Multiscale dynamical framework for high-level calculations of finite temperature ground and excited state properties

Karol Kowalski and Marat Valiev
Molecular Sciences Software Group, Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, WA 99352 USA
karol@kowalski@pnl.gov marat.valiev@pnl.gov

There is a pressing need for accurate calculations of finite temperature ground and excited state properties of nanoscale systems relevant to structural biology, hydrogen economy, environmental and material science problems. To address this challenging task we have designed a multiscale dynamical approach that combines the accuracy and computational complexity of coupled-cluster (CC) methods with the efficiency of classical molecular dynamics simulations. Our methodology is based on a seamless integration between the generic QM/MM interface, Tensor Contraction Engine module, and the classical molecular dynamics module of NWChem and offers an unprecedented ability for accurate large scale calculations of thermodynamics of ground and excited state properties. We illustrate our approach by large scale dynamical simulation of the excited state spectrum of the cytosine base in its native DNA environment using a variant of the completely renormalized equation-of-motion method with singles, doubles, and non-iterative triples (CR-EOMCCSD(T)).

1. Introduction
Theoretical methods for electronic structure calculations have made astounding strides, reaching a level of precision that, in many cases, approaches the chemical accuracy for the ground-state predictions. Over the last two decades, the coupled cluster approach (CC) has evolved into the dominant methodology for highly accurate electronic calculations in the gas phase[1,2]. Ground-state approaches such as the CC methods with singles, doubles, and non-iterative triples, CCSD[T] or CCSD(T) [3,4], have been successful in reducing computational costs associated with triply excited connected clusters necessary to attain chemical accuracy. This conspicuous success of the CC methodology was coupled with the development of efficient excited-state methodologies including one of the most established equation-of-motion CC (EOMCC) formalism[5,6]. In analogy to the ground-state situation, the non-iterative EOMCC approaches have several appealing features to a broad community of users including the “black-box” character of these methods and a relatively low computational cost. Traditionally CC methods have always been associated with gas phase model calculations. Yet we can anticipate that in many problems of practical interest the environment[7] and its dynamical fluctuations can significantly impact the ground- and excited-state properties. In this paper we discuss an efficient methodology for incorporating CC methods into large scale calculations of excited states[8] with a full account of finite temperature dynamical fluctuations of the environment. This multidisciplinary approach combines state-of-the-art ab-initio theory, classical molecular dynamics, and statistical resampling methods. The basic features of the approach are illustrated by the calculation of excited states of cytosine in its native DNA environment. The ultrafast
excited-state deactivation mechanism of cytosine, which is essential for the quenching of mutagenic photochemical reactions is still under debate [9].

2. Methodology

One of the greatest challenges in incorporating ab-initio methods in finite temperature simulations lies in the requirement to perform statistical averaging over the extended periods of time. This problem is resolved in our methodology using a non-Boltzmann resampling scheme [10] in conjunction with classical molecular dynamics and coupled cluster/molecular mechanics approaches. In this framework the dynamics of the entire system is driven by a lower level theory ($E_0$). The resulting trajectories are then processed with a higher level theory ($E$), albeit at a much lower rate, to calculate the actual observable ($\omega$) according to the following expression [10,11]:

$$\langle \omega \rangle = \frac{\langle \omega \rangle e^{\beta (E-E_0)} \rangle_0}{e^{\beta (E-E_0)}}.$$ 

In our particular implementation the lower level theory is based on an Amber-type molecular mechanics force field. The higher level theory is given by a combined coupled-cluster and molecular mechanics (CC/MM) approach described by the following Hamiltonian

$$H = H_{QM} + H_{QM/MM} + H_{MM},$$

Here $H_{QM}$ is a standard many-electron Hamiltonian describing the internal energy of the quantum mechanical (QM) region

$$H_{QM} = E_0^{(0)} + \sum_{\mu,\nu} f_{\mu,\nu} a_\mu^+ a_\nu + \frac{1}{2} \sum_{\mu,\nu,\delta,\kappa} v_{\delta,\kappa}^{\mu,\nu} a_\delta^+ a_\kappa a_\mu a_\nu,$$

