Protein Symmetry Detection

Nick Backes, Alex Dowling, Nilava Metya

Erdos Institute
Data Science Bootcamp

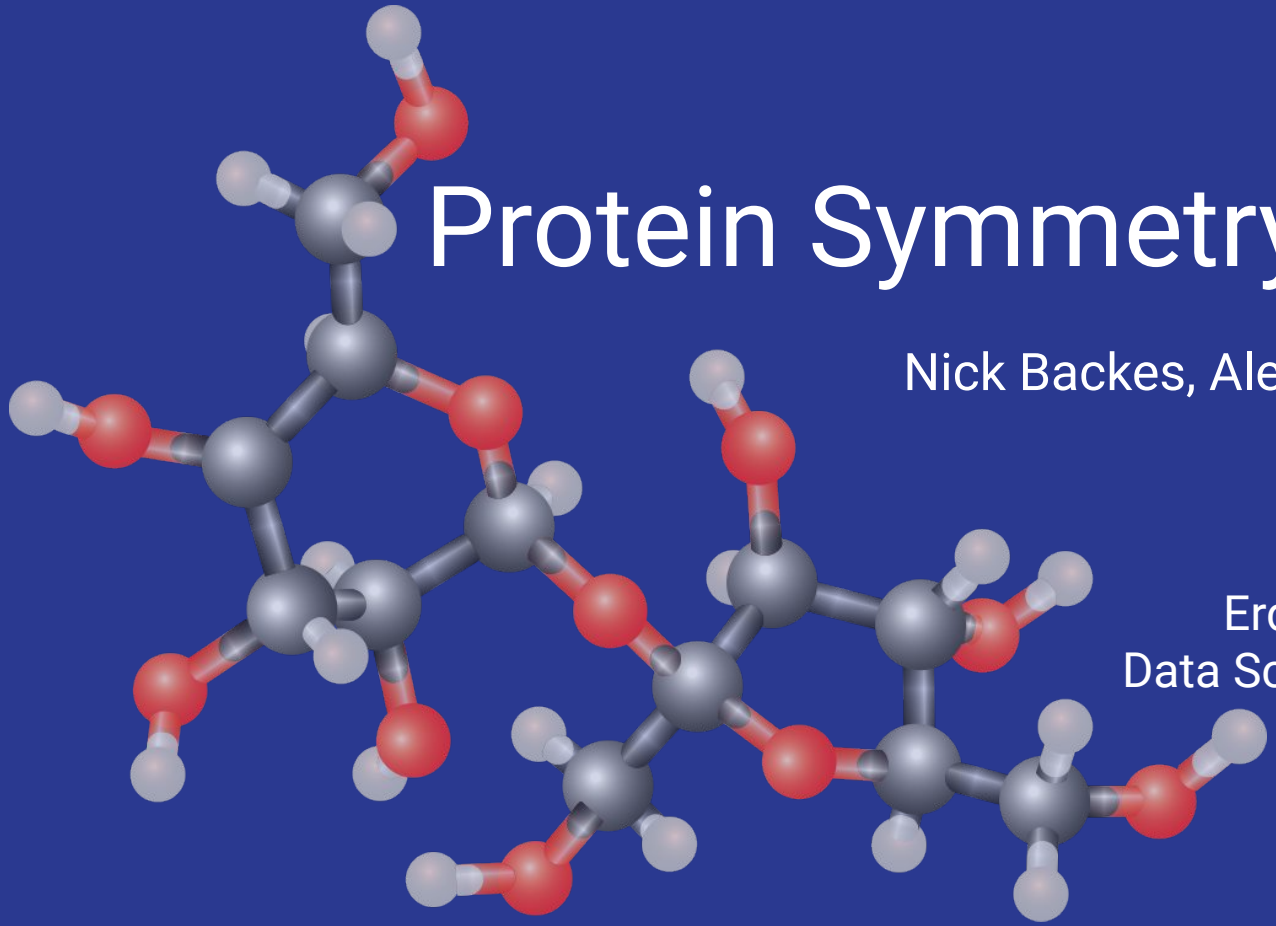


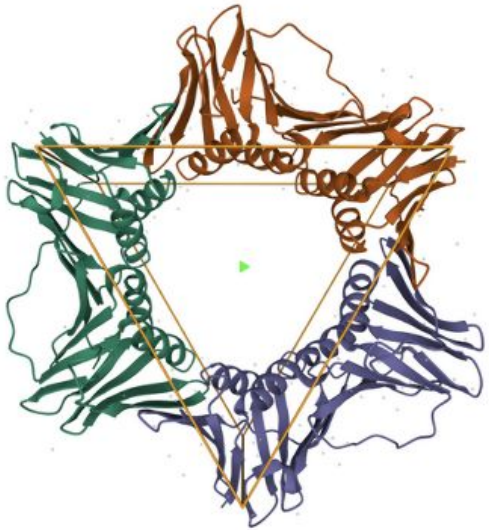
Image Credit: William Crochot
https://commons.wikimedia.org/w/index.php?curid=36728538

The Problem

Proteins consist of a sequence of amino acids assembled into a linear chain in 3d-space. It is easier for biological processes to synthesize proteins with some repetition; this leads to symmetries.

