac{\mu}{2} \text{Tr}[X^2] - \frac{\hbar \omega}{2} \text{Tr}[(LS - X)^2],$$

$$L^{BO}(R, \tilde{R}) = \frac{1}{2} \sum_{i=1}^{N} M_i \dot{R}_i^2 + E_{tot}(R_i),$$

where the potential energy $E_{tot}$ is defined by the self-consistent ground state density matrix at nuclear positions $R_i$ with masses $M_i$. The matrix $X$ is a sparse matrix with the range of the matrix $LS$, $\mu$ is the fictitious electronic mass, and $\omega$ is the curvature of the electronic harmonic potential. As shown in previous study [6], we use the $LS$ matrix as an auxiliary degree of freedom in the Born-Oppenheimer Lagrangian and the time-reversible Verlet scheme to obtain the matrix $X$. The matrix $X$ is time-reversible and evolves in a harmonic potential centered on the ground-state $LS$. In the practical numerical propagation of the matrix $X$, the equation of motion with a dissipative force is used to maintain the numerical stability of the matrix $X$ (see the details in ref. [6]).

3. Computational details

As a demonstration model in this study, we used the large-scale hydrated DNA system prepared by our previous study [7,22], which consists of DNA 10 base pairs (d(CCATTATG))$_2$ by PDB ID: 1WQZ) of 634 atoms, Mg$^{2+}$ ion of 9 atoms, and H$_2$O of 932 molecules, the total number of atoms in DNA system is 3,439 atoms, as shown in Figure 1. The initial structure was prepared by using the AMBER9 [23] with the force fields of PARM99 for the DNA atoms TIP3P for water molecules. The system was equilibrated with the constant-pressure MD and then the structure at last step was adopted. For the details of preparation of our DNA system, see our previous related work [22]. The linear-scaling FPMD simulations were performed by using our linear-scaling DFT code CONQUEST with PAO of single-zeta with polarization (SZP) basis set, Perdew-Burke-Ernzerhof (PBE) exchange-correlation functional [24], the DMM with self-consistent field (SCF) calculations, and the numerical integration grid cutoff of 75 Ha. For the setting of XL-BOMD conditions, we used the time-reversible Verlet scheme with a dissipative force correction of the order of 5 [4,5]. The MD time step and the initial temperature for the each atomic velocity were set as 0.5 fs and 300 K for NVE ensemble. For NVT ensemble, we performed three kinds of the temperature controlled MD simulations: MD simulations using the velocity scaling method with the temperature of 300 K and 600K, the stochastic velocity rescaling method with the temperature of 300K, and the Nosé-Hoover chain method with the temperatures of 300 K. As following the previous study [8], we used 5 chained thermostats, the 15th-order Yoshida-Suzuki integrator and the thermostat frequency of 500 cm$^{-1}$ in the Nosé-Hoover chain method. The fictitious thermostat mass of first thermostat is different from those of the rest (see the ref. [8]). In this study, although we limited the computational conditions for DFT calculations (PBE/SZP without dispersion correction) in order to know the energy behaviour of the thermostats, we can
improve the computational accuracy using such as the MSSF method [15,16] with higher quality of PAO basis set and the DFT-D dispersion correction [29]. In Appendix, we also showed the performance of three thermostats in CONQUEST code using silicon carbide system (see below).

Figure 1. Structure of hydrated DNA (PDB ID: 1WQZ).

4. Results and Discussion
First, we demonstrate our order-N FPMD simulations of hydrated DNA system in NVE ensemble. Figure 2 shows the energy profile obtained by the order-N FPMD simulations. In the early stage of the order-N FPMD simulation until around 50 MD steps, the potential and the kinetic energies of the hydrated DNA system fluctuate considerably. This fluctuations show the difference of the force fields between classical MD simulation and the order-N FPMD simulation. Around 100 MD steps or later, the hydrated DNA structure is relaxed sufficiently by the order-N FPMD simulation, and then towards the equilibration. The structure relaxation and equilibration process of the large-scale biomolecular system does not take many MD steps in the linear-scaling FPMD simulation. It is also seen that the order-N FPMD simulation shows good energy conservation and stable MD simulation in NVE ensemble.