where the indices $\nu, \mu, \lambda$, and $\kappa$ designate single-particle states, the elements $f_{\mu,\nu}^{\mu,\nu}$ and $v_{\delta,\kappa}^{\mu,\nu}$ represent one- and two-electron integrals, respectively, whereas $a_{\mu}^+$ ($a_{\mu}$) operators are the usual creation (annihilation) operators. The interaction between the QM region and its surroundings (MM region) is described by the second term $H_{QM/MM}$,

$$H_{QM/MM} = \sum_{\mu,\nu} \langle \mu | \frac{Q_n^{\mu}}{R_{\nu}^{\mu,n}} | \nu \rangle a_\mu^+ a_\nu + V(\{R^n\}_{\mu} \{R\}),$$

where $Q_n^{\mu}$ and $R_{\nu}^{\mu,n}$ denote charges and coordinates of the MM region. The $V(\{R^n\}_{\mu} \{R\})$ term represents the interaction between nuclei in MM and QM regions (\{R\} symbolically represents the set of nuclear coordinates of the QM region). The third term in Eq.(1), $H_{MM}$, describes the internal energy of the MM region which is similar to our lower level Hamiltonian ($H_0$) except that all interactions pertaining to only the QM region are eliminated. The Hamiltonian $\tilde{H}$, defined as

$$\tilde{H} = H_{QM} + H_{QM/MM},$$

effectively includes, through the $H_{QM/MM}$ term, the interaction of the environment with the QM region. The observable $\omega$ in our calculations refers to the vertical excitation energy calculated within the CC/EOMCC methodology as described below.

Our main motivation to pursue the non-iterative EOMCC approaches as ideal candidates to be combined with the MM part is their “black-box” character and moderate numerical cost. We will mainly focus on the recently introduced completely renormalized EOMCCSD(T) (CR-EOMCCSD(T)) approach that approximately accounts for the effect of triples. Here we recall only the salient features of this approach (for more detailed discussion see [12]). The CR-EOMCCSD(T) methods is a three-step procedure. In the first step we represent the ground-state wavefunction in the form of an exponential Ansatz that employs only singly- and doubly-excited cluster operators $T_2$ and $T_3$, respectively,
where the reference function \( |\Phi\rangle \) is customarily chosen as a Hartree-Fock Slater determinant.

The cluster operators \( T_1 \) and \( T_2 \) are defined as

\[
T_1 = \sum_{i,a} t_{ia}^a E_{ia}^a,
\]

\[
T_2 = \sum_{i,j,a,b} t_{ab}^{ij} E_{ab}^{ij},
\]

where the operators \( E_{ia}^a \) and \( E_{ab}^{ij} \) generate singly and doubly excited configurations \( |\Phi_{ia}^a\rangle \) and \( |\Phi_{ab}^{ij}\rangle \), respectively, when acting on the reference function. In all standard approaches the cluster amplitudes \( t_{ia}^a \) and \( t_{ab}^{ij} \) are obtained by solving the connected form of the Schroedinger equation projected onto a manifold of all singly and doubly excited determinants:

\[
\langle \Phi_{ia}^a | (\bar{H} - \bar{E}) | \Phi \rangle = 0,
\]

\[
\langle \Phi_{ab}^{ij} | (\bar{H} - \bar{E}) | \Phi \rangle = 0,
\]

where the subscript “C” designates the connected form of a given operator expression.

