Exercise: Dimensionality Reduction and Embedding for Molecular Dynamics Trajectories

Qian-Yuan TANG
(tangqianyuan@gmail.com)

In the research of molecular simulations, to clearly describe a biochemical reaction, we need to reduce the motions and fluctuations of a large amount of degrees of freedom into a few reaction coordinates (collective coordinates). These reaction coordinates are low-dimensional descriptors that can clearly represent the progress along a reaction pathway. Usually, appropriate reaction coordinates can extract the most relevant global motions and the dynamically meaningful slow variables. Besides, with appropriate reaction coordinates, the efficient sampling in the molecular simulations, and the calculations of the potential of mean force (PMF) along the biochemical reaction process can be accomplished. Therefore, a systematic and automatic method for determining appropriate coordinates would be of great importance in molecular simulations.

However, when we have reduced the dimensionality of the simulation trajectories, there will be some information lost in projection. For example, conformations that located in the same basin in some projection may have little structural or kinetic similarity. Thus, finding a better way in representing different conformations would be of great help to extend the low-dimensional descriptions of a biochemical process. To solve this problem, one may consider employing clustering algorithms by grouping the conformations with high structural similarity, but this cannot resolve the problem completely because the kinetic information is still missing. Therefore, scientists introduced the Markov State Models (MSMs) to simplify the simulation trajectory in a kinetically meaningful way. In a sense of Markovian description, a simulation trajectory should be understood as a time series of different staes, such as "…ABBABBCDABBBBA…", here, different letters denote for different states. Since the trajectory could be reduced as a sequence of conformations, similar to what we have learned about natural language processing (NLP) in this summer school, we could consider a low-dimensional embedding (such as word2vec) of such kind of sequences. This reduction maps the conformations into a low-dimensional manifold of features that maximizes the likelihood of preserving network neighboring properties (say, the kinetic and structural similarities between different conformations are preserved). Such a low-dimensional manifold could provide additional information than the traditional perspective of an energy landscape.

In this project, you will have a chance to use the algorithms (PCA, KPCA, autoencoders or word2vec) that introduced during this summer school to find a low-dimensional description for a long molecular simulation trajectory[1]. The project can be divided into two small exercises:

(I) Dimensionality Reduction

In this part of the exercise, you are going to reduce a long molecular dynamics trajectory by nonlinear methods (say, KPCA or autoencoders). In molecular dynamics simulation, a trajectory includes the snapshots of the atom coordinates of the molecule. More specifically, for protein molecules, the backbone dihdrals φ's and ψ's can be employed to determine the conformation of a protein. To reduce the dimensionality of a trajectory means to reduce the high-dimensional descriptions [$\varphi_1(t)$, $\psi_1(t)$, $\varphi_2(t)$, $\psi_2(t)$, …, $\varphi_{N-1}(t)$, $\psi_{N-1}(t)$] into low-dimensional descriptions, for example, a reduced 2D-description of the trajectory should be [$RC_1(t)$, $RC_2(t)$].

Compare the reaction coordinates determined by nonlinear methods with that determined by linear PCA and those order parameters which are commonly employed in protein studies such as RMSDs

[1] Molecular Simulation Trajectories Archive of a Villin Variant https://simtk.org/frs/?group_id=285 , or trajectories from DE Shaw's group.

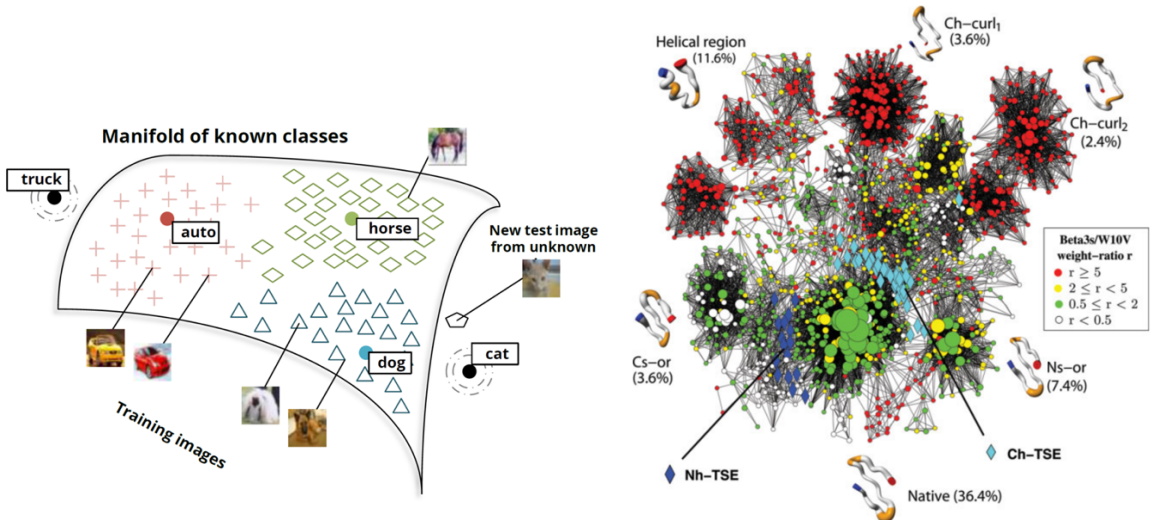
and the Q-values. Try to understand in what conditions or in what kinds of processes will nonlinear methods become essential.

Reference:

1. Wang J, Ferguson A L. Nonlinear reconstruction of single-molecule free-energy surfaces from univariate time series[J]. Physical Review E, 2016, 93(3): 032412. http://ferguson.matse.illinois.edu/resources/22.pdf
2. David C C, Jacobs D J. Principal component analysis: a method for determining the essential dynamics of proteins[J]. Protein Dynamics: Methods and Protocols, 2014: 193-226. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4676806/

(II) Network Embedding

When two conformations (or roughly speaking, two states) tend to transit from one to another, then we could say, the two conformations are "similar", then, when we preserve the similarities (local distances) between conformations, we can project high-dimensional data into a low-dimensional manifold (as illustrated in the figure below). Such kind of projection is just similar to what word2vec had done to human languages. Recently, there are a lot of algorithms have been developed to map social networks into low-dimensional manifold (such as deepwalk, LINE, node2vec, etc.). Many internet companies have employed the methods which is similar to word2vec to measure the similarities between different internet users. The conformational space of protein dynamics can also be recognized as a complex network (as shown below). Here, in this part of the exercise, we are going to reduce the dimensionality of simulation trajectories based on the similarities between the conformations. We would like to see whether such network embedding can provide us more information (e.g., the hierarchical structures in the folding landscape) than directly doing the reduction by KPCA or autoencoders, and we are going to reconstruct a model which is quite similar as MSM in a totally different way.



Reference:

1. Pande V S, Beauchamp K, Bowman G R. Everything you wanted to know about Markov State Models but were afraid to ask[J]. Methods, 2010, 52(1): 99-105. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2933958/
2. Perozzi B, Al-Rfou R, Skiena S. Deepwalk: Online learning of social representations[C]//Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2014: 701-710. https://pdfs.semanticscholar.org/bcea/4071cb8ad342618bf898a301025d9bd2c336.pdf
3. Grover A, Leskovec J. node2vec: Scalable feature learning for networks[C]//Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2016: 855-864. https://cs.stanford.edu/people/jure/pubs/node2vec-kdd16.pdf
4. 当机器学习遇上复杂网络：解析微信朋友圈 Lookalike 算法 http://www.36dsj.com/archives/75212