Protein Structure Clustering

Zhiyuan Song
Astronomy and Physics Department
Clemson University
Clemson, SC, 29634
zhiyuas@clemson.edu

Abstract

Different clustering methods were applied to differentiate protein structures with respect to the biochemical features, and further their atomic positions. PCA was applied for the dimension reduction of the features. k-means and spectral clustering (SC) can distinguish the structures though the latter performed less efficiently. Compared to hierarchy clustering (HC), both k-means and SC were sensitive to the initial selection of the centroids but easy to visualize the clustering pattern if features' dimension were less than 3 or reduced under 3 by PCA. HC result exhibited the significant stability though its clustering pattern was not straightforward to observe.

1 Introduction

A protein usually has a favorable and stable structure for specific identifications by other functional molecular groups during metabolism process. And it also undergoes reversible structural changes in performing its biological functions. By applying computational techniques,such as dynamics simulations [], the protein structures can be simulated and sampled with predictable Cartesian coordinates and momenta at any certain moment.
Consequently, biophysicists can predict protein structures in different scenarios. Although a protein is able to have multiple configurations under even same physical conditions, some of them are more possible to occur than others due to their lower free energies. Theoretically, the effective free energy of a protein structure can be evaluated using probability distribution as functions of biochemical features. Therefore, the problem becomes how we cluster protein structures using the features obtained from computational dynamic simulations and select the most typical ones to represent the entire cluster.
Multiple approaches were applied to achieve structural clustering, including dimension reduction with PCA, k-means, SC as well as HC. The data was obtained from my protein dynamics simulation results. Some essential features of the target protein were computed based on the atomic coordinates, including root mean square deviation (RMSD), radius of gyration (Rg), secondary structures and potential energy.

2 Data Collection

Aβ42 was selected to be the protein for clustering analysis. It is an intrinsic disordered protein that has no stable structure in the solution and is composed of 42 amino acids. Biochemical features of the Aβ42 were calculated from its computational simulations using atomic coordinates. 10,000 samplings were collected. RMSD, Rg, secondary structures and potential energy composed the features. Specifically, RMSD is used as a quantitative measure of similarity between two protein structures, defined as

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

$$\mathrm{RMSD} = \sqrt{\frac{1}{\mathrm{N}} \sum_{i=1}^{n} \|\mathbf{v}_i - \mathbf{w}_i\|_2^2} \tag{1}$$

Where $\mathbf{v}_i$ and $\mathbf{w}_i$ denote the position vectors of atom i in two different structures. Besides, N refers to the number of atoms of the protein. More importantly, structural alignment was taken into account before RMSD calculation, which is to minimize RMSD between a structure and a reference. Accordingly, the coordinates of the structure will be transformed to obtain such an intrinsic dissimilarity.

Rg describe the dimensions of a poly-peptide chain and defined as

$$\mathrm{Rg} = \sqrt{\frac{1}{\mathrm{N}} \sum_{i=1}^{n} \|\mathbf{p}_i - \bar{\mathbf{p}}\|_2^2} \tag{2}$$

Where N refers to the number of atoms of the protein. $\mathbf{p}_i$ and $\bar{\mathbf{p}}$ denote the position vectors of atom i and mean of the positions, respectively.

Secondary structures of a protein for amino acid components, including helices and β sheets, were determined by DSSP algorithm []. Moreover, the number of amino acids in a certain secondary structure (either a helix or a β sheet) was treated as a feature (ranges from 0 to 42). Besides, potential energy of the protein was evaluated based on the simulation force field [] and extracted to be the fifth feature. In total, 10,000 samples were involved, of which the sequence started from $0^t h$ to $9999^t h$. The features of the data were standardized before clustering because the features have different units and scales.

Furthermore, all atomic positions were extracted from the simulations to be features directly (381 atoms). The coordinates were transformed based on structural alignment with respect to the first structure. Besides, pair-wise RMSD for the entire samplings ($n(n-1)/2$) were calculated in order for HC.

3 Method

Euclidean distance (l_2 norm) was used in k-means and SC. Random seed = 0 was set to initialize the centroids in k-means in order for reproductions. Adjacency Matrices in SC were determined using nearest-neighbors method with 10. K-means was applied again after taking the eigen-decomposition on the Laplacian Matrices in SC. Owing to huge computational expenses of the eigen-decomposition on Laplacian Matrices, parallel job computations were used.

