Introduction

Molecular dynamics simulation(MD) is a popular approach to understanding complex biological systems like human RNA polymerase II. This approach models systems of up to 10^6 atoms evolving according to the laws of classical mechanics, providing insights on complex motions governing the function of the system. The simulation data usually contains snapshots of the system captured at fixed intervals of time and the number of these snapshots can be up to 10^{10}. Its extremely cumbersome to extract relevant information from such large datasets, so methods of dimensionality reduction are called upon to separate a few physically meaningful combinations of the input atom coordinates from the irrelevant motions.

In this work I compare MD data of a human RNA pol II with a 15 base DNA template and a 8 base RNA. I employ several dimensionality reduction methods to partition the data in order to select the method that produces the most structurally tight clusters and at the same time capturing important dynamics.

Methods

a) *Input Data*

MD simulation data consist of 44 100ns unbiased trajectories originally obtained by G. Wang form X. Huang lab. A total of 44000 snapshots were chosen for this work

b) *Feature extraction*

The systems of hundreds of thousands of atoms are unwieldy so first we use our knowledge of the system to extract only the features of atoms that are close to the active site, we extract the following important structural elements from the overall system: DNA, RNA, Bridge helix(BH) and trigger loop (TL). MD simulations of biomolecules can usually be partitioned into fast irrelevant to the function oscillations and motions happening at slower timescales usually associated with changes in dihedral angles. Thus we choose our data set to be confined to torsion angles of the protein and nucleic acid backbones as well as side chain dihedrals and bases dihedrals. This reduces the dataset to just 295 features. We use the sin/cos of the actual angles to allow our data to be centered.

c) *Dimensionality reduction*

Principal components analysis (PCA) is a method that uses the eigendecomposition of the covariance matrix to find a linear combination of the input features that form vectors along which the variance of the data is maximized. Maximizing variance in this case is justified as the largest motions are usually the slowest, although this does not necessarily holds for every case and depends heavily on the type of structure being analyzed. For example large motions can occur in the dangling ends of the structure which while having little functional relevance will contaminate the covariance matrix.

Time-lagged independent component analysis (tICA) is a popular approach to addressing the drawbacks of PCA by working with time-lagged correlation matrix instead. The time lag is optimized to achieve the most variance explained by the least number of obtained independent components.

These two methods yield linear combinations of the input features as the suggested dimensions vectors. Non linear methods can better describe more complex transitions which are expected to appear in such intricate systems.

Out of the numerous methods of nonlinear embedding we will focus on the **local linear embedding** class in its regular form, the modified form and the hessian form. The number of neighbours for calculating linear neighbourhood was chosen to be 40. In addition a **spectral clustering** approach is also employed to capture the manifold structure. These methods are generally more computationally expensive, thus we choose to subsample the data by 25 times to allow for reasonable computational times.

d) Clustering

After the number of dimensions has been reduced, the data points can be clustered to explore the nature of the obtained manifold. The K-means algorithm provides adequate performance for splitting the data into predefined number of clusters. The number of clusters was chosen based on the visual inspection of density plots produced by PCA and tICA.

We examine the 2D tica projection density plot to discover several clearly visibly peaks. We manually count them and use that number for selecting the number of clusters for k-means partitioning for the rest of the project.

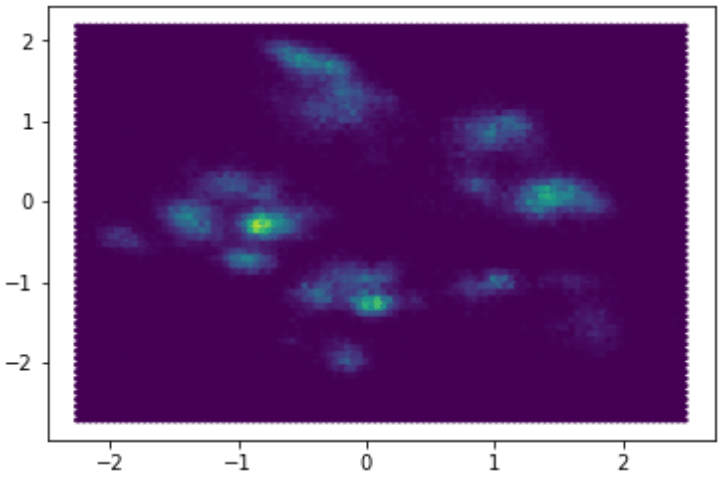


Fig.2 Density plot of 2D tICA projection

Root mean square deviation of atomic coordinates (RMSD) is a well defined measure of structural similarity. I use averages of RMSD within each cluster to select the method that produces the most structurally tight clusters.

