DSC 190 Final Project

Alexander Friend
apfriend@ucsd.edu

1 Backgound

Topology is the study of how geometric properties and spatial relations change or do not change, often focusing on features of different dimensions, from 0 dimensional points to higher dimensional "hyperobjects" that are impossible to truely imagine.

Persistent homology is the extension of the study of homology focusing on the way in which topological features change in sequences of spaces as different parameters are changed. Common tools used are persistence diagrams and barcodes, which show the "birth" and "death" of different topological features in different dimensions.

These studies are used in a great variety of ways, in topics such as computer graphics or GPS systems, all of which take advantage of various topological principles. One highly notable application is that of molecular biology. Since atoms can be abstracted as zero dimensional points in a multidimensional space, many of the techniques used in topological data analysis can be leveraged when studying some of the smallest building blocks of life.

Topological techniques can be leveraged in order to study the geometric properties of molecules. Topology classifies so called topological spaces as structures which have both an empty set, or the lack of any elements as weell having elements, which may form unions or intersect other elements within said set. They way in which these features interact with each other can have meaningful geometric properties that can be used when studying biological macromolecules.

The function of proteins, the fundamental building block of known life, is fundamentally determined by the physical form that these macromolecules exist in. There are four key structures that proteins are classified into, though sometimes they are only classified within the first three such structures. The first *Primary* structure is simply the linear sequence of amino acids, of which all proteins are formed of. Amino acids themselves are molecules in their own right, formed primarily out of Hydrogen, Nitrogen, Oxygen, and Carbon, though other elements are often present. The second *Secondary* Structure is determined by the three dimensional shape that these chains of amino acids form, primarily forming either, or more often a combination of both, α-helixes and β-sheets. The third *tertiary* structure of proteins is dictated by way in which this chain of amino acids with α-helixes and β-sheets bends in three dimensional space. If more than a single chain is present in a protein, then the way in which the multiple chains interact with each other

and bend due to intermolecular and intramolecular forces is defined to be a proteins *quaternary* structure [1].

In order to better study these proteins, their fundamental structure must be analyzed. It is in this task where topological data analysis can become an invaluable tool.

2 Task Description

For my task, I downloaded and extracted files on protein molecular structure from the Protein Data Bank. From this I compared the structure using topological methods, primarily using topological concepts such as simplicial complexes and simplex trees. A simplicial complex K is a collection of geometric simplexes such that if any geometric simplex $\sigma \in K$ then any face $\tau \subseteq \sigma$ is also in K. This is a useful task as topological simplexes are quite useful at describing important features of proteins that affect their function.

