LLNL-TR-738503

A geometric initial guess for localized electronic orbitals in modular biological systems

P. G. Beckman, J. L. Fattebert, E. Y. Lau, D. Osei-Kuffuor

September 15, 2017

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

A geometric initial guess for localized electronic orbitals in modular biological systems

Paul Gustav Beckman,[1,2] Jean-Luc Fattebert,[3] Edmond Y. Lau,[4] and Daniel Ossei-Kuffuor[1]
[1] *Center for Applied Scientific Computing, Lawrence Livermore National Laboratory*
[2] *University of Chicago*
[3] *Computational Sciences and Engineering Division, Oak Ridge National Laboratory*
[4] *Physical and Life Sciences, Lawrence Livermore National Laboratory*

(Dated: 14 September 2017)

Recent first-principles molecular dynamics algorithms using localized electronic orbitals have achieved $O(N)$ complexity and controlled accuracy in simulating systems with finite band gaps. However, accurately determining the centers of these localized orbitals during simulation setup may require $O(N^3)$ operations, which is computationally infeasible for many biological systems. We present an $O(N)$ approach for approximating orbital centers in proteins, DNA, and RNA which uses non-localized solutions for a set of fixed-size subproblems to create a set of geometric maps applicable to larger systems. This scalable approach, used as an initial guess in the $O(N)$ first-principles molecular dynamics code MGmol, facilitates first-principles simulations in biological systems of sizes which were previously impossible.

I. INTRODUCTION

First-principles molecular dynamics (FPMD) is a computational method for studying matter at an atomistic scale used in a variety of fields. FPMD typically requires solving the equations of density functional theory (DFT),[5] the Kohn-Sham equations, in order to compute a system's electronic structure and the forces acting on its atoms. Although they are general and accurate, FPMD simulations become computationally infeasible beyond a few hundred atoms for tens of picoseconds due to the $O(N^3)$ computational complexity of typical DFT solvers. In response to these limitations, many $O(N)$ complexity algorithms have been developed.[1]

One such algorithm is MGmol, a parallel FPMD code developed at Lawrence Livermore National Laboratory which uses spherically localized electronic orbitals in order to reduce the computational cost and global communications of DFT calculations.[7] However, determining the initial positions of the orbital centers during setup using MGmol requires computing the electronic structure using non-localized orbitals, resulting in a $O(N^3)$ cost which is intractable for large biological systems. In order to make simulations of many realistic protein, DNA, and RNA systems possible using MGmol, we present a geometric approach to approximate orbital centers in $O(N)$ time which can be used as an initial guess during setup.

II. METHODS

Proteins, DNA, and RNA all exhibit modular structures consisting of a linear sequence of residues chosen from a limited set - the 20 amino acids in proteins, and the 5 nucleobases in nucleic acids. These sequences are connected by specific bonds - the peptide bond in proteins and the phosphodiester bond in nucleic acids. Our approach to inexpensively compute orbital centers in these systems is to decompose the input into its constituent residues and bonds, approximate the orbital centers for these pieces separately, and recombine the results to yield the orbital centers for the entire system. The approximation of orbital centers in each isolated part is done by a set of geometric maps. These maps must preserve the local electronic configuration of the residue or bond and must be robust to conformational changes, as the residues are not rigid. Each geometric map is computed from a template consisting of the residue or bond's atoms and non-localized orbital centers. As generating these templates requires the use of non-localized orbitals, it is an $O(N^3)$ task.[4] However, this computation is only ever done once, after which the mappings can be applied to new input systems in linear time. Thus, for systems in which N is very large, this geometric method replaces an $O(N^3)$ computation with an $O(N)$ computation by using the results of several one-time computations on subproblems of a small, fixed size. FIG. 1 outlines the approach and its computational complexity.

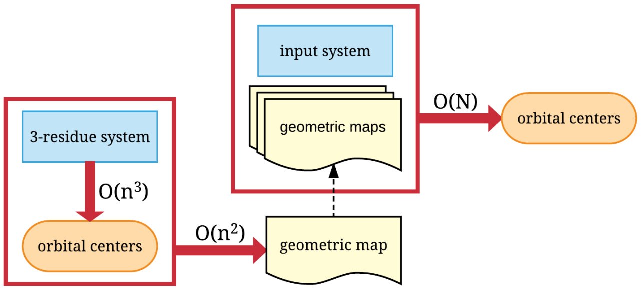

FIG. 1: **Geometric method overview**
Computations, shown as red arrows, are given with their respective costs, where n is the number of electrons in a 3-residue system and N is the number of electrons in an arbitrarily large input system.

