ToMATo for protein conformation

October 4, 2021

Abstract

The goal of this project is to analyze protein conformations using mode-seeking techniques, in order to detect metastable states and their proximity relations. The difficulty of recovering the metastable states stems from the fact that the clustering occurs in fairly high dimension (n can be of ther order of the hundreds or thousands), with data that are not sampled along linear structures and clusters that are nonconvex. In this project we will use the topology-based method ToMATo to cluster the conformations

- Collect the data:

  https://drive.google.com/file/d/1eMASNaHp4tcQrbJEcPOUhzpxHzjnf9mr/view?usp=sharing

  It contains the set of alanine dipeptide conformations: 3 coordinates per atom, 10 atoms per conformation, 1 atom per line (so 10 lines per conformation, seen as a 30-dimensional point).

- Implement and compute the RMSD distance matrix between the 30-dimensional conformations (RMSD = Root Mean Square Deviation). See

  https://en.wikipedia.org/wiki/Root-mean-square_deviation_of_atomic_positions

- Use MDS to produce an embedding of the data set in 2 dimensions for visualization.

- Get familiar with `Gudhi`'s implementation of ToMATo, e.g. try it out on the toy examples seen in class then play around with the parameters.

- Try applying ToMATo to the computed RMSD distance matrix. Beware that the data set is huge so you may want to consider applying it to subsamples of the data.

- Hopefully you will be able to recover the same kind of result as in this article:

  https://hal.inria.fr/inria-00389390/document

  Read this article (section 6.3) to get some more insight into the data and its interpretation.