Figure 2. Energy profile calculated by order-N FPMD simulation; total energy (green), potential energy (blue), and kinetic energy (red).

Here we investigate the temperature controlled FPMD simulations by using the velocity scaling, stochastic velocity rescaling, and Nosé-Hoover chain method. In this study, we focused on the temperature behavior in early stage of order-N FPMD simulation. Figure 3 shows the energy profile
obtained by the order-N FPMD simulation with the velocity scaling method. We set the controlled temperature as 300 K (Fig. 3(a)) and 600 K (Fig. 3(b)), respectively. In the velocity scaling calculations, we performed the velocity correction for each step. As seen in Figure 3, although the fluctuation of the kinetic energy is large in the early MD steps (that is, 0-50 MD steps), then the kinetic energy converges to around the setting temperature. We can expect that this rapid convergence in the early steps of FPMD simulation would be very useful.

Figure 3. Energy profile calculated by order-N FPMD simulation with velocity scaling method as (a) 300 K and (b) 600 K; total energy (green), potential energy (blue), and kinetic energy of 300 K (pink) and of 600 K (red).

Next Figure 4 shows the energy profile obtained by the stochastic velocity rescaling method. In the stochastic velocity rescaling method, we set the temperature as 300 K and also performed the velocity correction for each step. From Fig. 4, we can see that the behavior of the energy profile is as almost same as the velocity scaling method as 300 K (Fig. 3(a)). Table 1 shows the average and the variance of temperatures by the velocity scaling and the stochastic velocity rescaling methods within 100-300 MD steps. As see in Table 1, the two average temperatures show almost 300 K. Although the variance of temperature by the stochastic velocity rescaling method is larger than that by the velocity scaling method, it is very interesting that the energy profile of DNA system in early stage of FPMD simulation is similar behavior, this behavior is different from the case of small SiC system (see Appendix). We do not clearly understand why we have observed such a behavior in DNA system (large-scale biomolecular system), but we can expect that the use of the stochastic velocity rescaling method, which can sample the canonical distribution, would help to analyze the dynamics of a large-scale biomolecular system on FPMD simulations.

Figure 4. Energy profile calculated by order-N FPMD simulation with stochastic velocity rescaling method as 300 K for 0-300 MD steps (total energy (green open circle), potential energy (blue open circle), and kinetic energy of 300 K (pink open circle)).
Table 1. Average values and the variances of temperatures by velocity scaling and stochastic velocity rescaling methods within 100-300 MD steps.

| Method                     | Average | Variance |
|----------------------------|---------|----------|
| Velocity scaling           | 301.17 K| 1.15     |
| Stochastic velocity rescaling| 300.76 K| 28.02    |

Figure 5 shows the energy profile obtained by the Nosé-Hoover chain method. From Fig. 5 we can see that the temperature control by the Nosé-Hoover chain method in the early MD steps of FPMD simulations is relatively slow convergence in present conditions, unlike that of the velocity scaling method. This slow convergence of velocity by the Nosé-Hoover chain method takes many MD steps in FPMD simulations. In order to find the clear reason for this slow convergence of the temperature in the Nosé-Hoover chain method, we think that the evaluation of the calculation conditions (the number of thermostats, the fictitious thermostat masses and the system size) for biomolecular system in Nosé-Hoover chain method is necessary. In this study, we just applied the general parameter reported in previous study [8], hence we cannot describe the further discussion.

![Energy profile](image)

**Figure 5.** Energy profile calculated by order-N FPMD simulation with Nosé-Hoover chain method as 300 K for 0-400 MD steps (total energy (yellow green), potential energy (aqua blue), and kinetic energy of 300 K (bright pink)).

