Quantification of High-Dimensional Configuration of Ligand Biding Interactions using A Non-Linear Dimensionality Reduction Method of Manifold Learning

Hakim Mohd Azhan and Alex Dickson

Background

Molecular recognition, which is the process of biological macromolecules interacting with each other or various small molecules with a high specificity and affinity to form a specific complex, constitutes the basis of all processes in living organisms. Proteins, an important class of biological macromolecules, realize their functions through binding to themselves or other molecules. A detailed understanding of the protein-ligand interactions is therefore central to understanding biology at the molecular level. Moreover, knowledge of the mechanisms responsible for the protein-ligand recognition and binding will also facilitate the discovery, design, and development of drugs. There has been a broad study in the molecular recognition to help in developing drugs for diseases cure.

Bromodomain and PHD finger containing protein transcription factor (BPTF) is an epigenetic protein involved in chromatin remodeling and is a potential anticancer target. However, there is one reported small molecule inhibitor for BPTF bromodomain (AU1, ***rac-1***) (Kirberger et al.). The BPTF bromodomain ligand, AU1 is a specific ligand that was developed by the Pomerantz lab at the University of Minnesota. AU1 was developed to be a potential drug molecule. Their study has been on analyzing the selectivity, ligand deconstruction, and cellular activity of a BPTF bromodomain inhibitor (Kirberger et al.). There are a few advances made on the structure-activity relationship of a BPTF bromodomain ligand using a combination of experimental and molecular dynamics simulations leading to the active enantiomer (S)-1. Molecular dynamics (MD) has become a routine computational tool for elucidating the detailed process of ligand binding to a receptor from a high-dimensional configuration space of the ligand-protein dissociation. Due to the rapid development of this study, numerous amounts of conformations have been captured which accounts to about 1,195 conformations. Each conformation contain a high quantity of atomic information, e.g. about 30,000 atoms in the dataset, hence it could not guarantee the observation's quality. An efficient method is to reduce the dimensions of configurations and visualize them in two or three dimensional spaces, then some patterns may emerge, e.g. similar data would flock together to become clusters, and they could be easily observed in the graph.

For this project, I am proposing an application of a non-linear dimensionality reduction method known as the manifold learning to process high-dimensional conformations of ligand-protein interactions in a manner facilitating interpretation. Manifold learning can be thought of as an

attempt to generalize linear frameworks like Principal Component Analysis (PCA) to be sensitive to non-linear structure in data (*2.2. Manifold Learning — Scikit-Learn 0.20.2 Documentation*). This method enable to handle data's nonlinear property. In particular, I will cover manifold learning-based methods for looking at low dimension representations that show which atomic degrees of freedom contribute most to the ligand-binding interactions. The outcomes will be relevant for rationalizing the molecular mechanisms that are happening within the dataset.

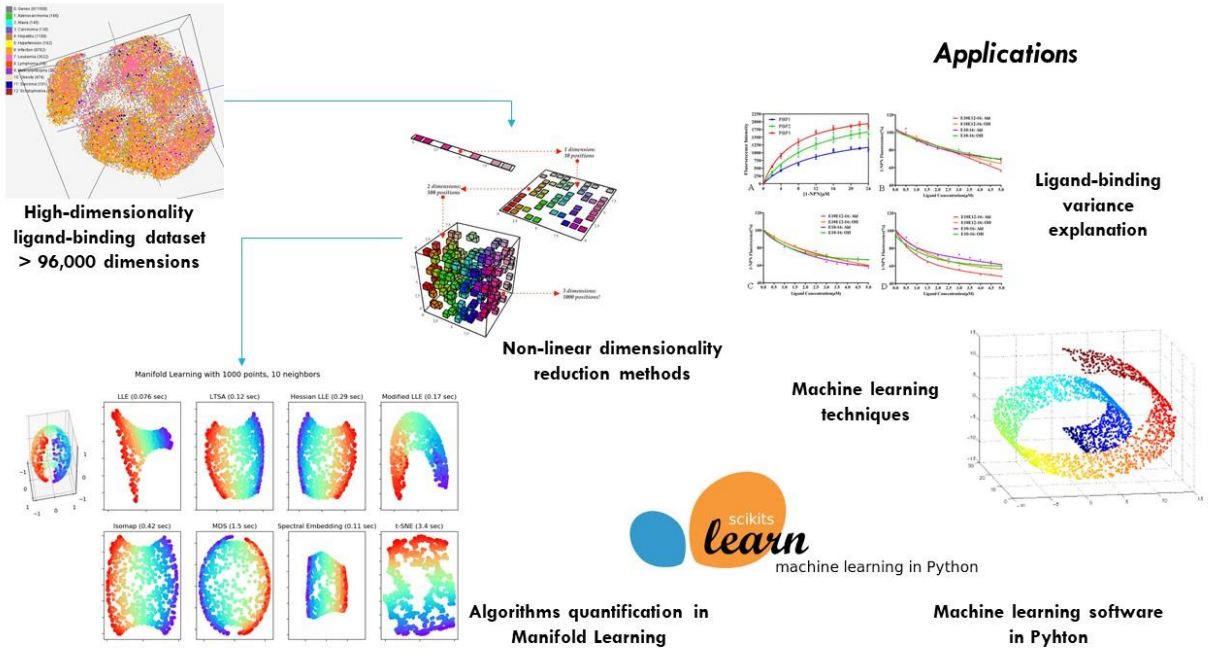
To validate these predictions, clustering algorithm K-means is applied on the dimensionality reduction results to quantify the dimensionality performance. The average clustering accuracy for all the algorithms will indicate that the proposed dimensionality reduction methods could preserved the underlying structure of molecular and the visualization results could reflect the relationships among molecular. I will compare the manifold learning results with the results of molecular simulations (specifically, the "committor probability" of each structure to either the bound or unbound state). Hence, I wish to demonstrate that manifold learning techniques can be helpful in the analysis of ligand diffusion landscapes and provide useful tools to examine structural changes accompanying rare events.

