



# Implementation of Fragment Orbital Method (FMO) for Highly Parallelized Quantum Chemical Calculations with CP2K

Peicho Petkov<sup>a</sup>, Petko Petkov<sup>b,\*</sup>, Georgi Vayssilov<sup>b</sup>, Stoyan Markov<sup>c</sup>

<sup>a</sup>*Faculty of Physics, University of Sofia, 1164 Sofia, Bulgaria*

<sup>b</sup>*Faculty of Chemistry, University of Sofia, 1164 Sofia, Bulgaria*

<sup>c</sup>*Natioanl Centre for Supercomputing Applications, Sofia, Bulgaria*

---

## Abstract

The reported work aims at implementation of a method allowing realistic simulation of large or extra-large biochemical systems (of  $10^6$  to  $10^7$  atoms) with first-principle quantum chemical methods. The current methods treat the whole system simultaneously. In this way the comput time increases rapidly with the size of the system and does not allow efficient parallelization of the calculations due to the mutual interactions between the electron density in all parts of the system. In order to avoid these problems we implemented a version of the Fragment Orbital Method (FMO) in which the whole system is divided into fragments calculated separately. This approach assures nearly linear scaling of the compute time with the size of the system and provides efficient parallelization of the job. The work includes development of pre- and post-processing components for automatic division of the system into monomers and reconstructing of the total energy and electron density of the whole system.

---

## 1. Main goals

The realistic simulation of various biochemical systems requires the use of a reliable first-principle quantum chemical method for extra-large systems (of  $10^5$  to  $10^6$  atoms) for long simulation times, from nano to microseconds. Despite the high performance that parallel computing systems make available, their use for such simulations is not trivial at least for two reasons:

- even the most appropriate first-principle methods based on density functional theory (DFT) scale with  $N^3$  ( $N$  is the number of electrons in the system), while appropriate scaling is linear;
- the possibilities for efficient parallelization of a quantum chemical system, if it is considered as one system, are limited due to mutual interactions between the electron density in all parts of the system (including Coulomb and exchange-correlation interactions).

There are different approaches [1-4] designed to provide essentially linear scaling of the computational power with the size of the system and efficient parallelization of the computational job. These approaches are based on division of the whole system into large number of fragments (*monomers*). One of these methods for division of a large system into fragments (denoted as Fragment Molecular Orbital method, FMO) to be calculated simultaneously on different nodes will be implemented in CP2K. The CP2K code shows scalability  $N \log N$  where  $N$  is the number of electrons in the system. By this reason we selected CP2K code for implementation of our method. This approach will provide nearly linear scaling with the size of the system and provide efficient parallelization of the computational job. The change in the scaling is achieved not by modification of the quantum chemical calculations but from this division of the system. In this case the computational time is proportional to the number of monomers  $n_f$  and the time necessary for calculation of the individual monomers, which still scales as  $N_e^3$  for DFT ( $N_e$  is the number of electrons in the monomer) method but not in the case of CP2K code as it was mentioned above:

---

\* Corresponding author. *E-mail address:* [ohpp@chem.uni-sofia.bg](mailto:ohpp@chem.uni-sofia.bg).

$$t_{mono} \sim n_f \times N_e^3 \quad \text{or} \quad t_{mono} \sim n_f \times N \log N \quad \text{in the case of CP2K} \quad (2)$$

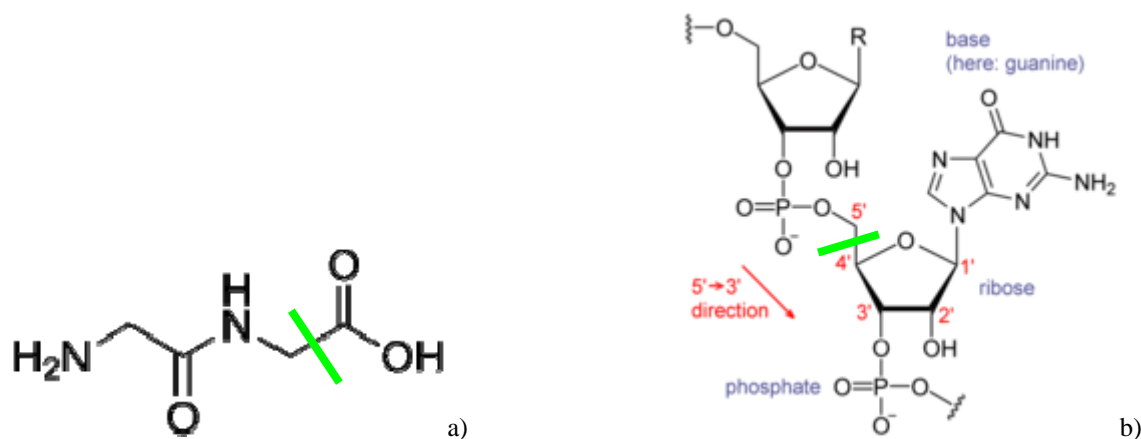
Thus, if the size of the studied system is extended not by increasing the size of the individual fragments but by adding new fragments, then the compute time will increase linearly with  $n_f$ , i.e. with the size of the system. Moreover, if the next fragments are calculated at additional nodes, then the real time of the calculation does not increase. Such perfect behaviour can be accomplished, however, only for non-interacting fragments, which is not of particular interest to chemistry. Different strategies to take into account the interactions between fragments have been developed. They are based on separation of the calculation for the whole system into three components:

