Geometric Methods in Data Analysis

Bastien BRIER - Andrei CONSTANTINESCU
bastien.brier@student.ecp.fr - andrei.constantinescu@student.ecp.fr

February 26, 2017

Mode-seeking for detecting metastable states in protein conformations

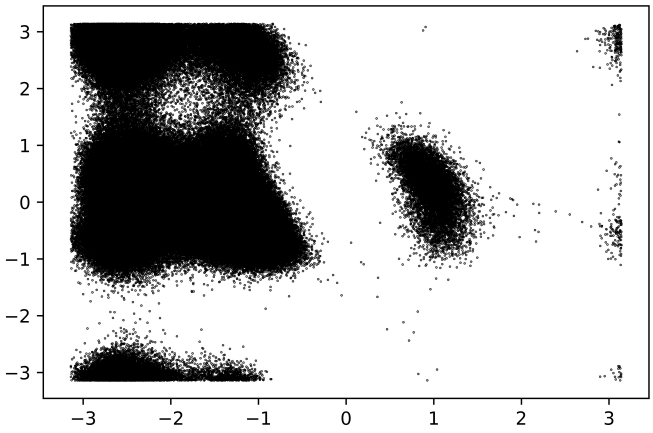
1 Introduction

In this project, we aim to analyse protein conformations using mode-seeking techniques. The goal is to find the different metastable states and their proximity relations. For protein analyses, the study of these states is primordial : they are relatively stable and the transition between them is highly unlikely. Thus, Markovian models can be used to describe long-timescale transitions between states. The main difficulty in clustering this data into different states is that the clustering happens in really high dimension: that is why mode-seeking techniques (like ToMATo) are really useful to perform a relevant clustering. Our dataset is composed of 14,207,380 different atom 3-d coordinates - 10 consecutive atoms forming a conformation. We can therefore see our dataset as 1,420,738 different 30-dimension conformations.
We will first compute the RMSD (Root Mean Square Deviation) of atomic positions between the different conformations, and then pass the resulting matrix to our mode-seeking algorithm, ToMATo.

2 Data Exploration

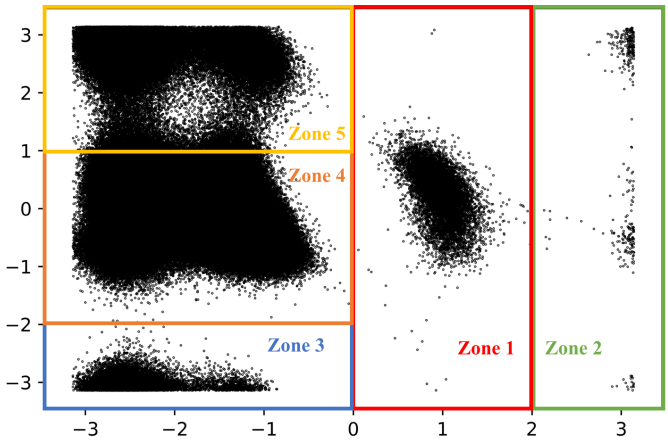
In addition to the alanine-peptide dataset, we are provided with the 'dihedral' dataset, which is the previous dataset conformations projected in 2 dimensions. This was possible because the alanine-dipeptide in fact have only two degrees of freedom, parameterized by the two angles ϕ and ψ.

Figure 1: Input point cloud after projection to the (ϕ, ψ) domain



Then, we used an eye-balling technique on the plotted data to assess zones of interest from which we would sample our points. This was to ensure that the cluster consisting of a relatively small number of points would not be missed in the sampling. If we just sampled randomly from the whole dataset, the cluster with the largest number of points would have eaten up all the other clusters and rendered the analysis unfruitful.

Figure 2: Division of the input space in 5 subspaces



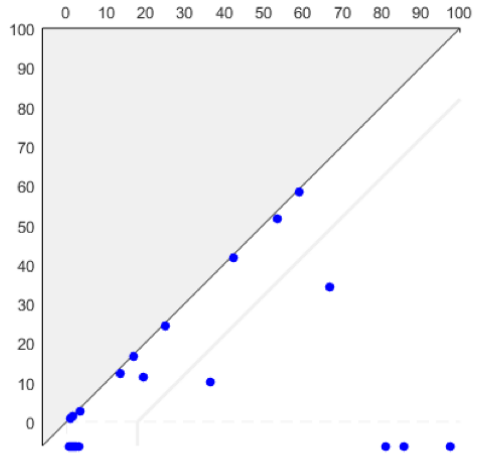
We will first proceed with an example with 5350 points evenly taken in the 5 zones presented and then generalize our model.