Questions:

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



Key Performance Indicators: Accuracies of the classifiers on a randomly selected test set, compared to a baseline of mode

Stakeholders: Scientists working in biology, biochemistry, and pharmaceuticals

The Data

Our source of data is the Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank (PDB). The PDB provides annotated data on the structure of over 90,000 proteins.

Each protein has an associated PDB file with thousands of lines of data, including:

- Names of scientists who discovered the protein
- Equipment used to perform the crystallography
- Number of repeated subunits in the protein
- Precise positions of atoms in the protein and their types
- Bonds between atoms in the protein and their types
- Amino acids and their sequencing

Data Formatting

We extracted two data tables:

The first consisted of 250 proteins, and only contained the positions of each atom in molecule.

The second consisted of ~5000 proteins, and the following features:

1. Total number of atoms of each type (hydrogen, carbon, oxygen, etc.)
2. Total number of amino acids of each type
3. Total number of bonds of each type (single, double, triple)
4. Oligomeric count (number of repeated subunits)

The last four form a table with 229 columns including symmetry.

KNN Model

Procedure (for each protein):

1. Extract the point cloud.
2. Shift so average lies at the origin.
3. Scale into a ball of radius 1.
4. Rotate so that furthest point is at (1,0,0).

The *distance* between clouds is the greatest distance from a point in one cloud to the closest point in the other cloud (Hausdorff metric).

Use a K-nearest-neighbors classifier to identify the symmetry.

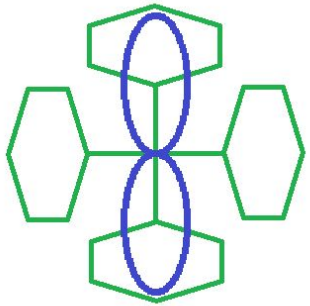


Fig 1: two molecules with different symmetry groups and high distance

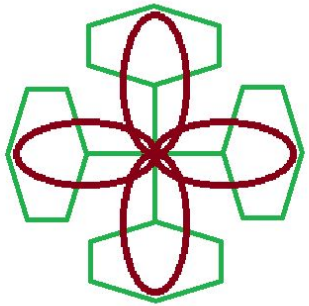


Fig 2: two molecules with the same symmetry group and low distance

KNN Model

We trained a k=14 nearest neighbors model on a training set of 200 proteins, and tested on 50 proteins.

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

Challenges:

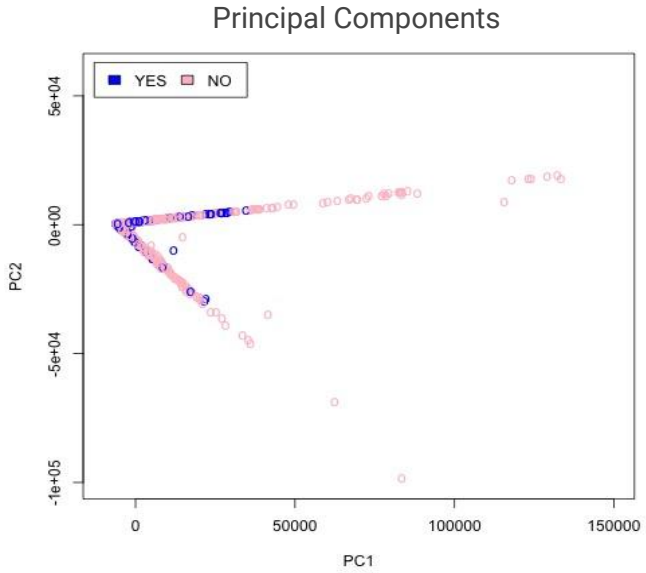
- Good alignment proved challenging
- Distance function selection
- Large computation time
- Lots of data needed for the proper averaging

Preliminary: PCA

Started with a Principal Component Analysis

- 99.95% of variance in first 3 components
- Features contributing included:
  - Atoms counts: N, C, O, H, P
  - Amino acid count: LEU

After identifying the principal components, did a linear regression



Model	Training Accuracy	Testing Accuracy
Baseline (Mode)	75.92%	76.15%
PCA + regression	75.72%	77.29%