- global Coulomb interactions of each fragment with the rest of the system, which are less time consuming than the quantum part of the calculations;
- first principle calculation for the quantum (exchange-correlation) interactions within the isolated fragment (in the field of the rest of the system);
- estimation of the quantum interaction of the fragment with other (neighbouring) fragments.

Since the Coulomb interactions are long-range, it is calculated completely in all methods. The main difference between different approaches comes from the division of the system into fragments and accounting of the quantum interactions between them.

## 2. Work done in the project

The Fragment (or localized) Molecular Orbital method (FMO) is designed for bio-molecular and other covalent system where the fragments are formed by breaking a chemical bond. For this reason the method includes special treatment of the border atoms of the fragments that allows the quantum chemical calculation of each to converge, even if the bonds at the border are not correctly described. Further improvement can be achieved if the interaction between fragments is accomplished by calculations of "dimers" — the pairs of fragments, which corrects the description of the bonds between fragments (where they exist) and also allows for the accounting of the exchange-correlation interactions between fragments. In order to perform calculations of large bio-molecules at DFT level, a concept for division of the whole system into fragments (for proteins and nucleic acids) and a number of pre-processing modules were developed. The system is divided into fragments as shown on Fig. 1. The dangling bonds are saturated by H atoms placed on the C-C bond.



**Figure 1.** Recommended bonds to be cleaved in the structure of proteins (a) and DNA/RNA (b).

A pre-processing set of programs for FMO with the CP2K software package was developed. The first one, called FRAGMENTEN, divides the system into fragments, generates an input file for CP2K for each fragment; the second one, called GRIDPOT, calculates the electrostatic field of the rest of the system into the space of each fragment.

The FRAGMEN program accepts an input file where the fragmentation procedure parameters are set. All possible parameters are:

- `project_name`: sets the base of filenames of possible inputs
- `input_type`: possible values - pqr, pdb, pdbpsf, xyz
- `number_of_fragments`: <optional> number of fragments
- `max_atoms_per_fragment`: <optional> maximum number of atoms per fragment, default 200
- `fragment_prefix`: prefix of fragments file names
- `res_distance_cut`: <optional> maximum distance between subsequent residues in the fragment, units [angstrom], default 7.0
- `fragments_residue_count`: distribution of the number of residues along the fragments, counted in the same order as in the input sequence. The `number_of_fragments` option has to be defined.
- `fragments_charges`: distribution of the net charge of along the fragments ignoring the charge from input files. The `number_of_fragments` option has to be defined.
- `delta`: sets grid cell size

In general the program FRAGMEN reads the protein's atom coordinates and charges from a file in pqr format and, depending on the fragmentation parameters, divides the protein into fragments according to either **fragments\_residue\_count** or the average number of residue per fragment. One can simply put an upper limit for the number of atoms per fragment and the software will create fragments where the atom numbers do not exceed the limit. The fragmentation is followed by termination of fragments sharing carbons and alpha carbons connected by bonds. The atom coordinates and charges for every fragment are written to separate files, that are subsequently read by GRIDPOT and CP2K. At the end, FRAGMEN creates a parameter input files for GRIDPOT and CP2K.

The GRIDPOT program reads the parameter files written by FRAGMEN. It calculates the electrostatic potential on the grid and writes the values to disk later to be read by the CP2K package.

