BMI 203 Winter Quarter 2021

Homework Assignment 2

Miguel Guardado

02/19/2021

1. Explain the distance metric you utilized to calculate the similarity/dissimilarity between small molecules
2. Use a dimensionality reduction algorithm (PCA, t-SNE, UMAP, etc) to generate a

2D visualization of the small molecule dataset. Each point should represent a

single molecule. (Note: you may have to plot a subset of your data, depending

on which dimensionality reduction algorithm you choose.)

1. Cluster the small molecules using your implementation of a partitioning clustering

algorithm. Visualize this clustering by coloring clusters on the 2D visualization

generated in question 2.

1. Explain your choice of partitioning clustering algorithm. Visualize this clustering in the same way as question 3
2. Cluster the small molecules using your implementation of a hierarchical clustering algorithm. Visualize this clustering in the same way a question 3
3. Explain your choice of hierarchical clustering algorithm. Is it sensitive to initialization conditions? How do you select the number of clusters?
4. Evaluate the quality of both clustering using your implementation of a clustering quality metric. Explain your choice of quality metric. Which clustering preformed ‘best’ according to your metric?
5. Compare the two clustering using your implementation of a clustering quality metric. Explain your choice of quality metric. Which clustering preformed ‘best’ according to your metric
6. For the “best” clustering, as determined by your quality metric, visualize the distribution of Auto dock Vina scores in each cluster. Do members of the same cluster have similar docking scores? Why or why not
7. Select the top scoring molecule from each cluster. This Is your list of cluster heads. Visualize the top 5 by score in PyMOL and pick your favorite. Are they structurally diverse?