Methods

This problem can be addressed by using the non-linear dimensionality reduction method. Unlike the linear dimensionality reduction, this method is used to minimize information loss for feature extraction, compact coding and computational efficiency. The data can be transformed into "good" representations for further processing, constraints among feature variables may be identified, and redundancy eliminated. A novel strategy of dataset reconstruction using manifold learning has been proposed for dealing with noise in the interaction interface data, which in this case it relates to the information from the protein complexes databases, specifically in the interaction between the BPTF bromodomain protein and ligand.

This work is based on an assumption that there is a low-dimensional manifold, where interface and non-interface can be differentiated from each other. In this study, a longitudinal analysis of different manifold learning algorithms is planned to look at the degree of freedom contributes to most of the ligand-binding interactions representation. Isomap, one representative of manifold

learning method. Isomap is an extension of MDS that tries to maintain the intrinsic geometry of by adopting an approximation of the geodesic distance on the manifold, where the geodesic distance is calculated by summing the Euclidean distances along the shortest path between two nodes. Another algorithm that will be used to capture the geometry structure and apply eigendecomposition to maintain the structure in a lower dimensional embedding of the data are locally linear embedding (LLE). LLE assumes each sample could be represented as the linear combination of its local neighbor samples and tries to find an embedding that could preserve the local geometry in the neighborhood of each data point. Some other methods are proposed to improve LLE's quality, such as modified locally linear embedding (MLLE), Hessian eigenmapping (HLLE), spectral embedding, local tangent space alignment (LTSA) and t-distributed stochastic neighbor embedding (t-SNE) algorithms. To facilitate the chemical interpretation, the multi-dimensional dataset is represented by the most occurring substructure for each feature. For visualizing the feature space, the principle components on the most important features are extracted using the scikit-learn library in Python. This step had been proven beneficial for active learning in preliminary investigations.



Results and discussion

Datasets and preliminary analysis

This dataset is associated with a paper that analyzes the This dataset contains big high-dimensional data as it contains 1195 molecular conformations, each with 32213 atoms with positions in the three spatial dimensions (x, y, z). Each conformation contains the protein, the ligand and water molecules resolved at atomic resolution. The dataset is formatted in pickle file. Pickle is commonly used for serializing and de-serializing a Python object structure. Pickling is a way to convert a python object like list and dictionary into a character stream which enables all the necessary information to reconstruct the object in another python script (*11.1. Pickle — Python Object Serialization — Python 2.7.16rc1 Documentation*).

A preliminary data analysis has been done on this dataset. The pickle file contains the data of capturing 1195 molecular conformations. Hence, using the MD Traj module, this file has been superposed with the PDB file, which contains all the atomic information of the molecular conformations. This enables for me to observe the molecular dynamics (MD) simulation of the BPTF and its inhibitor as shown in the figure 1. The ligand is moving around the protein showing how the ligand interacting with the protein. In Figure 2, the graph shows that there are a lot of hydrogen atoms that exist in this ligand-protein binding are surrounded by water. Figure 3 shows that the dataset does not contain any data about the B-factor that relate to the temperature information.