DFT calculations with predefined external electrostatic field are originally implemented in the CP2K software package but one can only set an analytical function to calculate the external field. If the field map can not be described by a simple formula, which is the case with FMO protein calculation, one needs a possibility to load the electrostatic field from an external source. A module reading the precalculated electrostatic field and storing its data distributed across the computing nodes was developed and added to the CP2K source code in order to overcome this issue. The third party library called Global Array Toolkit [5] was used to distribute and access the electrostatic field data in an easy way. The electronic structure of each element is calculated by CP2K and the electronic density and energy is written to files for further processing.

A post-processing module, POSTFRAG, which calculates the total electron density as a sum of electron density of all monomers and the total energy of the system was developed.

### 3. Results

As a case study and in order to check the functionality of the pre-processing modules, the structure of Trp-Cage Miniprotein, consisting of 304 atoms (pdb code: 1L2Y), was divided into 10 fragments. The calculations were performed on 8 cores Intel architecture system. It was taken as an example because it is one of the smallest proteins and one can calculate the whole structure of the protein at DFT level. Using the pre-processing modules, the structure was divided in 10 fragments and their electronic structure was calculated separately. The time for single point calculation of the whole structure of the protein on 8 CPUs is ~2500 s, while the single point calculation per fragment takes ~200 s and generation of the electrostatic field takes additional ~160 s.

Using 128 CPU partition (512 cores) of Blue Gene/P machine the structure of Human Interferon Gamma was tested. It is a pleiotropic cytokine endowed with multiple biological activities such as antiviral, antibacterial, antiproliferative, antitumor, immunoregulatory and gene activity regulation. The mature form of hIFN- is a non-covalent homodimer, consisting of two 17 kDa monomers. Each monomer comprises 143 amino acids, organized in six alpha-helices.[6] For this test the hIFN-gamma structure without c-termini was extracted for PDB crystallographic structure under ID 1FG9 that contains 127 amino acid residues (4190 atoms). The whole structure of this protein was submitted with the CP2K code on 512 cores but the calculation failed. Than the

structure of the hIFN was divided in 14 fragments with the pre-processing module FRAGMEN. The FRAGMEN option `max_atoms_per_fragment` was set to 300. The calculation of the electrostatic potential per fragment takes ~4 min, the electron density was calculated for 10 min. The total time for calculation of the electrostatic potential and the electron density of all 14 fragments will take approximately 196 min, if these fragments are calculated subsequently. However if the fragments are calculated simultaneously than the total time is approximately 15 min. .

The main advantage of the method proposed here is that the calculation of the electronic structure of a real protein, which consists of  $\sim 10^6$  atoms, would be possible at DFT level if it is divided in fragments in such a way that each fragment will be in range of several hundred atoms. The division of a large system of many subsystems which do not communicate during the calculation, provides conditions for scaling of the proposed method in the petascale regime. The future improvements include extension of the software allowing DNA and RNA fragmentation and better treatment of the exchange – correlation effects by calculation of dimmers (dimer = combination from two monomers which are closer in the space than predefined cut-off distance ) and implementation of faster algorithm for calculation of the electrostatic potential.

### Acknowledgements

This work was financially supported by the PRACE project funded in part by the EUs 7th Framework Programme (FP7/2007-2013) under grant agreement no. RI-211528 and FP7-261557. The work was achieved using the PRACE Research Infrastructure resources Blue Gene/P, National Centre for Supercomputing Applications, Bulgaria.

### References:

1. W. A. Goddard, R. Biswas, D. Srivastava, L. H. Yang, A. Nakano, R. K. Kalia, K. Nomura, A. Sharma, P. Vashishta, F. Shimojo, A. C. T. van Duin, De Novo Ultrascale Atomistic Simulations On High-End Parallel Supercomputers, *Int. J. High Perform. Comput. Appl.* 22 (2008) 113.
2. F. Shimojo, R. K. Kalia, A. Nakano, P. Vashishta, Divide-and-conquer density functional theory on hierarchical real-space grids: Parallel implementation and applications, *Phys. Rev. B* 77 (2008) 085103.
3. Z. Zhao, J. Meza, B. Lee, H. Shan, E. Strohmaier, D. Bailey, L.-W. Wang, The linearly scaling 3D fragment method for large scale electronic structure calculations, *J. Phys.: Conf. Ser.* 180 (2009) 012079.
4. D. Fedorov, K. Kitaura, *The Fragment Orbital Method*, CRC Press, Boca Raton, 2009.
5. Global Arrays Toolkit official site. <http://www.emsl.pnl.gov/docs/global/>
6. D.J. Thiel, M. –H. le Du, R.L Walter, A. D’Acry. C.Chene, M. Fountoulakis, G Gorata, F.K. Winkler, S.E. Ealick, *Structure* 8(9) (2000) 927