A. Extracting templates

In order to obtain a template for each residue with the electronic configuration that would be found in a realistic biological system, we calculate maximally-localized Wannier functions (MLWF)[6] from which we extract orbital centers in systems consisting of 3 residues, where the middle residue is the portion of interest. In this way, we avoid possible alterations to the electronic configuration caused by adding terminating charge groups and removing the bonds that link residues.

Once the MLWF solution is obtained for the 3-residue system, we extract the atom name, residue name, residue number, and coordinates of each atom in the middle residue to be used as a template. We also extract all orbital centers whose nearest atoms belong to the middle residue for that residue's template.

B. Generating and applying maps

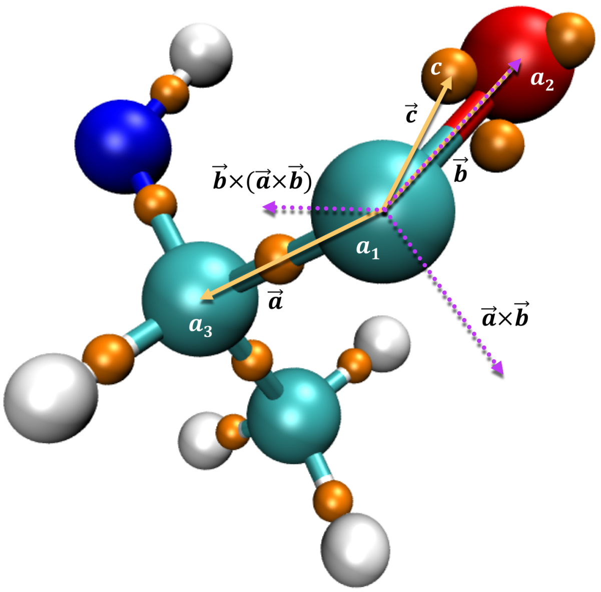

FIG. 2: **Residue template for alanine**
Orbital centers, shown in orange, calculated in $O(N^3)$ time using non-localized orbitals.

Given a template containing the locations of atoms and orbital centers, we hope to find a local basis that allows us to define the location of orbital center c in relation to the locations of nearby atoms. We first determine the locations and names of the three atoms a_1, a_2, and a_3 closest to c, ordered here from closest to farthest. We then take the primary bond vector $\vec{b}$ from atom a_1 to atom a_2, the auxilary bond vector $\vec{a}$ from atom a_1 to atom a_3, and the orbital center vector $\vec{c}$ from atom a_1 to c. We then calculate the orthogonal matrix A representing a local orthonormal basis.

$$A = \left[\frac{\vec{b}}{||\vec{b}||}, \frac{\vec{a}\times\vec{b}}{||\vec{a}\times\vec{b}||}, \frac{\vec{b}\times(\vec{a}\times\vec{b})}{||\vec{b}\times(\vec{a}\times\vec{b})||}\right]$$

We can then define $\vec{c}$ in this local basis by

$$\vec{c}_A = A^{-1}\vec{c} = A^T\vec{c}$$

In order to calculate the location in an unknown input system of an orbital center c_i which is analogous to c in the template, we locate atoms $a_{1,i}$, $a_{2,i}$, and $a_{3,i}$ in the input and define vectors $\vec{b_i}$ and $\vec{a_i}$ as before. We then calculate an orthogonal matrix A_i that defines a local basis in the input analogous to the basis used in the template

$$A_i = \left[\frac{\vec{b}_i}{||\vec{b}_i||}, \frac{\vec{a}_i\times\vec{b}_i}{||\vec{a}_i\times\vec{b}_i||}, \frac{\vec{b}_i\times(\vec{a}_i\times\vec{b}_i)}{||\vec{b}_i\times(\vec{a}_i\times\vec{b}_i)||}\right]$$

Then $\vec{c}_i$, which defines the location of the orbital center in the input, is given by

$$\vec{c}_i = \ell A_i\,\vec{c}_A \qquad \ell = \frac{||\vec{c}||\ ||\vec{b}_i||}{||\vec{b}||\ ||\vec{c}_i||}$$

where the length factor ℓ allows $\vec{c}_i$ to scale with the length of the bond $\vec{b}_i$. Thus, for each orbital center c_i, we have a mapping from a set of input atoms $\{a_j\}_{j=1}^N$ within a given residue to c_i that is fully defined by $\vec{c}$ in the local basis, a length scaling factor, and three atoms names

$$M_i : \{a_j\}_{j=1}^N \to c_i \qquad M_i = \left\{\vec{c}_A, \frac{||\vec{c}||}{||\vec{b}||}, (a_1, a_2, a_3)\right\}$$