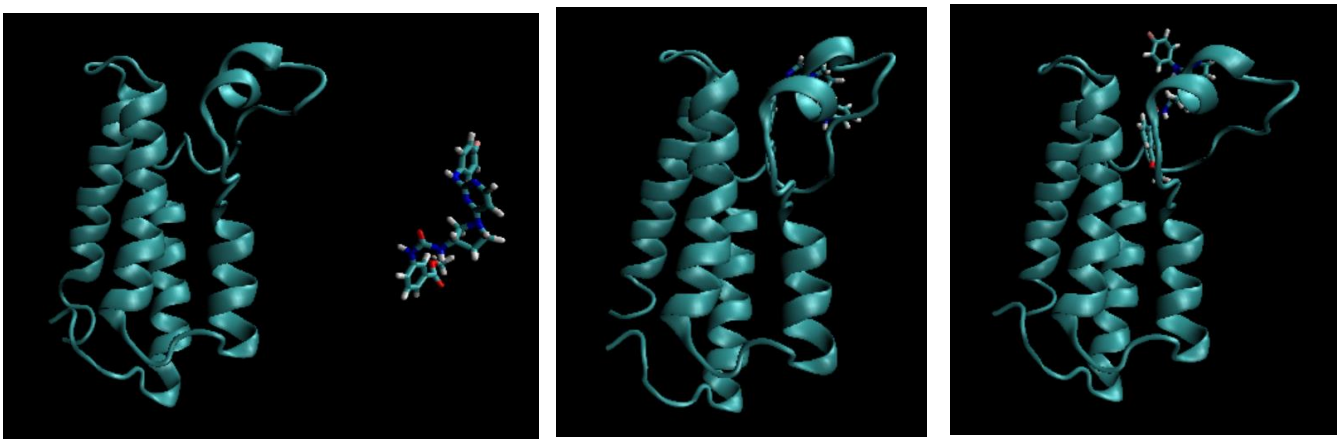


Figure 1: These are three consecutive conformations of the ligand-protein binding of BPTF bromodomain ligand and its inhibitor, AU1.

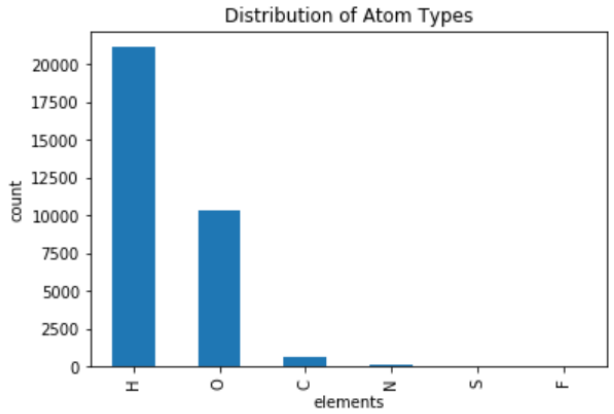


Figure 2: Graph representation of all types of atoms that available in the dataset. Hydrogen (H), oxygen (O), carbon (O), nitrogen (N), Sulphur (S), and fluorine (F).