Next, having obtained the cluster operators from CCSD equations we represent the wavefunction of the \( K \)-th (\( K=1,2,\ldots \)) excited state, using the EOMCCSD formalism, as

\[
|\Psi_K\rangle = (R_{k,0} + R_{k,1} + R_{k,2}) e^{T_1 T_2} |\Phi\rangle.
\]

The excitation operators, \( R_{k,j} \), along with the corresponding excited-state energies are obtained by diagonalizing the similarity transformed Hamiltonian, \( \bar{H} = e^{-T_1 T_2} \bar{H} e^{T_1 T_2} \), in the space spanned by the reference function and all singly and doubly excited configurations. It is well documented that the EOMCCSD approach provides a good description of singly excited states with average departures from the experimental/exact (full CI) results within 0.3 eV. This picture no longer holds for doubly excited states. In such situations errors on the order of a few eV have been documented [13]. To alleviate these problems inclusion of triples is prerequisite. However, compared to the CCSD/EOMCCSD \( N^6 \) (\( N \) symbolically defines the system size) scaling, the \( N^8 \) cost of the CCSDT/EOMCCSDT approaches make these methods useless for large scale calculations. A good alternative to full EOMCCSDT is offered by non-iterative completely renormalized EOMCCSD(T) approaches which provide nearly EOMCCSDT level of accuracy for doubly excited states. We will pursue one variant of the CR-EOMCCSD(T) methods which directly corrects the EOMCCSD excitation energies by adding the correction due to triples \( \delta_k (IA) \) to excitation energy obtained in EOMCCSD calculations, \( \omega_{k,\text{EOMCCSD}} \), i.e.,

\[
\omega_{k,\text{CR-EOMCCSD(T)}} = \omega_{k,\text{EOMCCSD}} + \delta_k (IA),
\]

where the \( \delta_k (IA) \) correction is expressed as

\[
\delta_k (IA) = \left\langle \Psi_k (IA) | M_{k,3} (2) | \Phi \right\rangle / D_k.
\]

In the above formula the \( |\Psi_k (IA)\rangle \), \( M_{k,3} (2) \), and \( D_k \) represent the trial wavefunction, three-body moments of the EOMCCSD equations, and overlap between trial and EOMCCSD wavefunctions for the \( K \)-th state (for details See Ref.[12]). The \( IA \) index refers to the form of the denominators used in constructing the three-body cluster/excitation operators [12]. As demonstrated on several examples, corrections of this type work well for singly and mixed singly-doubly excited states.

All high-level calculations were performed using TCE [14] implementation in the NWChem[15]. Some changes have been made to improve the performance of the CCSD code and a few classes of diagrams have been fused together to reduce the need of multiple calls to the most expensive parts of the two-electron integrals. For example, by introducing the intermediate quantity \( C_2 = T_2 + \gamma T_1 \) the
4-particle 2-electron integrals that appear in the \( \langle \Phi_i^a \mid V_n \left( T_2 + \gamma T_1^2 \right) \mid \Phi \rangle \) terms are invoked only once. This also entails reduction of the memory requirements for recursive intermediates. Currently, the size of largest intermediate equals approximately the \( n_o n_u \) which should be compared to \( n_o n_u \) required by the previous version of the CCSD method (\( n_o \) and \( n_u \) refer symbolically to the number of occupied and unoccupied spin-orbitals respectively).

3. Results and Discussion

An excellent illustration of our method is provided by the calculation of the two lowest states of cytosine base in the native DNA environment. Since ultrafast internal conversion to the ground state and very short excited-state lifetime (subpicosecond) of cytosine serve as a protection against undesired UV induced photochemical reactions, first-principle studies along this line assume particular importance. Much of extensive theoretical work on this problem was based on ab-initio calculations in the gas-phase. However, model DFT studies of hydrated cytosine indicated that the effect of the aqueous solution may be quite sizable. In this light having a chance to include the effect of fluctuating environment in conjunction with state-of-the-art excited-state methods is a very unique opportunity. We used our formalism to calculate the thermal average of the \( n_o \pi^* \) and \( \pi\pi^* \) states of cytosine. The system considered in this calculation consisted of a 12-mer fragment of B-DNA (3'-TCGCGTTGCGCT-5') solvated in a rectangular box (51.51 x 69 Å) of SPC/E water. To neutralize the charge, 22 sodium ions were also added to the system resulting in a total 18060 atoms, see Figure 1. After initial optimization the system was brought to equilibrium by warming up in stages (50 K increments) over the course of 60 ps of classical molecular dynamics simulation. Excitation energies of the quantum representation of cytosine base capped with hydrogen link atoms in the field of the entire DNA-water complex (18048 point charges) were calculated every 0.5ps in the context of combined CR-EOMCCSD(T)/MM approach. CR-EOMCCSD(T) calculations were based on cc-pVDZ basis sets [16] with frozen core electrons. Our results clearly demonstrate the qualitative impact of the fluctuating environment. After thermal averaging based on formula (4) we obtained excitation energies of 5.01 and 5.79 eV for the \( n_o \pi^* \) and \( \pi\pi^* \) states, respectively (see Figure 2). These values should be contrasted with the gas-phase (carried out for the ground-state optimized geometry on the B3LYP level) calculations for these states which are equal to 4.76 and 5.24 eV, respectively. Therefore, significant “blue shift” of 0.25 and 0.54 eV was detected for these states in our simulations.