A collection of these maps, one for each orbital center, defines a mapping from the set of input atoms $\{a_j\}_{j=1}^N$ within a given residue to the set of all its orbital centers $\{c_i\}_{i=1}^K$

$$R : \{a_j\}_{j=1}^N \to \{c_i\}_{i=1}^K \qquad R = \{M_i\}_{i=1}^K$$

In order to calculate approximate orbital centers for an entire input system, we decompose the set of input atoms into isolated residues and bonds and apply the corresponding map R to each component.

III. RESULTS

In the following section we study MGmol's performance when using the geometric method described above to generate an initial guess for the orbital centers.

MGmol computes electronic structure by minimizing the Kohn-Sham energy functional for a set of non-orthogonal electronic orbitals $\{\phi_i\}_{i=1}^N$ represented on a uniform finite difference mesh. By formulating each orbital as a MLWF confined to a strictly local region of

TABLE I

Peptide	Initial energy using atom-centered orbitals (Ha)	Initial energy using geometrically-generated orbitals (Ha)	Final energy (Ha)
GPG tripeptide	3733.43	-128.56	-153.68
GRG tripeptide	4700.49	-161.97	-192.79
GWG tripeptide	5062.77	-165.93	-199.60
10-residue peptide	18036.42	-479.17	-570.60

fixed radius, the evaluation of the energy functional and its gradient are done in $O(N)$ operations.[2] MGmol computes an approximation of the inverse of the Gram matrix S defined by $S_{ij} = \int_\Omega \phi_i(\mathbf{r})\phi_j(\mathbf{r})$ in $O(N)$ time by calculating the interactions only between orbitals with centers within some fixed cutoff radius.[7] We now examine the accuracy and scalability of these approximations in the context of modular biological systems.

A. Accuracy

When no explicit orbital centers are provided, MGmol uses atom-centered orbitals as an initial guess, after which an iterative steepest descent algorithm is used to converge to a stable electronic configuration. We compare the initial energy of a system using atom-centered orbitals and using geometrically-generated orbitals to the final energy of the system after the solver has converged to a low energy groundstate. TABLE I shows these results for a number of peptide systems. We see that the geometric method provides orbital centers with a much lower initial energy than the atom-centered orbitals and relatively close to the final converged energy. This indicates that the geometric approach gives a significantly better first guess than atom centering, and yields an electronic structure close to the converged groundstate.

Osei-Kuffuor and Fattebert[7] demonstrated that MGmol achieves an error on forces between atoms which decays exponentially with both the orbital confinement region radius and the cutoff radius between orbital centers used when computing the inverse Gram matrix. We replicate these results in a peptide chain consisting of 148 atoms, shown in FIG. 3. This confirms MGmol's accuracy in biological systems when using geometrically-generated orbitals as an initial guess.

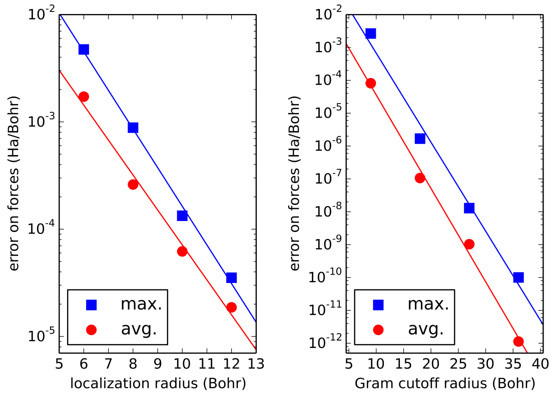

FIG. 3: **Error on forces**
Error shown as a function of the orbital confinement region radius (left) and as a function of the Gram cutoff radius (right).

B. Scaling

We test MGmol's parallel scaling on a system consisting of a single straight peptide chain in a continuous dielectric solvent[3] with Dirichlet boundary conditions. We begin with 493 atoms, which results in 666 doubly occupied orbitals in a 314 x 39 x 41 Bohr domain. We then double the peptide's length along the x-axis in each subsequent test. The number of processors is scaled proportionally to the problem size so that the number of mesh points per processor is held constant. 20 iterations of the DFT solver are run at each MD step to update the orbitals. Wall clock times per MD step are shown in FIG. 4.