On the other hand, as one of the possible improvements of this slow convergence of the temperature control in the Nosé-Hoover chain method, we tried combining the velocity scaling method with the Nosé-Hoover chain methods. In early MD steps in order-N FPMD simulations, we used the velocity scaling method, and then we switched to the Nosé-Hoover chain method. Figure 5 show the energy profile combined with these two method. In this energy profile, we set the velocity scaling method from 0 to 100 MD steps, and then the Nosé-Hoover chain method. From Figure 6, the kinetic energy in the Nosé-Hoover chain method at around 200 MD steps and later converged to the setting temperature of 300 K. We can expect that this combined temperature controlled protocol in NVT ensemble should be very effective in the early steps of the temperature controlled FPMD simulations.
5. Conclusion

In this study, we have performed the order-N FPMD simulation of hydrated DNA system using order-N DFT code CONQUEST. First, we have demonstrated the NVE ensemble of hydrated DNA system and have found that the order-N FPMD simulation show the rapid equilibration using initial structure obtained by classical MD simulation and also keeps the good energy conservation. We have also investigated the order-N FPMD simulation with the velocity scaling method, the stochastic velocity rescaling method, and the Nosé-Hoover chain method in NVT ensemble. Here we have just tested one case of calculation condition in the Nosé-Hoover chain method. In the initial stage of temperature control in our order-N FPMD simulations, we have found that the velocity scaling method was very effective to converge to the setting temperature. We also showed that the stochastic velocity rescaling method was also the similar behaviour of the velocity scaling method in DNA system. Apart from this, we can expect that the stochastic velocity rescaling method would be one of the possible choices on total FPMD simulations of large-scale biomolecular system, which can sample the canonical distribution. On the other hand, although we need to explore the further calculation conditions of the Nosé-Hoover chain method, the Nosé-Hoover chain method showed the slow convergence compared with the velocity scaling method. We also showed the combined protocol of these temperature control method. These combined temperature controlled protocol in NVT ensemble was very effective for the temperature control in early stage on the order-N FPMD simulations and for reducing the real computational time. We expect that the temperature controlled order-N FPMD simulations by CONQUEST code will be useful for understanding the large-scale biomolecular system. Further numerical tests for above points will be presented in a future publication.

Acknowledgements

This work was partly supported by the JSPS KAKENHI projects: Grant Numbers, 22790122, 16KT0168. The calculations were performed in part on the NIMS material numerical simulator and the COMA (PACS-IX) system of University of Tsukuba. The authors also acknowledge Dr. T. Miyazaki (National Institute for Materials Science) and Prof. D. R. Bowler (University College London) for critical comments and fruitful discussion, and Prof. M. J. Gillan (University College London), Dr. M. Arita (Tokyo University of Science), Dr. N. Okimoto (RIKEN), and Dr. H. Saito (RIKEN) for useful discussions.
Appendix: The behaviour of SiC system by three thermostats in CONQUEST code

Here, we show the order-N FPMD simulations on a silicon carbide (SiC) system using three methods such as the velocity scaling, the stochastic velocity rescaling, and the Nosé-Hoover chain method in NVT ensemble. The SiC system is consisted of 64 atoms. The simulation cell is cubic cell of length 8.716 Å with periodic boundary conditions. The PAO of single-zeta was used with an energy grid cutoff of 80 Ha. These conditions are almost same as the literature of [8]. A time step of 1.0 fs was used and the total MD steps was just 500 steps. The initial temperature was set as 300 K.

Figure A1 show the energy profiles of kinetic energy and the time-averaged temperatures obtained by the order-N FPMD simulations with three methods. From Fig. A1(a), we can see that the fluctuation of kinetic energy by the velocity scaling method is small because the kinetic energy is forced to be exactly equal to target energy (target temperature). While the fluctuation of kinetic energy by the stochastic velocity rescaling method is rather large, where the target energy (target temperature) is selected to be satisfy the canonical equilibrium distribution. As seen in Fig. A1(b), the time-averaged temperatures show the convergence to the target temperature of 300 K. On the other hand, the kinetic energy by the Nosé-Hoover chain method seems to fluctuate slowly around the target temperature.

Figure A1. (a) Energy profiles of kinetic energy and (b) Time-averaged temperatures of SiC system by order-N FPMD simulation with three thermostats; Nosé-Hoover chain (blue), stochastic velocity rescaling (red), and velocity scaling (green). The initial temperature is set as 300 K.

Figure A2 show the energy profile of total energy (that is, kinetic energy and potential energy) and the time-averaged total energy obtained by the order-N FPMD simulations with three methods. As expected from the kinetic energy profile, although the fluctuation of total energy by the stochastic velocity scaling method is large, the time-averaged total energy is as almost same behavior as that of the velocity scaling method.

