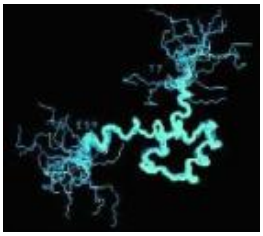
Lab.: Inria Sophia-Antipolis-Méditerranée
Group: Algorithms-Biology-Structure
Supervisor: Frederic.Cazals@inria.fr
Web: `http://team.inria.fr/abs`



Twenty conformations of a protein, showing a well defined core and disordered tails. Nobel lecture of K. Wütrich, 2002.

PhD thesis proposal

Deep geometric analysis of high dimensional potential energy surfaces

Keywords: high-dimensional spaces, dimensionality reduction, importance sampling, deep learning, concentration phenomena, molecular simulation.

Context: Understanding biological functions at the molecular/atomic level requires computing average properties of very large (in fact continuous) ensembles of molecular conformations. As of today, except in rare cases where massive molecular dynamics or Monte Carlo simulations are used [1], these time scales remain out of reach.

However, recent work has shown that it is not only possible to learn important features on energy surfaces of such systems [2], but also to compute efficiently so-called partition function for such surfaces [3], by exploiting geometric features is regions concentrating the (probabilistic) mass.

Goals: Since each atom has three Cartesian coordinates, the conformation of a molecule with n atoms is described by $3n$ coordinates and $d = 3n - 6$ degrees of freedom or dof. (One removes the dof for 3D translations and rotations, whence $3n - 6$.) All properties of the molecule are thus determined by an energy surface of dimension d, typically several thousands, called the potential energy surface (PES). Despite intensive research, the exploration and the characterization of such surfaces is currently an open problem.

From a computer science / applied mathematics standpoint, the problems faced as well posed, though. From the structural standpoint, stable structures correspond to local minima of the PES. From the thermodynamic viewpoint, meta-stability is characterized by occupancy probabilities (for Boltzmann's distribution) of selected basins of the PES. Finally, dynamics may be modeled by a Markov state model involving metastable states.

The goal of this PhD thesis will be to develop novel models for these high dimensional PES, via a process mixing dimensionality reduction [4] and importance sampling [3]. The developments foreseen also bear connexions with selected analysis currently developed in the field of deep learning, where it has been realized that energy based models govern the behavior of e.g. classifiers [5].

Software developments will be integrated to the Structural Bioinformatics Library (http://sbl.inria.fr), a state-of-the-art environment providing both low level methods (in generic C++) and specific applications in molecular modeling.

Background. Master in theoretical computer science, or applied mathematics, or physics, or bioinformatics/biophysics.

References

[1] D. E. Shaw, P. Maragakis, K. Lindorff-Larsen, S. Piana, R. O. Dror, M. P. Eastwood, J. A. Bank, J. M. Jumper, J. K. Salmon, Y. Shan, and W. Wriggers. Atomic-level characterization of the structural dynamics of proteins. *Science*, 330(6002):341–346, 2010.

[2] Frank Noé, Simon Olsson, Jonas Köhler, and Hao Wu. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science*, 365(6457):eaaw1147, 2019.

[3] A. Chevallier and F. Cazals. Wang-landau algorithm: an adapted random walk to boost convergence. *J. of Computational Physics*, NA(NA), 2020.

[4] R.R. Coifman, S. Lafon, A.B. Lee, M. Maggioni, B. Nadler, F. Warner, and S.W. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *PNAS*, 102(21):7426–7431, 2005.

[5] Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. In *ICLR*, Addis Ababa, 2020.