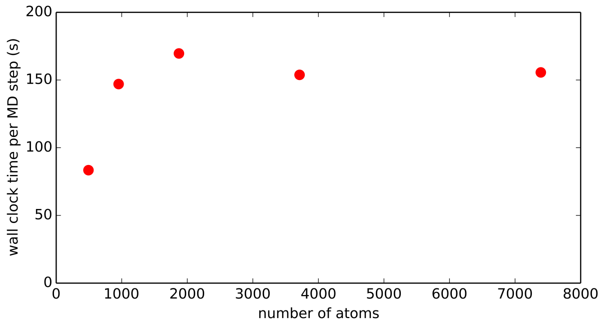

FIG. 4: **Parallel weak scaling**
Wall clock time per MD step on an Intel Xeon EP X5660 Linux cluster with high-speed interconnect (InfiniBand QDR QLogic). The number of mesh points per processor is held constant in each simulation.

C. Dense system simulation

In order to test MGmol's performance using geometrically-generated initial orbitals on a dense test system, we simulate a system consisting of the amyloid forming peptide GNLVS from the eosinophil major basic protein[8] solvated in water. In this test we use periodic boundary conditions and do not use a dielectric solvent. FIG. 5 shows the converged electronic structure of the 9.00 x 31.79 x 67.58 Bohr unit cell.

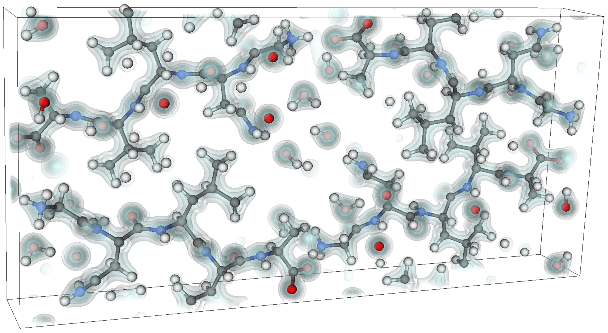

FIG. 5: **Unit cell**
Visualization of the electronic groundstate of the unit cell in the GNLVS system.

We compared MD runs on the unit cell starting from the electronic structure computed by the $O(N)$ solver with a geometric initial guess, and by the $O(N^3)$ solver, and saw that the maximum orbital center movement in the first MD step in the $O(N)$ case was 0.2358 Bohr, and in the $O(N^3)$ case was 0.2849 Bohr. This demonstrates that the groundstate computed by the $O(N)$ solver, given a geometric initial guess, is comparable in stability to the $O(N^3)$ groundstate at the start of an MD simulation.

[1] DR Bowler and Tsuyoshi Miyazaki. Methods in electronic structure calculations. *Reports on Progress in Physics*, 75(3):036503, 2012.

[2] J.-L. Fattebert and F. Gygi. Linear-scaling first-principles molecular dynamics with plane-waves accuracy. *Phys. Rev. B*, 73:115124, Mar 2006.

[3] Jean-Luc Fattebert and François Gygi. Density functional theory for efficient ab initio molecular dynamics simulations in solution. *Journal of computational chemistry*, 23(6):662–666, 2002.

[4] François Gygi, Jean-Luc Fattebert, and Eric Schwegler. Computation of maximally localized wannier functions using a simultaneous diagonalization algorithm. *Computer physics communications*, 155(1):1–6, 2003.

[5] W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. *Phys. Rev.*, 140:A1133–A1138, Nov 1965.

[6] Nicola Marzari and David Vanderbilt. Maximally localized generalized wannier functions for composite energy bands. *Phys. Rev. B*, 56:12847–12865, Nov 1997.

[7] Daniel Osei-Kuffuor and Jean-Luc Fattebert. Accurate and scalable o(n) algorithm for first-principles molecular-dynamics computations on large parallel computers. *Physical review letters*, 112(4):046401, 2014.

[8] Alice Soragni, Shida Yousefi, Christina Stoeckle, Angela B Soriaga, Michael R Sawaya, Evelyne Kozlowski, Inès Schmid, Susanne Radonjic-Hoesli, Sebastien Boutet, Garth J Williams, et al. Toxicity of eosinophil mbp is repressed by intracellular crystallization and promoted by extracellular aggregation. *Molecular cell*, 57(6):1011–1021, 2015.