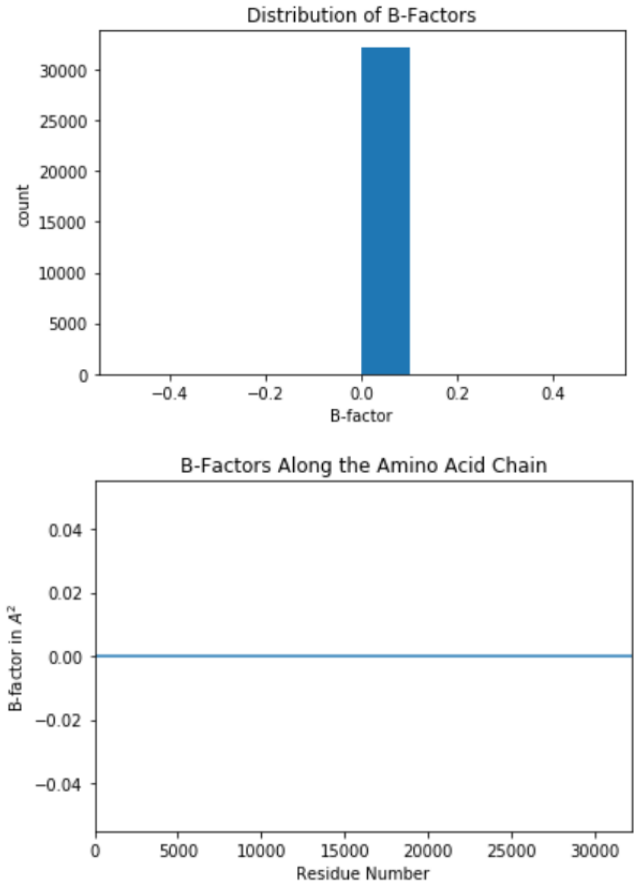


Figure 3: Both graph representation shows the preliminary analysis of the datasets on B-Factors. Constants zero on both representations shows that the dataset does not provide any data on the B-Factors.

Dimensionality reduction results

I conduct dimensionality reduction on the type of proteins individually for the ligand and the protein binding-site and then combine both proteins with the Manifold Learning algorithms. These combinations with different sizes and dimensions are reduced to two dimension vectors. Then, the reduction results are visualized in graphs. Figure 4 and Figure 5.a illustrate the 2-D visualization results for ligand and protein binding site respectively. Each subfigure shows the visualization of one different non-linear dimensionality reduction algorithm used.

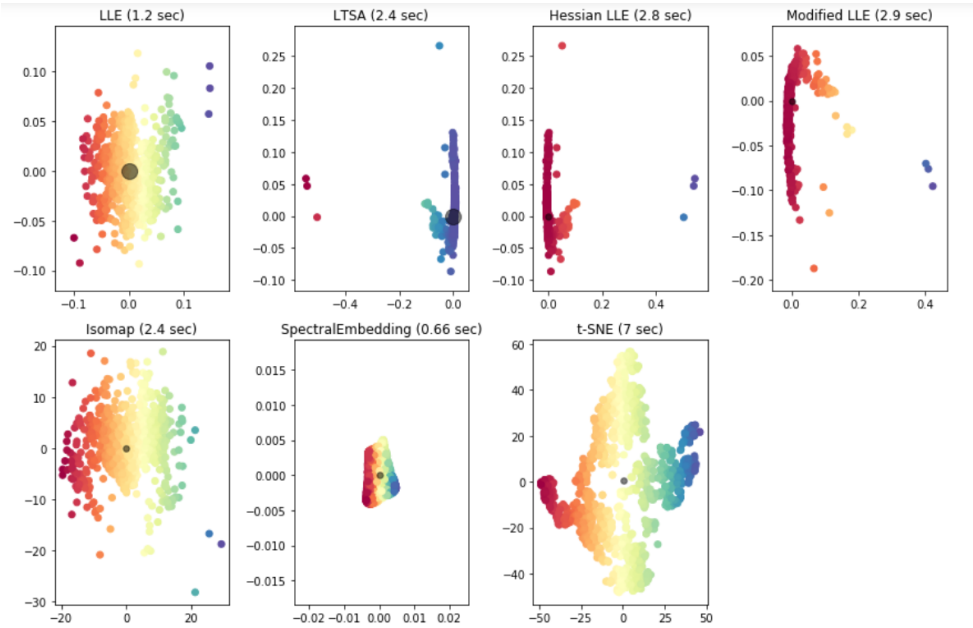


Figure 4: 2-D visualization for the dimensionality reduction by applying the Manifold Learning algorithms. This visualization composed of combination of atoms that make up the ligand. The small black dot shows the center of the cluster calculated by the K-Mean algorithm, K=1.

The modification of the LLE which are Hessian LLE and modified LLE give a different visualization as it become more condense than LLE. The shape of Isomap is similar like LLE but the measures are different for both visualizations. It can be observed that all the graph visualizations have a center point that has been calculated by the K-mean algorithm.