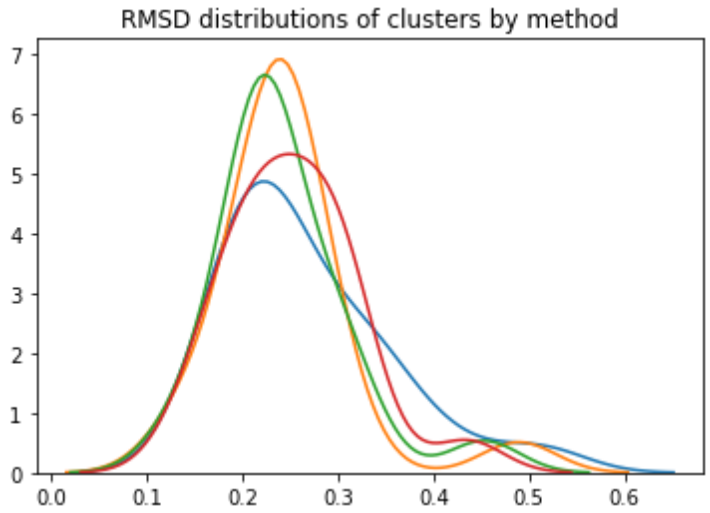
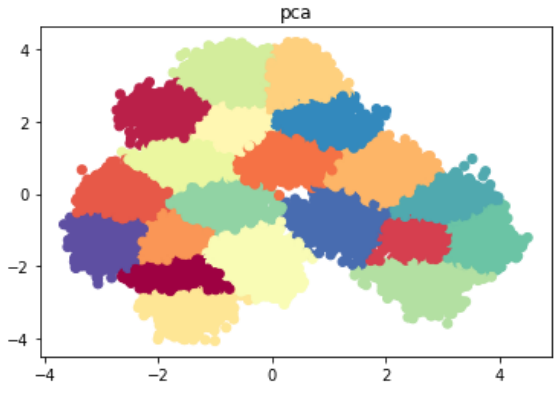
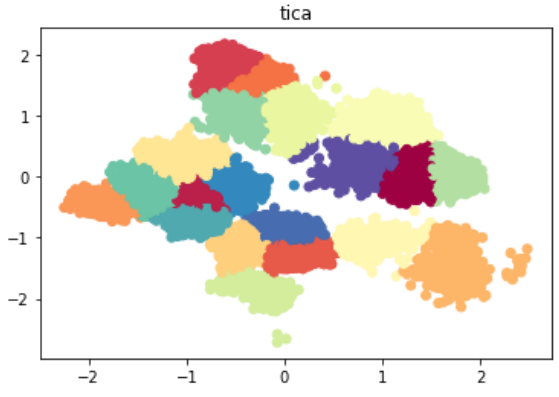


Fig 1. Mean pairwise RMSD distributions inside clusters

The peak at ~0.2 in RMSD distributions show that generally all clusters are structurally tight, with the least tight being the PCA (blue line). Hence we can use cluster centers for our analysis of structures.

2D projection	Structures of cluster centers
	
	

