Hybrid coupled cluster and molecular dynamics approach: Application to the excitation spectrum of cytosine in the native DNA environment

Marat Valiev$^a$ and Karol Kowalski$^b$

$^a$Molecular Sciences Software Group, Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Washington 99352

$^b$Electronic mail: marat.valiev@pnl.gov

(Received 19 September 2006; accepted 8 November 2006; published online 7 December 2006)

Evolution of the excited state energies of cytosine base in the native DNA environment was investigated using a hybrid coupled cluster and classical molecular dynamics approach. The time averaged excitation energies obtained with the variant of the completely renormalized equation-of-motion with singles, doubles, and non-iterative triples approach that includes a bulk of the correlation effects for excited states, are compared with the analogous calculations in the gas phase. Significant blue shifts for the two lowest singlet excitation energies can be observed as a result of the interaction of the quantum system with the surrounding environment. © 2006 American Institute of Physics. [DOI: 10.1063/1.2403847]

The coupled-cluster (CC) method$^{1-4}$ has become a standard tool for highly accurate calculations of the ground- and excited-state properties in quantum chemistry (for recent advances of the CC method in nuclear physics see Refs. 5 and 6). Still, the majority of CC-based calculations for molecular systems are performed in the gas phase and, for the most part, neglect any environmental effects let alone dynamical fluctuations. Yet we can anticipate that in many problems of practical interest the environment and dynamical features can have a significant impact on the calculated properties. Recent studies$^7$ clearly indicate that interactions with the surrounding environment even at the static level have a large influence on properties, such as excitation energies. The challenge still remains of how to efficiently reconcile significant computational overhead of CC methods with a proper accounting of the environment and its dynamical fluctuations. Clearly the use of high-level ab initio methods is imperative in attaining the desired level of accuracy for vertical excitation energies and excited-state potential energy surfaces. This is best epitomized by the gas-phase studies of excited states of cytosine where several mutually exclusive scenarios of ultrafast internal conversion have been proposed (for details, see Refs. 8–12 and references therein) to elaborate the vital problem of photostability of DNA. Given that the inclusion of the environment may result$^7$ in as much as 0.8 eV shifts in vertical excitation energies, these effects should be at least approximately included in excited-state studies of biological systems. The main objective of this paper is to address these issues by an approach that allows calculation of dynamically averaged excited state properties in the context of large scale simulations. The approach rests on a combination of a non-Boltzmann sampling scheme,$^{13}$ an efficient parallel implementation of the coupled-cluster method, and classical molecular dynamics simulations. In this approach, dynamics of the entire system is driven by a lower level theory ($E_0$). The resulting trajectories are then processed with higher level theory ($\tilde{E}$), albeit at a much lower rate, to calculate the actual observable ($\omega$) according to the following expression:

$$\langle \omega \rangle = \frac{\langle \omega e^{-\beta E-E_0} \rangle}{\langle e^{-\beta E-E_0} \rangle}.$$  (1)

Given our intent to calculate excitation energies in the biological environment (DNA), we choose the lower level Hamiltonian ($H_0$) to be represented at the classical molecular mechanics level with an Amber-type force field.$^{14}$ The higher level description is given by a combination of coupled-cluster and molecular mechanics (CC/MM) descriptions with the following Hamiltonian:

$$H = H_{QM} + H_{QM/MM} + H_{MM},$$  (2)

where $H_{QM}$ is the standard many-electron Hamiltonian describing the internal energy of the QM region

$$H_{QM} = E_{QM}^{(0)} + \sum_{\mu,\nu} f_{\mu\nu}^{a\dagger}a_\mu + \sum_{\mu,\nu,\lambda,\kappa} \nu_{\mu\lambda\nu\kappa} a_{\lambda}^{\dagger}a_{\kappa}^{\dagger}a_\mu a_\nu,$$  (3)

where indices $\mu$, $\nu$, $\lambda$, and $\kappa$ designate single-particle states. The interaction between the QM region and its surroundings (MM region) is contained in the second term $H_{QM/MM}$.

$$H_{QM/MM} = \sum_{i,\mu,\nu} \langle \mu | \frac{Q_i^{\mu} | \mathbf{R}_m - \mathbf{r} | \nu \rangle a_\mu^{\dagger}a_\nu + V(|\mathbf{R}^\mu\rangle, \{\mathbf{R}\}),$$  (4)

where $Q_i^{\mu}$ and $\mathbf{R}_m$ denote charges and coordinates of the MM region. The third term in Eq. (2), $H_{MM}$, 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

$$\tilde{H} = H_{QM} + H_{QM/MM},$$  (5)

effectively includes, through the $H_{QM/MM}$ term, the interaction of the environment with the QM region and forms the...
basis for our electronic structure calculations within CC/MM description of ground (E) and excited state (ω) energies.