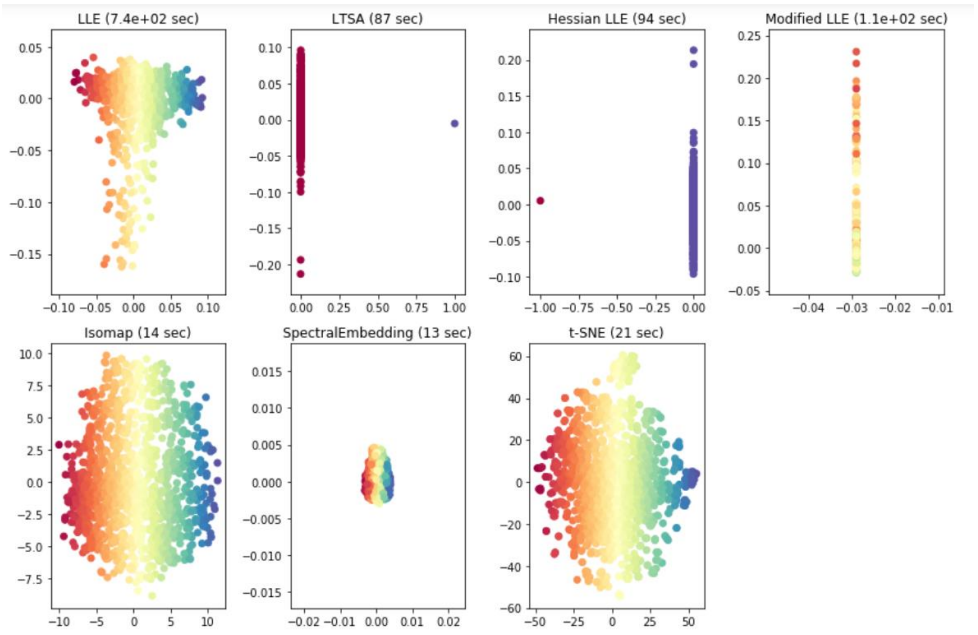


Figure 5.a : 2-D visualization for the dimensionality reduction by applying the Manifold Learning algorithms. This visualization composed of combination of atoms that makes up the protein-binding site.

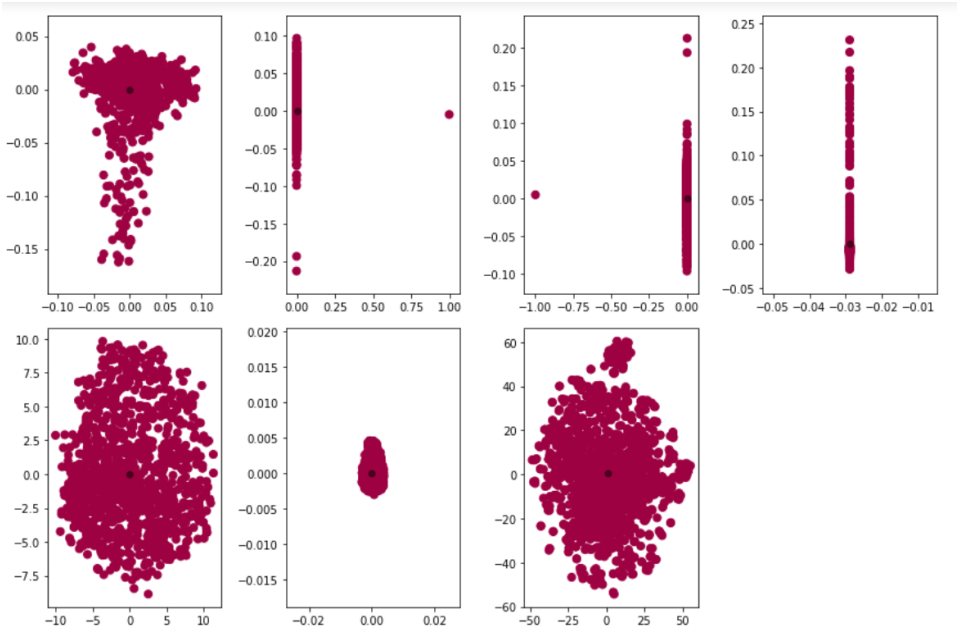


Figure 5.b : This 2-D visualization shows the clusters classification implied by the K-mean clustering algorithm, K=1. The small black dot shows the center of the cluster.

Among all the algorithms, the results generated by the Spectral Embedding are more concentrated than any other algorithms. This may because Spectral Embedding applies Gaussian kernel on the similarity matrix, and some similarities may become zero after this process, which makes the points in the graph more concentrated. Since the K-mean algorithm is calculated for protein, there visualization is colored in red to represent a cluster.

