Protein Conformation Classification

Song, Zhiyuan
CUID: C15433990

Motivation

A protein usually undergoes reversible structural changes in performing its biological function. The alternative structures of the same protein are referred to as different conformations, and transitions between them are called conformational changes. By applying computational techniques, the protein conformations can be simulated and sampled with predictable cartesian coordinates at any certain moment. Although a protein is able to have multiple conformations, some of them are more favorable than others due to their lower free energies. Theoretically, the effective free energy of a protein conformation can be evaluated using probability distribution as functions of certain features of protein structures. Therefore, the problem becomes how we classify protein conformations using relevant features obtained from computational dynamic simulation.

Method

Multiple cluster approaches can be applied to achieve conformational classification, including dimension reduction with PCA or LDA, *K*-means, spectral clustering as well as SVM. More importantly, it is necessary to compare their outcomes and investigate the best method to perform the classification. Besides, hierarchical clustering has been exploited in current research work and can serve as benchmarks for other classification.

Intended Experiments

I will use data from my current research work, which contains coordinates of all atoms of a certain protein derived from equilibrated simulation sampling. One way is to calculate some comprehensive quantities of a protein structure first, *e.g.*, root mean squared value (RMSD), radius of gyration (RG), and secondary structure content. Then, I will use these quantities as high-dimensional data sets to classify protein conformations. The dimension reduction approaches can be applied to exclude possible unessential features. Alternatively, I will treat coordinates of all atoms of the protein as features to distinguish its conformations. Such an attempt will require Structural alignment of all protein samplings, which is similar to centering data points beforehand. Both strategies will be examined with multiple clustering methods using the same protein structure samplings. Moreover, I will compare their outputs as well as the result from the method I'm currently using (hierarchical clustering). The ultimate goal is to find out a best way to classify protein conformation and apply it in my future research work.