As a brief summary of the salient features of the CC approach, we note that it exploits an exponential parametrization of the ground-state wavefunction |Ψ0⟩=e|^Φ⟩, where T is the cluster operator and |Φ⟩ is a reference function most frequently chosen as a Hartree-Fock (HF) determinant. In the approximate approaches, for example the CCSD method (CC with singles and doubles), the expansion for the cluster operator T is truncated on the low-order many-body components, i.e., T=T1+T2, where singly and doubly excited components T1=∑i,j=1,2,⋯|ti,j⟩⟨ai| and T2=∑i,j,k=1,2,⋯|ti,j⟩⟨ai|aj| defined through the singly and doubly excited cluster amplitudes tijk and ti,j, respectively. As always indices i,j,k,⋯ (a,b,c,⋯) are used to denote occupied (unoccupied) spin orbitals in the reference |Ψ⟩. The cluster amplitudes are calculated by solving the equations ⟨Φ|^H|^Φ⟩=0 and ⟨Φ|ab|^H|^Φ⟩=0, where ^H=e−T^H^Te and |Φ⟩ are the singly and doubly excited Slater determinants. The CCSD energy is calculated from the expression E_{CCSD} = \langle Φ|\hat{H}|Φ⟩. Combined with the energy of the MM region, this is the energy used in the non-Boltzmann resampling expression E = E_{CCSD} + E_{MM} [see Eq. (1)].

The excitation energies [ων in Eq. (1)] were calculated according to the excited-state extension of the ground-state CC theory, the equation-of-motion coupled-cluster (EOMCC) formalism (or linear response CC approach), which is based on a particular simple form of the K-th state wavefunction |ΨK⟩=e|^Φ⟩, where RK is the excitation operator. For the EOMCC formalism with singles and doubles (EOMCCSD), the K-th state is given explicitly by the expression |ΨK⟩=e(RK0+RK1+RK2)e|^Φ⟩, where RK0, RK1, and RK2 are zero-, one-, and two-body components of the RK operator. The EOMCCSD excited-state energies for the K-th excited state (or excitation energies ω_{CCSD}^ν) and corresponding RK operator are obtained by diagonalizing the RK operator in a space spanned by the reference function and all singly and doubly excited determinants. As discussed in the literature, the EOMCCSD approach provides a satisfactory description of the excited states dominated by single excitations with respect to the reference |Ψ⟩.

In order to further improve the accuracies for singly excited states, one needs to include the effect of triply excited configurations either in iterative or, in a more economical manner, non-iterative fashion. For the purpose of this paper we adopted the δ(IA) variant of completely renormalized EOMCCSD approach with non-iterative triples (CR-EOMCCSD(T)), and the corresponding excitation energy defined as:

$$ω^K_{\text{CR-EOMCCSD(T)}} = ω^K_{\text{EOMCCSD}} + δ^K_{\text{CR-EOMCCSD(T)}},$$

where

$$δ^K_{\text{CR-EOMCCSD(T)}} = \sum_{i<j<k,a<b<c} Z^{abc}_{K,ijk} M^{ijk}_{K,abc} / D^K.$$  (7)

In Eq. (7), the quantities M^{ijk}_{K,abc} refer to triply excited moments of the EOMCCSD equations, whereas the tensor Z^{abc}_{K,ijk} is defined as \langle Φ| \left[R^K_0 \left(T_1 \epsilon^+ T_2 + 1 / 6 \epsilon^2 T_3 \right) + R^K_1 \left(T_2 + 1 / 2 \epsilon^2 T_3 \right) + R^K_2 \left(T_3 \right) \right] |Φ⟩^{abc} with \vec{R}^{abc}_{K,3} representing an approximation of the exact, triply excited RK₃ operator, where the amplitudes, \vec{R}^{abc}_{K,3}, are set equal to \sum M^{ijk}_{K,abc} / (ω^K_{EOMCCSD} + \epsilon^+ + \epsilon^j + \epsilon^k + \epsilon^+ - \epsilon^a - \epsilon^b - \epsilon^c) with ε’s corresponding to the HF orbital energies. The D_K quantity is the overlap between trial and EOMCCSD wavefunctions for the K-th state. The overall scaling of the method is λn_i λ_n_i, where n_i and n_a are the number of occupied and unoccupied orbitals correspondingly.

The CC/MM approach described above was recently implemented by us in the NWChem computational chemistry package. It is based on a seamless integration between the generic QM/MM interface, Tensor Contraction Engine module, and the classical molecular dynamics module and provides a highly scalable parallel framework for accurate large scale simulations. This approach was recently applied to the calculation of the two lowest excited states of a water molecule described by the d-aug-cc-pVTZ basis set and embedded in a 30 Å cubic box of classical SPC/E water molecules. A average of 40 ps classical molecular dynamics simulations with 0.5 ps intervals for an excitation energy calculation at the EOMCCSD level produced a blue shift on the order of 0.63 eV. This compares well with a similar calculation performed by Kongsted et al., giving strong encouragement in favor of our approach.