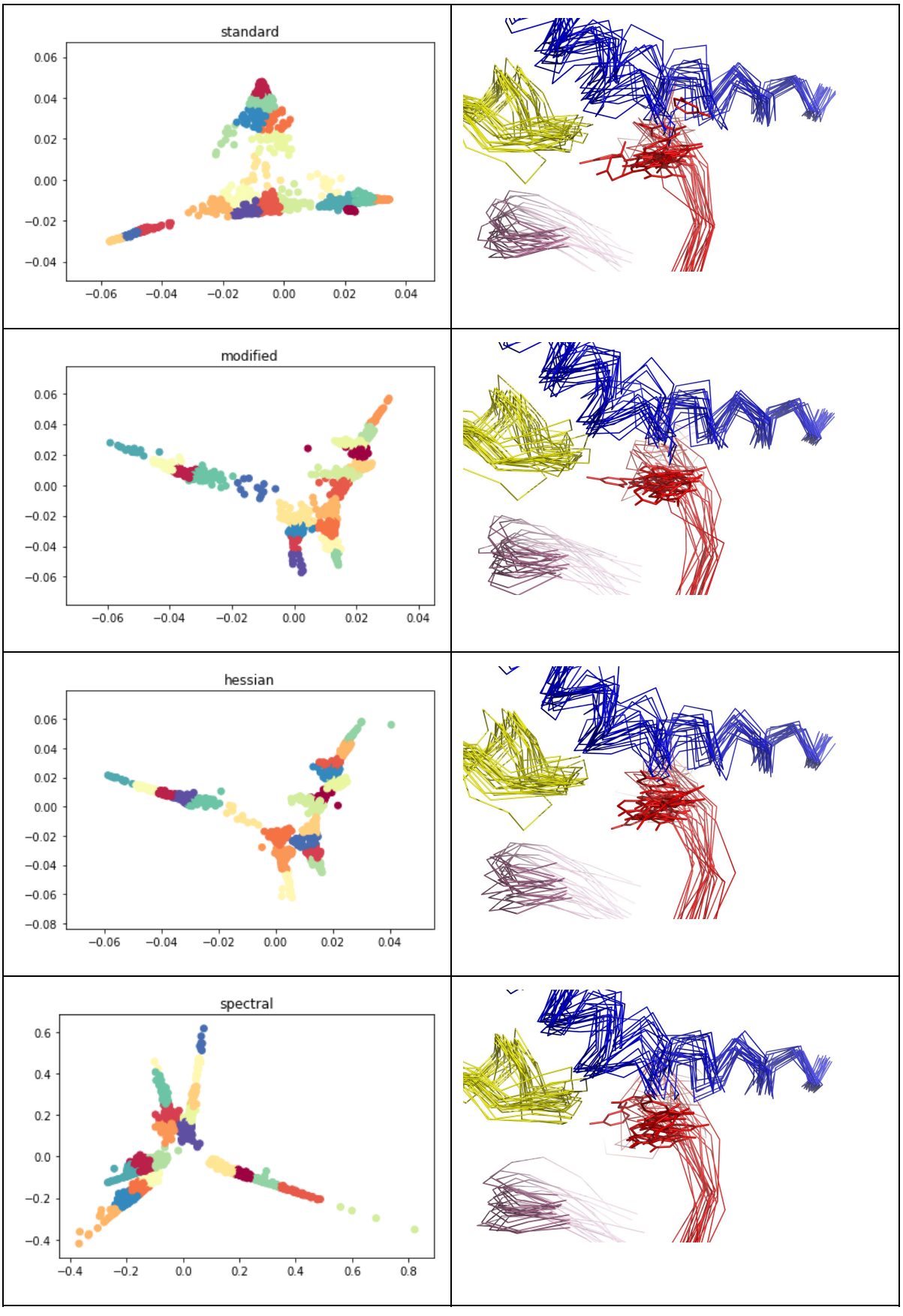

standard
modified
hessian
spectral

Fig 2. Analysis of clusters, 2D projections are depicted on the right, the structures are in the left column. Blue and yellow are the protein parts of the system: BH and TL respectively, pink and red are the RNA and DNA respectively. DC15 base is depicted with sticks.

The structural observation of the cluster centers show that PCA cluster centers have the most narrow distribution of BH(blue) and TL(yellow) backbone, which indicates that DC15 (red sticks) is a critical base in this system as it governs the positioning of the incoming NTP to be incorporated by the RNA polymerase into the growing RNA chain. The clustering that captures the most diversity in its position should provide the best explanation of the dynamics. By visual inspection all methods resolve several different positions of this nucleotide, except hessian LLE, which generally captures only a single position.

Conclusion

Currently less popular methods for partitioning MD simulation data provide reasonable partitioning of the trajectories. A more careful and robust analysis is needed to ensure that the methods considered here are indeed fit for further use, for example, to construct and study kinetic models such as Markov State Models.