Protein superfamily classification using Machine/Deep Learning Approaches

Proteins are large, complex biomolecules that play an essential role in the body. Proteins are made up of thousands of tiny units connected in a long chain called amino acids. Amino acids are simply represented by alphabets. There are 20 different most common and 5 uncommon types of amino acids that can be combined to form a protein. They are arranged randomly in sequences that build millions of unique proteins. Each protein has a unique 3-dimensional structure and its specific functions, which are governed by a sequence of amino acids. A protein family is a group of proteins that share a common evolutionary origin, reflected by their functions and similarities in sequence or structure. Usually, proteins are arranged in hierarchies' designs where proteins share a common ancestor subclassified into smaller groups. A large group of distantly related proteins is the superfamily, whereas a small group of closely related proteins is called a subfamily. The proteins with similar structural and functional domains are classified as members of a specific family. The traditional method to identify protein families is complicated and time-consuming. So, this project explores many machine learning and deep learning methods that can be used to classify protein families.

The dataset is downloaded from Kaggle, which is initially retrieved from Protein Data Bank (PDB). The data contains details on protein classification, extraction methods with more than 400000 protein structures. With these proteins having different family types, this project aims to determined protein families based on their sequences using machine learning techniques. Moreover, this project compares several methods and come up with the one way that shows most success to classify protein family based on the accuracy and many other skill metrics.