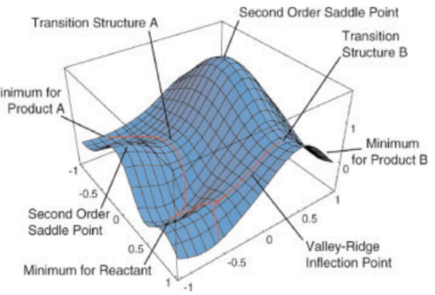
Lab.: Inria Sophia-Antipolis-Méditerranée
ABS http://team.inria.fr/abs
Ecuador, http://www-sop.inria.fr/tropics
email: Frederic.Cazals@inria.fr (ABS)
email: Laurent.Hascoet@inria.fr (Ecuador)

Master internship proposal



Locating critical points on potential energy surfaces of molecular systems requires partial derivatives. From [Sch03].

Killing two birds with one stone: algorithmic differentiation of C/C++ force fields used in biophysics

Keywords: algorithmic differentiation, software engineering, molecular simulation, multivariate energy functions.

Context: Structural biology is concerned by the study of the relationship between the structure and the function of biomolecules in general and of proteins in particular. In studying this relationship, a central problem is the description of the potential energy landscape (PEL) of the molecular system, namely the function associating a potential energy to each conformation. Indeed, PEL are defined over high dimensional conformational spaces – a molecule with n atoms has $3n$ Cartesian coordinates. In describing the PEL, of paramount importance are critical points, in particular local minima and (index one) saddles, as these can be used to derive the thermodynamic and kinetic properties of the system [Wal03].

Goals: A key requirement of PEL exploration algorithms is the ability to perform gradient descent to minimize the energy of the system–see the original basin-hoping algorithm [LS87] and its recent improvements by the Inria ABS project-team [RDRC16]. Given the high dimensionality of conformational spaces, the manual computation can be daunting and error prone. An efficient and elegant alternative is algorithmic differentiation (AD) [HP13] (http://www-sop.inria.fr/tropics/tapenade.html), a collection of techniques computing partial derivatives when the functions are given in the form of a computer program. Currently, the Tapenade system developed by the Inria Ecuador project-team handles Fortran and C code.

In this context, the goal of this internship will be twofold:

1. The partial derivatives of standard force fields used in molecular modeling, namely (AMBER, CHARMM, MARTINI) have been computed analytically using computer algebra systems. This manual step is a clear limitation if one wishes to change these force fields, by adding / removing terms. Also, these force fields are embedded in complex software packages, which makes their comparison difficult.

The first goal will therefore be to compute first and second order derivatives of C implementations of these force fields, using algorithmic differentiation tools. This step will make it possible to update the force fields easily, and also to compare their performances in a coherent software framework.

2. C++ generic programming provides a versatile framework to design generic code using so-called templates. Templates are of special interest for PEL, as the same exploration algorithm can be instantiated using different force fields. However, automatically handling generic C++ code is challenging for AD tools. The second goal will therefore be to specify a subset of C++ amenable to AD. This functionality will be used to process generic implementation of these force fields, currently being integrated within the Structural Bioinformatics Library (http://sbl.inria.fr), a reference library for computational structural biology developed by ABS. Phrased differently, this new ability will make it possible to use all classical force fields in the context of a generic C++ library (the SBL), whose tenet is precisely to decouple algorithms and the biophysical models used to represent molecules, using C++ templates.

On a more general perspective, AD of a well specified subset of C++ code will enlarge the community of users of Tapenade.

Difficulties: As explained above, for a molecule with n atoms (n in the range 50...10,000), a conformation is given by $3n$ Cartesian coordinates, or equivalently $d = 3n - 6$ internal coordinates (bond lengths, valence angles, dihedral angles). Thus, the potential energy is a real valued function of d variables. This function is generally defined by of the order of hundreds of parameters, which correspond to atom types and make up the different contributions to the potential energy. In this context, the difficulties to compute partial derivatives in a fast and numerically robust way are twofold:

- *Computational efficiency and numerical accuracy.* For systems with a large number of degrees of freedom, factoring out the code to be differentiated so as to obtain efficient calculations is challenging. Likewise, since floating point numbers are used, the code design must be such that rounding errors are minimized, another ambitious endeavor.

- *Data structures and programming language issues.* As mentioned earlier, Tapenade performs algorithmic differentiation for Fortran and C code. When C++ code is used, the question of the data structures used (iterators, direct access) comes into the play. Similarly, the naming of functions raises difficulties (prefixing by ::, namespaces). We wish to specify a subset of C++, from which we shall automatically generate C code directly amenable to AD by Tapenade. This is another non trivial task.

Background. Master in (theoretical) computer science or applied mathematics or bioinformatics/biophysics.

Conditions. CDD SMIC.

References

[HP13] L. Hascoet and V. Pascual. The tapenade automatic differentiation tool: principles, model, and specification. *ACM Transactions on Mathematical Software (TOMS)*, 39(3):20, 2013.

[LS87] Z. Li and H.A. Scheraga. Monte carlo-minimization approach to the multiple-minima problem in protein folding. *PNAS*, 84(19):6611–6615, 1987.

[RDRC16] A. Roth, T. Dreyfus, C.H. Robert, and F. Cazals. Hybridizing rapidly growing random trees and basin hopping yields an improved exploration of energy landscapes. *J. of Computational Chemistry*, 37(8):739–752, 2016.

[Sch03] H.B. Schlegel. Exploring potential energy surfaces for chemical reactions: an overview of some practical methods. *Journal of computational chemistry*, 24(12):1514–1527, 2003.

[Wal03] D. J. Wales. *Energy Landscapes.* Cambridge University Press, 2003.