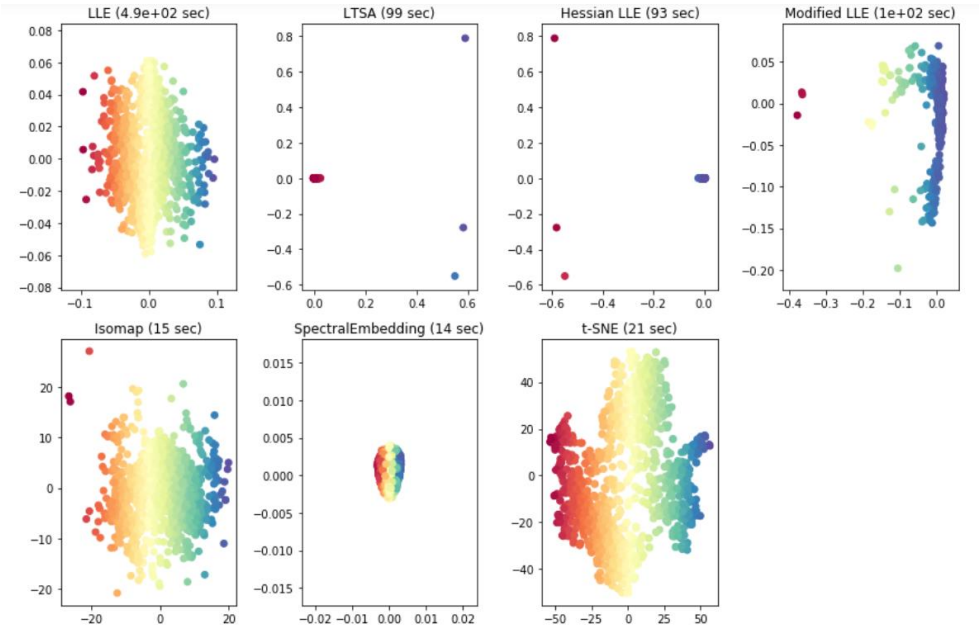


Figure 6.a : 2-D visualization for the dimensionality reduction by applying the Manifold Learning algorithms. This visualization composed of combination of atoms that makes up the protein-binding site and the ligand.

The dimensionality reduction on the combination of ligand and protein binding site results in Figure 6.a and 6.b. Just like before, Spectral Embedding has the most concentrated data points than any other algorithms. However, LTSA and Hessian LLE show a bit weird reduction when both have lesser data points in the graph visualization. In Figure 6.b, I can observe that both ligand and binding protein site have been clustered into two clusters. Plus, there are small black points that shows the centers of the clusters.

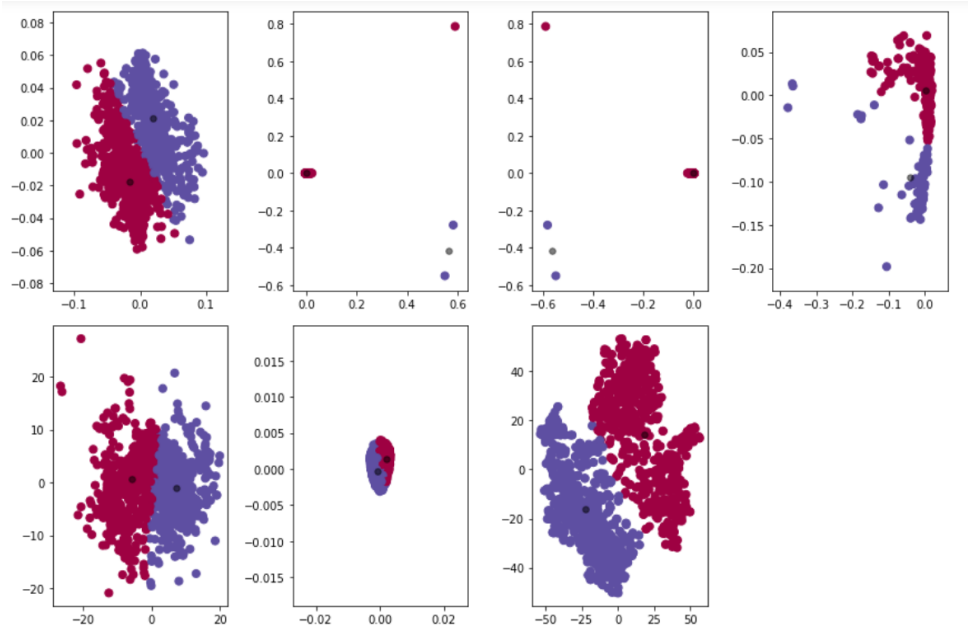


Figure 6.b : This 2-D visualization shows the cluster classification implied by the K-mean clustering algorithm, K=2. There are two small black dots shows the centers of the clusters separately.

