High-dimensional machine learning using multi-level parallellisation

Background

The Alexandria Chemistry Toolkit (ACT) is a software package under development at the department of cell and molecular biology, Uppsala university. Its goal is to refine a physical model of organic molecules, a so-called force field[1]. The foundation for the toolkit is an open-data repository, the Alexandria library[2], available from Zenodo[3]. A first version of the code has been described in an application where some atomic properties were optimized to reproduce molecular electrostatics[4]. Now we are preparing the ACT to refine the remaining terms and to complete an entire force field from scratch[1]. To do so, it would be interesting to introduce more powerful algorithms.

Computational challenges

A force field that can model for instance a protein in solution will need on the order of 1000 parameters. Some of these parameters are tightly coupled to more or less independent physical observables, which simplifies the problem somewhat. The algorithm used in ACT is a Markov Chain Monte Carlo (MCMC) based on a stochastic simulation in parameter space is performed (Figure 1). That means a χ^2 function is constructed that is to be minimized. The MCMC algorithm simulates parameters at a certain "temperature" which means the system as a whole can cross "energy" barriers. This avoids getting stuck in the nearest local minimum.

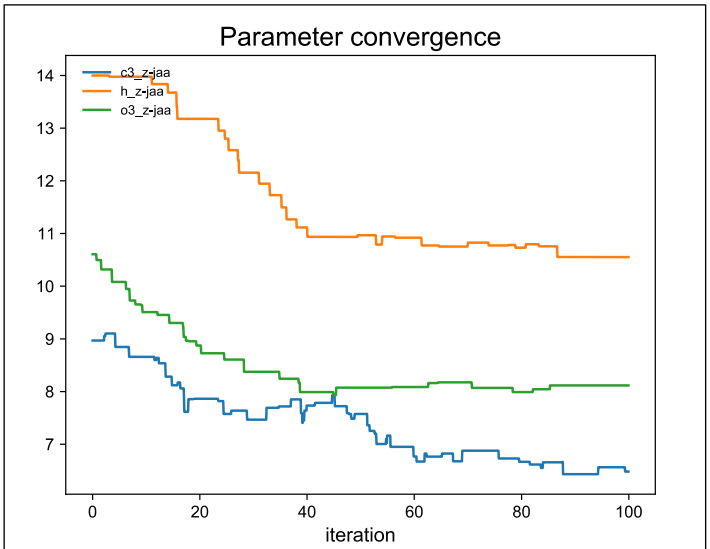

Figure 1 Example MCMC simulation for three parameters.

Nevertheless, due to the high complexity of the parameter space, finding a global minimum of the parameters is difficult. Therefore, the aim of this project is to introduce a genetic algorithm (GA) into the ACT. Presently, we perform each MCMC simulation on all cores of a node in a HPC environment, for instance, 32 cores at tetralith. We do a number of independent such simulations and monitor their progress. As a result, we find that some of the simulations do better for some parts of chemical space than others. Therefore combining e.g. 10 simulations into one, governed by a GA might improve overall results, allowing us to come closer to the global minimum.

About the ACT

The Alexandria Chemistry Toolkit is a large C++ code, forked off from the GROMACS simulation engine some years ago. There are about 40,000 lines of code in the ACT that are not in the parent package, GROMACS. The code is strictly object-oriented, but does not use

the most complex features of C++. At present the parallelization scheme used is based on the message passing interface (MPI), which allows for fine-grained control of the parallelism. This method is very portable and allows achieving high parallel efficiency but at the expense of coding time and complexity.

References

1. van der Spoel D: **Systematic design of biomolecular force fields**. *Curr Opin Struct Biol* 2021, **67**:18–24.
2. Ghahremanpour MM, Van Maaren PJ, Van Der Spoel D: **Data Descriptor: The Alexandria library, a quantum-chemical database of molecular properties for force field development**. *Sci Data* 2018, **5**:180062.
3. Ghahremanpour MM, van Maaren P, van der Spoel D: **Alexandria Library**. 2017, doi:10.5281/ZENODO.1170597.
4. Ghahremanpour MM, Maaren PJ van, Caleman C, Hutchison GR, van der Spoel D: **Polarizable Drude Model with s-Type Gaussian or Slater Charge Density for General Molecular Mechanics Force Fields**. *J Chem Theory Comput* 2018, **14**:5553−5566.