2P105

Improving the elongation method: the intermediate electrostatic field for DNA and proteins via genetic algorithms

○Denis Mashkovtsev[1], Wataru Mizukami[1], Jacek Korchowiec[1,2], Anna Stachowicz-Kuśnierz[2], Yuriko Aoki[1]

[1] *Department of Molecular and Material Sciences, Kyushu University, Japan*
[2] *Department of Chemistry, Jagiellonian University, Poland*

[Abstract]

Quantum chemical methods are computationally expensive, especially for large molecules. Therefore, a number of approximate reduced-scaling methods are proposed. In this work Elongation, one of such methods, based on simulating the polymerization process, is considered. In view of the linearity of the polymerization process during the simulation of the growing chain, not all fragment-fragment interactions are included. A possible improvement is the introduction of atomic charges in the places of the following fragments for the inclusion of electrostatic interactions. For this purpose, Charge Sensitivity Analysis, a QSAR/QSPR method, is considered. Because of its structure analytical procedure cannot be used for parametrization; thus the genetic algorithm was utilized with the fitness-function based on a comparison of atomic charges in molecules from a training set with the reference charges, calculated by several quantum chemical methods. Finally, obtained models were tested in calculations via Elongation on several peptides and DNA chains.

[Introduction]

Quantum chemical methods theoretically allow one to calculate properties of any many-electron system with any desired accuracy. However, in practice, computational costs with an increase in the size of the system grow so tremendously that calculation of properties is possible only for systems consisting of several hundred atoms. In this case, a significant part of molecules of biological interest, the simulation of which is necessary for an understanding of processes in living beings and *de novo* drug-design, are left behind. Such a state of affairs challenges the creating of reduced-scaling quantum chemical methods. The main approach in existing solutions is the division of the entire systems into fragments. In this work, Elongation is considered, one of such methods, based on simulating the polymerization process [1]. In this method, only a small part of a molecule, called active, is treated. During the calculation, the active part is moving along polymer chain, so the first fragment is being deleted to frozen part, and next fragment is being added to the active part. However, in this process, every fragment is influenced only by already considered fragments. Therefore, a reproducibility of the electronic structure decreases. A possible improvement is the introduction of intermediate electrostatic field (IEF), i.e. introduction of atomic charges in the places of the following fragments for the inclusion of electrostatic interactions. However, accurate calculation of atomic charges leads to loss of benefits from fragmentation of the system, therefore it is required to utilize approximate methods. One of such methods, Charge Sensitivity Analysis (CSA), is considered [2]. CSA is based on electronegativity equalization principle and allows one to obtain atomic charges knowing only the structure of the molecule and its total charge. However, preliminary CSA has to be parametrized. This is a typical optimization problem, but the mathematical basis of CSA does not allow to find parameters via any analytical procedure. For this reason, parametrization was carried out by the genetic algorithm, heuristic method based on evolution process.

[Methods]

Derivation of parameters for CSA was performed via the genetic algorithm using the concept of Darwinian evolution with full elitism approach to building a new generation. Evolution was 500 generations long; every generation consisted of 20000 individuals. Final parameters were averaged on 5 evolution cycles. Parents were chosen from a tournament on 3 random individuals. Probabilities of crossover, individual and allele mutation were 1.0, 1.0, and 0.7. The training set consisted of 287 small molecules with 6 atom types: H, C, N, O, P, and S. Testing set consisted of 27 molecules. All molecules in both sets were optimized at the B3LYP/6-31G(d) level of theory in Gaussian 16. Thereafter, reference atomic charges using 4 techniques (**NPA**, **MK**, **CM5**, **APT**) with different theoretical basis were carried out in a single point calculation on the same level of theory. Obtained parameters were used for calculation of atomic charges in 4 biopolymers: 2 peptides (α-helix built by glutamic acid and alanine: *gluala*, PDB: *2m0w*) and 2 DNA chains (DNA chain with random sequence: *randDNA*, PDB: *1ap1*). Considered biopolymers were calculated via the Elongation method with and without IEF via the Hartree-Fock method with 6-31G basis set in GAMESS 2012 package. Electronic energies of biopolymers obtained from the full-molecule calculation on the same level of theory were used as a reference.

[Results and Discussion]

Coefficients of linear correlation between atomic charges calculated via CSA and reference atomic charges (RAC) obtained via different calculation techniques are presented in **Table 1**. Comparison of r^2 for training and testing sets shows that correlation coefficients are similar and obtained parameters, in general, reproduce charges. However, low values of r^2, especially for **MK**, indicate poor transferability of charges, but bigger values for the testing set than for training set can be also evidence of a too small number of degrees of freedom in the model, i.e. 12 parameters are not enough for utilized sets.

Charge model used for RAC	r^2 for training set	r^2 for testing set
NPA	0.8800	0.9162
MK	0.8068	0.8137
CM5	0.8924	0.9034
APT	0.9085	0.9195

Table 1. Coefficients of linear correlation between CSA charges and RAC

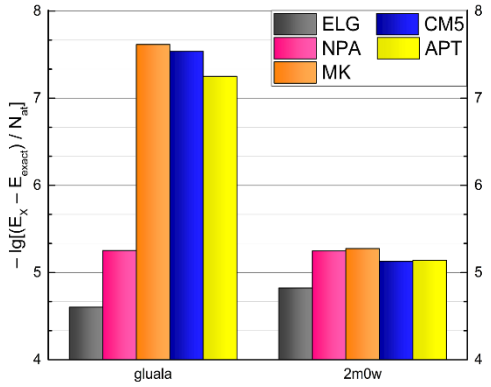

Fig. 1. Negative logarithm of per atomic difference in electronic energies via Elongation w/o and w/ IEF and for full-molecule

Negative logarithms of per atomic difference between electronic energies of two peptides (*gluala*, *2m0w*) without (**ELG**) and with (**NPA**, **MK**, **CM5**, **APT**) IEF and full-molecule results are presented in **Fig. 1**. Since negative logarithms are calculated, a higher bar means better. Comparison of results with atomic charges against **ELG** shows that the introduction of electrostatic field improves results. Noticeably larger improvement is observed for more linear *gluala*, since in linear structure long-range interactions have bigger impact than the interaction between adjacent fragments in opposite to three-dimensional structures. Thereby, the use of IEF is justified, however, it is still necessary to improve a lot in terms of close interactions in the future.

Results for DNA chains will be discussed during the presentation.

[References]

[1] A. Imamura et.al., *J. Chem. Phys.* **95**, 5419 (1991).

[2] A. Stachowicz et.al., *J. Mol. Model.* **17**, 2217 (2011).