Protein Symmetries

Nick Backes, Alex Dowling, Nilava Metya
Erdos Institute DS Bootcamp Spring 2024

Overview

Proteins consist of a sequence of amino acids assembled into a linear chain. To function, it must be realized into a three dimensional structure, which typically exhibits some form of symmetry between identical or similar subunits. Identifying the symmetries of a protein structure is useful in understanding the function of the protein. We aim to address the following:

- Which features of proteins predict their symmetry?
- Given experimental data on a protein, how accurately can we predict if the symmetry is group is cyclic C2?

This data is useful to biochemists, pharmaceutical scientists, and applied mathematicians.

Data Collection

The source of data for this project is the Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank (PDB), which provides thousands of lines of annotated data on the structure of over 90,000 proteins and other macromolecules. We extracted two labeled data sets from these files:

- 250 point clouds each storing the positions of all atoms in space for a protein (centered at the origin with farthest point scaled and rotated to (1, 0, 0)),
- Tabular data containing the number of atoms of each type, number of bonds of each type, number of amino acids of each type, and oligomeric count for ~5000 proteins.

Modeling & Results

Our approach to handle the point cloud data was a k-nearest neighbors classifier equipped with the Hausdorff metric, which measures the greatest distance from a point in one cloud to the closest point in the other cloud. For a training set of 200 proteins, we set k=14 to achieve:

Model	Train Accuracy	Test Accuracy
Baseline (Mode)	66%	66%
Point Cloud KNN	69%	58%

We found that the computational complexity of the metric combined with the high dimensionality of the space of protein point clouds rendered this classifier ineffective.

For the tabular data, our final model used the software XGBoost to generate a gradient boosted decision tree model. This contained 229 features (described in the data collection section) on roughly 5,000 proteins. The performance of the model is shown below.

Model	Train Accuracy	Test Accuracy
Baseline (Mode)	75.92%	76.15%
Simple Regression Tree	87.97%	87.91%
XGBoost	99.97%	90.63%
XGBoost without oligomeric count	99.82%	77.73%

Performing a feature importance analysis on the XGBoost model revealed that oligomeric count (number of repeated subunits in the protein) was the dominant feature in predictions.

Conclusions & Future Directions

The molecular structure of proteins is extremely complicated, with intricate chemical information occurring over thousands of atoms for each protein. Our methods did not effectively incorporate the atom point clouds into a compelling model. However, coarser features like oligomeric count proved effective in identifying C2 symmetry via the XGBoost model.

Methods from point cloud registration and topological data analysis could improve the computational efficiency of the KNN algorithm, while the XGBoost model can be improved by considering more features from biology. Restricting to features which are easier to determine experimentally would make the classifier more useful in practice.