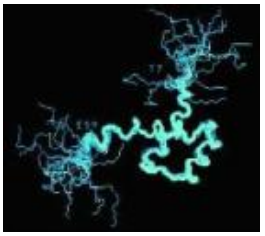
Lab.: Inria Sophia-Antipolis-Méditerranée
Group: Algorithms-Biology-Structure
Supervisor: Frederic.Cazals@inria.fr
Web: `http://team.inria.fr/abs`



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

Master internship proposal

Combining dimensionality reduction and feature detection, with applications to molecular data analysis

Keywords: molecular simulation, high-dimensional spaces, dimensionality reduction, feature detection, novelty detection, clustering, local models.

Context: Molecular simulations based on molecular dynamics and Monte Carlo simulations generate massive data in a sequential manner – that is one gets a time series of conformations, ranging in size from millions to billions. These data are inherently high-dimensional, as a molecule with n atoms (say $n = 5000$ for a medium sized protein) has a configuration space of dimension $3n$ in cartesian coordinates.

To exploit these data, a classical strategy is dimensionality reduction (DR) on the pooled dataset, so as to find out a small number of collective coordinates accounting for macroscopic properties of the system. For example, in using principal components analysis, these collective variables are linear combinations of the original coordinates. Naturally, the non-linear nature of the data prompted the application on non linear DR methods, such isomap [DMS$^+$06] and diffusion maps [CLL$^+$05, RZMC11, NAKH14].

Goals: Currently, DR methods are used on the pooled dataset. The goal of this internship will be to design more local DR methods, by combining ideas from dimensionality reduction, feature detection [MJ15], and discrepancy analysis [CL15, LC15]. These methods will also be aware of distance concentration phenomena [CTP11], which may jeopardize the analysis carried out. In doing so, the goal will be to single out regions of the configuration space where specific variables are at play. In a second step, these models will be used to monitor the novelty production of a simulation, to check whether new regions of the conformational space are being discovered along time.

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

Misc. Ideally, the MSc will be followed-up by a PhD thesis.

References

[CL15] F. Cazals and A. Lhéritier. Beyond two-sample-tests: Localizing data discrepancies in high-dimensional spaces. In P. Gallinari, J. Kwok, G. Pasi, and O. Zaiane, editors, *IEEE/ACM International Conference on Data Science and Advanced Analytics*, Paris, 2015. Preprint: Inria tech report 8734.

[CLL$^+$05] 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.

[CTP11] M. Ceriotti, G. Tribello, and M. Parrinello. Simplifying the representation of complex free-energy landscapes using sketch-map. *PNAS*, 108(32):13023–13028, 2011.

[DMS$^+$06] P. Das, M. Moll, H. Stamati, L. Kavraki, and C. Clementi. Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction. *PNAS*, 103(26):9885–9890, 2006.

[LC15] A. Lhéritier and F. Cazals. A sequential non-parametric two-sample test. *Submitted*, 2015. Preprint: Inria tech report 8704.

[MJ15] J.W. Mueller and T. Jaakkola. Principal differences analysis: Interpretable characterization of differences between distributions. In *Advances in Neural Information Processing Systems*, pages 1693–1701, 2015.

[NAKH14] L. Nedialkova, M. Amat, I. Kevrekidis, and G. Hummer. Diffusion maps, clustering and fuzzy markov modeling in peptide folding transitions. *The Journal of Chemical Physics*, 140, 2014.

[RZMC11] M. Rohrdanz, W. Zheng, M. Maggioni, and C. Clementi. Determination of reaction coordinates via locally scaled diffusion map. *J. of Chemical Physics*, 134(12), 2011.