Predicting compound properties using machine learning

Shakiba Fadaei

Master thesis in Bioinformatics

Computational protein structure and function prediction has been a subject of research for many years. Predicting protein functions and structures helps capture the biological roles of proteins, and leads to filling biological gaps. Throughout the years, many tools have been developed for this purpose. With the emergence of high-throughput sequencing technologies, most of these methods use the sequence information in order to extract these properties. Nevertheless, a noticeable proportion of proteins are enzymes. As a result, some studies have been performed on capturing protein properties based on the reaction mechanisms they catalyze. Enzyme Commission (EC) number is a very popular way of identifying enzyme functions. A large number of tools that predict enzyme functions use EC numbers as a descriptor.

In this project, the main purpose is to find enzyme structures using reaction fingerprints. Reaction fingerprints contain the information of substrate reactive sites, atom and bond changes, and surrounding atoms in a reaction. So, it captures the 2D representation of a reaction. Because of this, it can be expected that the reaction fingerprints could also capture the enzyme structures.

The parsing steps and extraction of the feature vectors from the reaction fingerprints, using neural networks, are explained in details. Moreover, the feature vectors extracted are going to be further validated in a model, mapping reaction fingerprints to their corresponding EC numbers. This way we can conclude if the feature vectors contain all the necessary information as the raw fingerprints. Eventually, the feature vectors are used for mapping reaction fingerprints to CATH superfamilies. CATH stands for Class, Architecture, Topology, and Homologous superfamily. Hence, in this project, a model is suggested that can predict the structure of an enzyme that catalyzes a reaction. Deep learning methods have been used throughout this project. The architecture of the networks, and the chosen parameters are explained and justified.

Supervisor: Prof. Daniel Wegmann