The system considered in this work (see Fig. 1) consists of 1,2-mer fragment of B-DNA (3’T-CGCGTGTGCCT-T5’) solvated in a rectangular box (51 × 51 × 69 Å) of SPC/E water. To neutralize the charge, 22 sodium ions were also added to the system resulting in a total of 18 060 atoms. 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.

FIG. 1. (Color) The 12mer B-DNA fragment studied in this work. The inset shows the quantum region containing the cytosine base fragment.
The production run was performed at constant temperature and pressure (298.15 K, 1.025 × 10^5 Pa) for 60 ps using a 15 Å cutoff. The autocorrelation analysis has shown that after 0.5 ps the snapshots from classical MD trajectory became essentially uncorrelated. Thus, the excitation spectra of the cytosine base was calculated every 0.5 ps within the context of the combined CR-EOMCCSD(T)/MM approach (see Fig. 2) using the cc-pVDZ basis set.\textsuperscript{22} These calculations were based on a quantum representation of cytosine base capped with a hydrogen link atom in the field of entire DNA-water complex (18 048 point charges). The simulation was performed in parallel and took a total of 105 hours (52 min per single snapshot) utilizing 256 processors on the HP/Linux Itanium-2 cluster. We focused our attention on the two lowest excited singlet states that can be identified with the valence type π→π* (the 1ππ* state) and nO→π* (the 1nOπ* state) transitions. The statistical errors in the average excitation energies are estimated to be less than 0.01 eV. As explained by the authors of Ref. 11 once the dynamic correlation is properly taken into account by the CASPT2 method (second order perturbative correction for the complete active space reference function) the conical intersection between the ground state and the 1ππ* state is responsible for the ultrafast decay of the singlet excited state of cytosine. This is in sharp contrast to the CASSCF predictions, which favor the 1nOπ* state as the lowest excited state. The results of our calculations are collected in Table I. Using Eq. (1), we have obtained thermally averaged CR-EOMCCSD(T)/MM vertical excitation energies of 5.01 and 5.79 eV (5.27 and 6.01 with the EOMCCSD/MM approach) for the 1ππ* and 1nOπ* states respectively. This should be compared to the gas phase

<table>
<thead>
<tr>
<th>EOMCCSD</th>
<th>CR-EOMCCSD(T)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1ππ*</td>
<td>1nOπ*</td>
</tr>
<tr>
<td>5.02</td>
<td>5.44</td>
</tr>
<tr>
<td>4.76</td>
<td>5.24</td>
</tr>
<tr>
<td>5.27</td>
<td>6.01</td>
</tr>
<tr>
<td>5.01</td>
<td>5.79</td>
</tr>
</tbody>
</table>

From Table I, one can also see that the environment effectively stabilizes the lowest excited state by increasing the gap between the 1ππ* and 1nOπ* states to 0.77 eV (compare this with the analogous gap of 0.48 eV obtained in the gas-phase calculations). This fact may amplify the role played by the first excited state in the ultrafast nonradiative conversion. Another issue is the impact of the environment on the structure of the excited-state wavefunctions. Detailed analysis of the largest R amplitudes obtained in our simulation shows that for the 1nOπ* state the leading singly-excited components are accompanied by increasingly large values of doubly-excited contributions compared to the gas phase calculations. This situation cannot be described by CIS or TD-DFT methods and in fact requires the use of methods accounting for the effect of triples.

In conclusion, we have developed an approach for highly accurate characterization of excited states in large biomolecular systems in a finite temperature setting. The described methodology may provide a useful alternative to conventional QM/MM simulations of excited states based on DFT.\textsuperscript{25} We have illustrated our approach by calculating excited states of cytosine base in the native DNA setting taking into account not only the presence of the protein environment but also its dynamical fluctuations. Significant blue shifts for the two lowest singlet excitation energies were observed compared to gas phase calculations. The inclusion of the environment also stabilizes the lowest excited state by increasing the gap between the first and second excited states. These results give us a strong motivation to develop credible and efficient \textit{ab initio} methods for molecular dynamics simulations,\textsuperscript{29,28} as well as sophisticated models for QM-environment interactions\textsuperscript{26,27} that would enable more accurate dynamical studies of large biochemical systems.

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).

\textsuperscript{1} F. Coester, Nucl. Phys. 7, 421 (1958).
\textsuperscript{2} F. Coester and H. Kümmel, Nucl. Phys. 17, 477 (1960).
\textsuperscript{4} R. J. Bartlett, \textit{in Modern Electronic Structure Theory}, edited by D. R.
References