Figure A2. (a) Energy profiles of total energy and (b) Time-averaged total energy of SiC system by order-N FPMD simulation with three thermostats; Nosé-Hoover chain (blue), stochastic velocity rescaling (red), and velocity scaling (green). The initial temperature is set as 300 K.
References

[1] For example, Novel prize of chemistry 2013, Martin Karplus, Michael Levitt, Arieh Warshel, “for the development of multiscale models for complex chemical systems”.

[2] D. R. Bowler and T. Miyazaki, Rep. Prog. Phys. 75, 036503 (2012).

[3] See the web page; http://www.order-n.org/.

[4] A. M. N. Niklasson, C. J. Tymczak, M. Challacombe, Phys. Rev. Lett. 97, No. 123001 (2006).

[5] A. M. N. Niklasson, Phys. Rev. Lett., 100, No. 123004 (2008).

[6] M. Arita, D. R. Bowler, T. Miyazaki, J. Chem. Theory Comput., 10, 5419–5425 (2014).

[7] T. Otsuka, M. Taiji, D. R. Bowler, T. Miyazaki, Jpn. J. Appl. Phys. 55, 1102B1 (2016).

[8] T. Hirakawa, T. Suzuki, D. R. Bowler, T. Miyazaki, J. Phys.: Condens. Matter 29, 40591 (2017).

[9] E. Hernández, M. J. Gillan, C. M. Goringe, Phys. Rev. B, 53, 7147–7157 (1996).

[10] D. R. Bowler, T. Miyazaki, M. J. Gillan, J. Phys.: Condens. Matter 14, 2781–2798 (2002).

[11] D. R. Bowler, R. Choudhury, M. J. Gillan, T. Miyazaki, Phys. Status Solidi B, 243, 989–1000 (2006).

[12] E. Hernández, M. J. Gillan, and C. M. Goringe, Phys. Rev. B 55, 13485 (1997).

[13] O. F. Sankey and D. J. Niklewski, Phys. Rev. B 40, 3979 (1989).

[14] J. M. Soler, E. Artacho, J. D. Gale, A. García, J. Junquera, P. Ordejón, and D. Sánchez-Portal, J. Phys.: Condens. Matter 14, 2745 (2002).

[15] A. S. Torralba, M. Todorović, V. Brázdová, R. Choudhury, T. Miyazaki, M. J. Gillan, and D. R. Bowler, J. Phys.: Condens. Matter 20, 294206 (2008).

[16] A. Nakata, D. R. Bowler, and T. Miyazaki, J. Chem. Theory Comput., 10, 4813 (2014).

[17] A. Nakata, D. R. Bowler, and T. Miyazaki, Phys. Chem. Chem. Phys., 17, 31427 (2015).

[18] R. McWeeny, Rev. Mod. Phys. 32, 335 (1960).

[19] X. P. Li, R. W. Nunes, and D. Vanderbilt, Phys. Rev. B 47, 10891 (1993).

[20] D. R. Bowler and M. J. Gillan, Comput. Phys. Commun. 120, 95 (1999).

[21] D. R. Bowler and M. J. Gillan, Chem. Phys. Lett. 325, 473 (2000).

[22] T. Otsuka, T. Miyazaki, T. Ohno, D. R. Bowler, and M. J. Gillan, J. Phys.: Condens. Matter 20, 294201 (2008).

[23] Case D A et al. 2006 AMBER 9 University of California, San Francisco.

[24] J. P. Perdew, K. Burke, M. Ernzerhof, Phys. Rev. Lett., 77 3865 (1996), Errata: Phys. Rev. Lett., 78, 1396 (1997).

[25] G. Bussi, D. Donadio, and M. Parrinello, J. Chem. Phys. 126, 014101 (2007).

[26] S. Nosé, J. Chem. Phys. 81, 511 (1985).

[27] W. G. Hoover, Phys. Rev. A 31, 1695 (1985).

[28] G. J. Martyna, M. E. Tuckerman, D. J. Tobias, and M. L. Klein, Mol. Phys. 87, 1117 (1996).

[29] S. Grimme, J. Comp. Chem. 27, 1787 (2006).