3 RMSD Computation and First Example

In order to compute the RMSD, we use the existing code indicated in the instructions : https://github.com/pandegroup/IRMSD.
In order to use this package, we had to reshape our dataset. We reshaped it axis-major order, which means we had to transform our data to a matrix with shape (number of conformations, number of coordinates, number of atoms per conformations), i.e. in our case, a (1420738, 3, 10) matrix.
We then have to pad out our structure by adding zeros in order for our matrix to have a "padded" number of atoms that is a multiple of four. Finally, we can create our conformations object and calculate the RMSD, beginning with our 5350 sample points.

This gives us a n * n matrix, where n is the number of points chosen in the sample: here 5350. Such a matrix can be easily given to ToMATo, and computations are fast. The difficulty relies in choosing the right hyperparameters: the number of neighbors (because we are also estimating the density) and the Rips radius. After a few trial and errors, we tuned these values to 20 and 0.3 respectively. We mainly judged of the rightness of these hyperparameters by looking at the persistence diagram, and ensuring it made sense: i.e. not too many points, indicated a too fine-grained approach, nor too few indicating a too whole-grained one.

Figure 3: Persistence diagram obtained

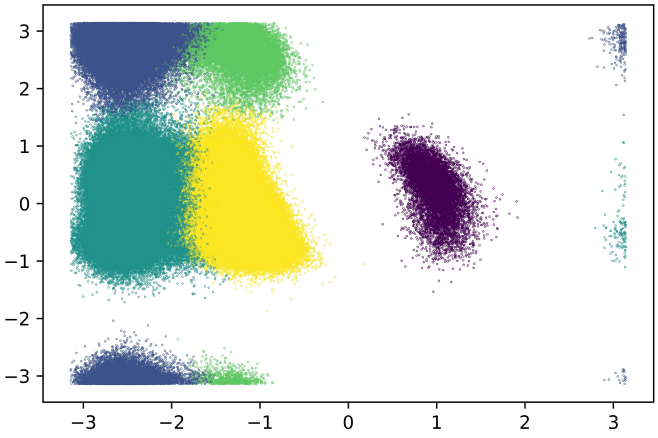


From this diagram, we can deduce that 5 clusters is a good cut-off. If we ploted the sampled point into their projection on two dimensions, and colored them by their cluster, this gave us a the result we expected. We now need to extend and generalize these results.

4 Generalization and Results

We now want to use ToMATo on our whole dataset. As computing and storing the results of a (1420738, 1420738) matrix is really long and expensive, we decided to select 1000 reference points evenly distributed in the 5 subspaces we defined above. Moreover, as we had to use ToMATo on a Linux virtual machine, our computational power was very limited. We then had to divide our dataset in 100 evenly distributed smaller datasets. For this, we chose randomly 1/100th of the points of each subspace, to ensure that each sub-dataset is representative of the bigger one.
We used the same hyperparameters as above and a threshold tau of 40 or 90 depending on the samples. We obtained similar results in each of the dataset. We then merged the 100 results file and obtained the final clustering below.

Figure 4: Final clustering with ToMATo



The conformations tend to regroup in 5 different metastable states, hereby represented by the 5 different clusters. As the coordinates of our dataset in the 2D representation are angles in radiant, the extremities of the graph are in fact really close to each other, which explains this segmentation.

This work allowed us to get a good grasp of how ToMATo works and made us apply it to a real-life example with conclusive results. Possible next steps would be to explore the possibility of a soft segmentation, or to apply this algorithm to other datasets that also exist in high dimensions, where vanilla clustering algorithm like k-means, or even spectral clustering do not give relevant results.

5 References

[1] F. Chazal, L. Guibas, S. Oudot, and P. Skraba. Persistence-based clustering in riemannian manifolds. J. ACM, 60(6):1-38, 2013.
[2] J. Chodera, W. Swope, J. Pitera, and K. Dill. Long-time protein folding dynamics from short-time molecular dynamics simulations. Multiscale Modeling and Simulation, 5(4):1214-1226, 2006.