4 Clustering Analysis

First, RMSD and the number of β sheets were selected for k-mean clustering because they are typically the more significant features of a protein structures, as shown in **Fig.1A**. According to the data distribution, the samples appeared to be not separable. Consequently, a binary clustering was determined for k-mean on the two selected features. The blue structure located in highly distributed region according to the data distribution. While the orange structure exhibited a significant dissimilarity with the former. To avoid the influence from manual selection of the features, PCA was applied for high dimensional feature sets and able to reduce the dimension to 2 for direct visualization in 2D plots. Both k-mean and SC were applied and the clustering results were show in **Fig1B&C** as well as the data distribution. In the case of PCA + k-mean, a binary clustering was determined as well due to data distributions. However, the two centroid structures were similar. Differently, 4 clusters were determined because a cluster including data located in the central highly populated region was expected. We can observe that the green cluster meet our expectation. 4 representative structures were distinctive, inferring PCA + SC performed better than PCA + k-means. As for computational expenses, it took much less time for k-means than SC even if SC incorporated with parallel job computations. asdasdasd

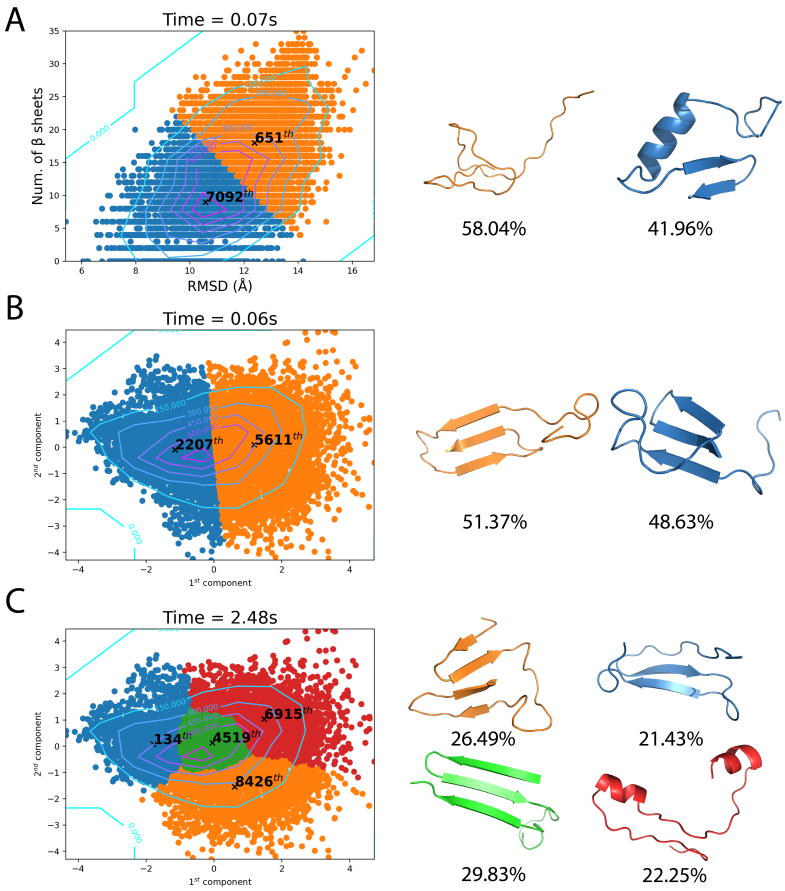

Figure 1: **Feature-based k-mean and SC Analysis** A) k-means with respect to RMSD and the number of β sheets. B-C) k-means and SC with respect to 2 principal components of all 5 features reduced by PCA. Data distributions were displayed using contours. Cluster centroids were colored and labeled with the corresponding sample sequence. The centroid structures showed together with their populations on the panel right.

5 Methods Comparison

6 Conclusion

References

7 Appendix

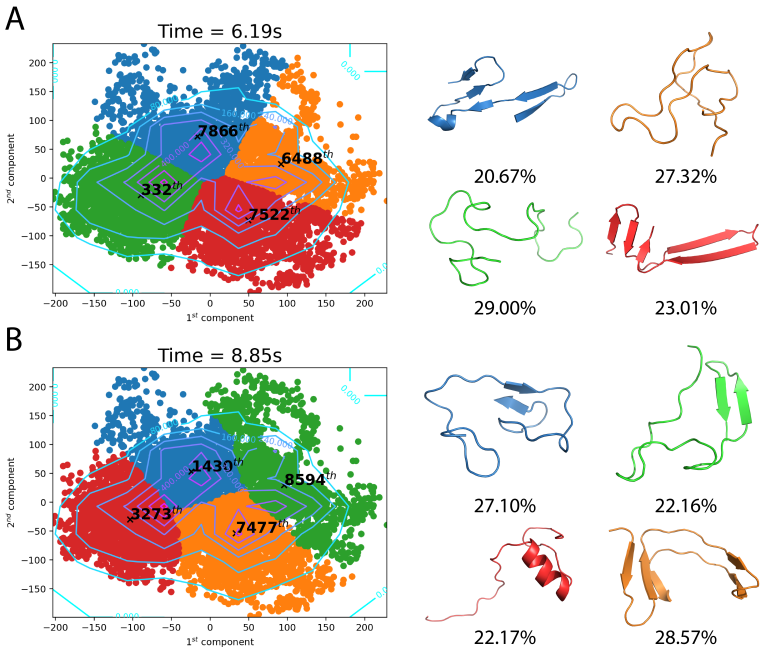

Figure 2: **Feature-based k-mean and SC Analysis** A) k-means with respect to RMSD and the number of β sheets. B-C) k-means and SC with respect to 2 principal components of all 5 features reduced by PCA. Data distributions were displayed using contours. Cluster centroids were colored and labeled with the corresponding sample sequence. The centroid structures showed together with their populations on the panel right.