Time performance

I also test the time performance of all the algorithms. The time required to compute the algorithms performance has been included in the manifold learning. The time performance will tell about the sensitivity of the algorithm towards the data size. Table 1 shows the results.

Table 1 *Comparison of Manifold Learning algorithms in terms of time consumption*

	Ligand	Protein-binding site	Ligand and Protein binding site
LLE	1.2s	7.4e+02s	4.9e+02s
LTSA	2.4s	87.0s	99.0s
Hessian LLE	2.8s	94.0s	93.0s
Modified LLE	2.9s	1.1e+02s	1.0e+02s
Isomap	2.4s	14.0s	15.0s
Spectral Embedding	0.66s	13.0s	14.0s
t-SNE	7.0s	21.0s	21.0s

LLE denotes Locally Linear Embedding, LTSA denotes Local Tangent Space Alignment, t-SNE denotes t-distributed Stochastic Neighbor Embedding.

Spectral Embedding consumes much less time than other algorithms. Most different step for the other algorithms is to calculate the similarity matrix. Isomap and Locally Linear Embedding needs to apply knn algorithm and shortest path algorithm to achieve the similarity matrix, while Spectral Embedding also known as the Laplacian Eigenmaps only apples Gaussian kernel on the edit distance matrix. Higher complexity calculation for the Isomap and Locally Linear Embedding than Spectral Embedding results in larger time consumption. For the combinations of ligand and protein binding site atoms, it takes longer time consumption due to larger amounts of atoms molecules exist with higher dimensions. The time of Isomap and Locally Linear Embedding increases which means the time performance of these algorithms is sensitive about the data size (Yang et al.).

Take-out Lesson

This individual computational project has thought me a lot on the thought process for conducting a computational biology project. These are the main lessons that I can apply for future computational biology research if I were to conduct one:

1. Require basic computational technique like the usage of computational tools and concept like the broad computational concepts and algorithms. For this limitation, a lot of research and reading have to be done in order to have a better understanding of the computational biology.

2. There can be more than one tools that can be used to solve this problem. Based on another article that I read, instead of using Python, Matlab also can be used to conduct the dimensional reduction of large dataset.

3. Large dataset should be run on a better and compatible machine like HPCC to avoid any crashing of the computer. This will also consume more time to carry an analyzation of the data.

If I were to start over this project, I would spend more time to understand the algorithms of the Manifold Learning. Not all the algorithms that exist in the Manifold Learning should be used to

get the results that are wanted. I would still to obtain the clustering validation using the accuracy score and compare the dimensionality result with the committor probability. Among all the algorithms, multi-dimensional scaling (MDS) does not working that making me to omit that. I wish that I can look into this because as far as I concern it has to do with the matrices and that relates to statistic.

Conclusion

Manifold Learning is a useful tool to reduce higher dimensional to a lower data representation. The best algorithm of Manifold Learning is the Spectral Embedding because it applies Gaussian kernel on the similarity matrix. Plus, the time performance of the Spectral Embedding is faster than other algorithm like LLE, hessian LLE, modified LLE, Isomap and t-SNE. These other algorithms have lower time performance due to the shortest path algorithm that has to be calculated to find the similarity.

References

2.2. Manifold Learning — Scikit-Learn 0.20.2 Documentation. https://scikit-learn.org/stable/modules/manifold.html. Accessed 18 Feb. 2019.

11.1. Pickle — Python Object Serialization — Python 2.7.16rc1 Documentation. https://docs.python.org/2/library/pickle.html. Accessed 24 Feb. 2019.

Kirberger, Steven E., et al. "Selectivity, Ligand Deconstruction, and Cellular Activity Analysis of a BPTF Bromodomain Inhibitor." *Organic & Biomolecular Chemistry*, vol. 17, no. 7, 2019, pp. 2020–27. *Crossref*, doi:10.1039/C8OB02599A.

Yang, Jiaoyun, et al. "Nonlinear Dimensionality Reduction Methods for Synthetic Biology Biobricks' Visualization." *BMC Bioinformatics*, vol. 18, no. 1, Jan. 2017, p. 47. *BioMed Central*, doi:10.1186/s12859-017-1484-4.