Interestingly enough, during our simulation we observed an increasingly more important, compared to the gas-phase calculations, role played by doubly excited configurations in the wavefunction expansion of the \( n_o \pi^* \) state. This fact, which should be attributed to the effect of the surrounding environment, has interesting theoretical implications. Since such cases are simply beyond the scope of TD-DFT methods which are intrinsically built upon single excitation model, the high-level ab-initio approaches that properly account for the dynamical and non-dynamical correlation effects are the only methods that can provide highly predictive results.

We believe that the approach described in this manuscript along with its application to an important problem of photostability of DNA will be of interest to a general physics/chemistry community and will facilitate further integration of high-level methods into a large-scale dynamical calculations.

4. Acknowledgement

This research was performed in part using the MSCF in EMSL, a national scientific user facility sponsored by the U.S. DOE, OBER and located at PNNL. We would also like to acknowledge support to M.V. from the Advanced Scientific Computing Research program of the U.S. Department of Energy, Office of Science (DE-AC06-76RLO 1830).
Figure 1. The graphical representation of the cytosine base fragment in 12mer B-DNA fragment.

Figure 2. Time evolution of the two Lowest excited states of cytosine.

References
[1] Paldus J and Li X 1999 Adv. Chem. Phys. 110 1
[2] Bartlett R J 2005 in Theory and Applications of Computational Chemistry The First Forty Years, edited by C.E. Dykstra, G. Frenking, K.S. Kim, G.E. Scuseria, p.1191.
[3] Urban M, Noga J, Cole S J, Bartlett R J 1985 J. Chem. Phys. 83 4041.
[4] Raghavachari K, Trucks G W, Pople J A, Head-Gordon M 1989 Chem Phys Lett 157 479.
[5] Geertsen J, Ritthby M, Bartlett R J 1989 Chem. Phys. Lett. 164 57.
[6] Stanton J F, Bartlett R J 1993 J. Chem. Phys. 98 7029.
[7] Kongsted J, Ousted A, Mikkelsen K V, Christiansen O 2003 J. Chem. Phys. 118 1620.
[8] Valiev M, Kowalski K 2006 Phys. Rev. Lett. submitted.
[9] Abo-Riziq A, Grace L, Nir E, Kabelac M, Hobza P, de Vries M S 2005 Proc Natl Acad Sci USA 102 20.
[10] Wood R H, Liu W B, Doren D J 2002 J. Phys. Chem. A 106 6689.
[11] Torrie G M, Valleau J P 1974 Chem. Phys. Lett. 28 578.
[12] Kowalski K, Piccuch P 2004 J. Chem. Phys. 120 1715.
[13] Watts J D, Bartlett R J 1999 Spectrochim. Acta Part A 55 495.
[14] Hirata S 2003 J. Phys. Chem. A 107 9887.
[15] Apra E, Windus T L, Straatsma T P, Bylaska E J, de Jong W, Hirata S, Valiev M, Heckler M, Pollack L, Kowalski K, Harrison R, Dupuis M, et. al. “NWChem, A Computational Chemistry Package for Parallel Computers, Version 4.7” 2005 Pacific Northwest National Laboratory, Richland, Washington 99352-0999, USA.
[16] Dunning Jr T H, 1989 J. Chem. Phys. 90 1007.