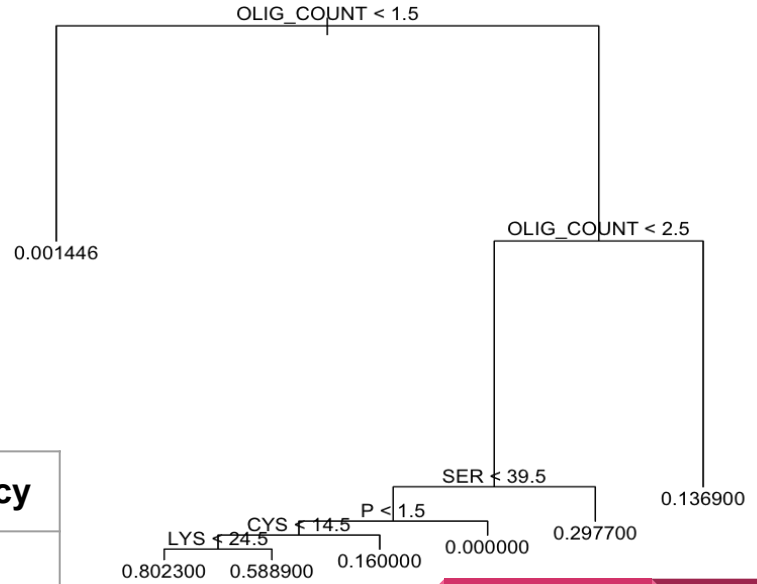
Preliminary: Regression Tree

Features selected by the tree:

- Oligomeric count (OLIG_COUNT)
- Amino acid counts: SER, CYS, LYS
- Phosphorus atom count

Without oligomeric count, the tree performed much worse

Model	Training Accuracy	Testing Accuracy
Baseline (Mode)	75.92%	76.15%
Tree with OC	88.52%	87.84%
Tree without OC	77.29%	75.77%

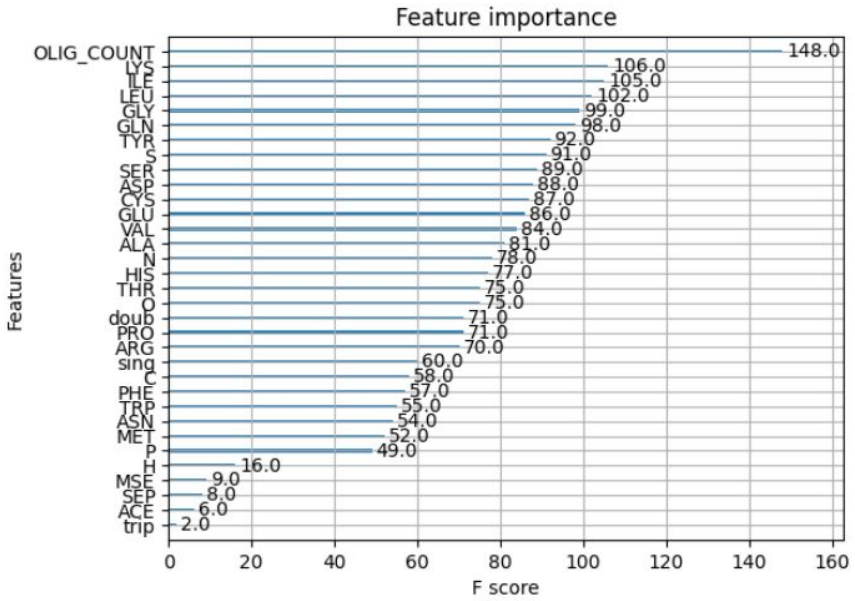


XGBoost Model

We used XGBoost to generate a gradient boosted decision tree model for the tabular data.

Model	Train accuracy	Test accuracy
Baseline (Mode)	75.92%	76.15%
XGBoost	99.97%	90.63%
XGBoost without OC	99.82%	77.73%

Most important feature: oligomeric count (OC)



Results

KNN (K-Nearest-Neighbors) Model:

- 200 training proteins
- Accuracy: 69% train / 58% test
- Features: atom point clouds

Regression Tree Model:

- 4000 training proteins
- Accuracy: 87.9% train / 88.0% test
- Features:
  - Total number of atoms of each type
  - Total number of amino acids of each type
  - Total number of bonds of each type
  - **Oligomeric count**

XGBoost:

- 4000 training proteins
- Improved tree-based model
- Accuracy: 99.97% train / 90.63% test
- Features:
  - Total number of atoms of each type
  - Total number of amino acids of each type
  - Total number of bonds of each type
  - **Oligomeric count**

Future Directions

Explore more symmetries!

KNN model:

- Point cloud registration to better align molecules
- Better distance function that captures the geometry of the molecules
  - Rewarding similarities vs punishing differences
  - Sensitivity to noise
- Downsampling or parallel processing to improve computation time

XGBoost:

- Constraints in feature selection induced by lab-based science

Other model types:

- Neural network or other machine learning style model

Thanks For Watching!

Special thanks to our group mentor, Kash Bari, for excellent words of advice and encouragement.

References:

[1] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The Protein Data Bank. Nucleic Acids Research, 28(1):235–242, 01 2000.