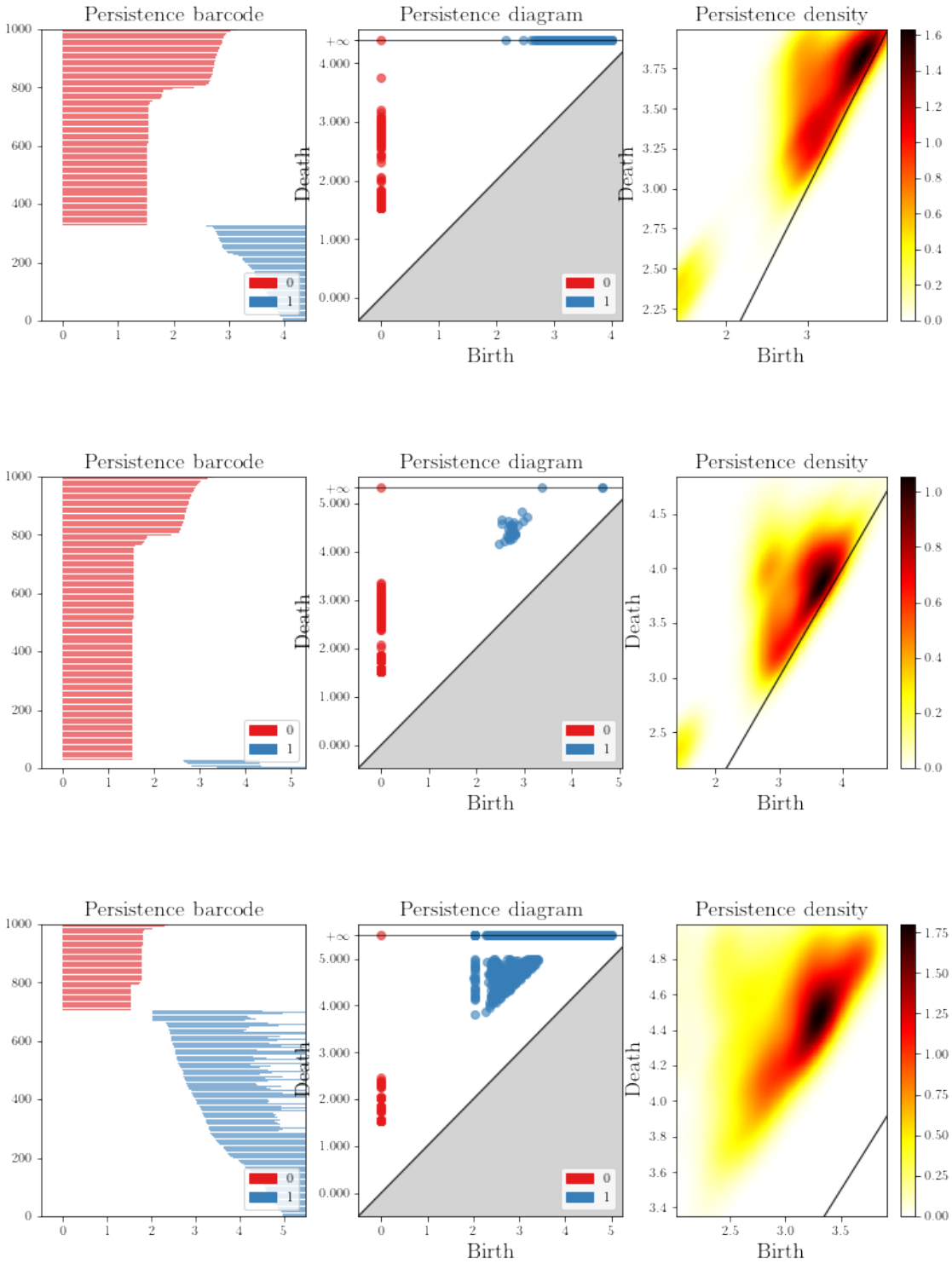
3 Methodology

In order to obtain data, I wrote a script to parse pdb files found in educational Protein Data Base's web portal, category selection. This was done in so that I would have a way to quickly categorize different proteins in bins in such a way to increase the likelihood of being able to successfully categorize and cluster proteins by their topological profiles.

The two groups of proteins that I ended up most closely studying were proteins that were hormones and viral. The hormone protein primarily was made up of various recombinant human insulin proteins, bovine insulin, and glucagon. Comparatively viral proteins more closeley studied were proteins from the SARS-COV virus, the SARS-COV-2 virus, the Ebola virus, and the human rhinovirus, more colloquially known as the common cold. Other proteins more closely studied were receptor proteins for insulin as well. These groups were chosen under the assumption that viral proteins would differ greatly from proteins associated with the endocrine system. The two groups were closely related to each other in function however, so the hypothesis was that these proteins would cluster more easily.

3.1 Exploratory Data Analysis

In order to better explore the data I plotted persistence barcodes as well as persitence diagrams, such as Figure 1,. Figure 2, and Figure 3 below, as can be seen the Sars-Cov-2 protein has many more 1-dmensional simpexes that persist than compared to the recombinant insulin or insulin receptor



4 Results

Much time was spent simply trying to create simplex tree objects for some of the protein structures. This is a result of the very high amount of atoms in some proteins, meaning

that there were numbers of vertices in the magnitude of the hundreds of thousands, which was very computationally and memory intensive. This also limited me to only being able to analyze a handful of the .pdb files I had parsed and extracted to point cloud data.

This also affected my ability to create a classification model, as I simply did not have enough data to properly train any kind of model.

5 Future Work

Givven more time I would like to explore other methods of producing persitence profiles. One method I found in late in my research is that of calculating a so-called *dynamic distance*, described by the equation:

$$D_{ij} = 1 - \|C_{ij}\|$$

where C_{ij} is the cross-correlation matrix of molecular oscillation[2]. This method was developed in order to account for the variance that is inherent in measureing molecules, as they vibrate due to changes in their environment such as heat.

Bibliography

[1] Alberts B, Johnson A, Lewis J, et al. Molecular Biology of the Cell. 4th edition. New York: Garland Science; 2002. The Shape and Structure of Proteins. Available from: https://www.ncbi.nlm.nih.gov/books/NBK26830/]

[2] Kovacev-Nikolic V, Bubenik P, Nikolić D, Heo G. Using persistent homology and dynamical distances to analyze protein binding. Stat Appl Genet Mol Biol. 2016 Mar;15(1):19-38. doi: 10.1515/sagmb-2015-0057. PMID: 26812805.

All source code and notebooks have been uploaded to the GitHub Repo: