DSAA6000M Project Proposal

Ying Wang Tingyu ZHU Jinwang LU

11. October 2024

Any chemical substances, including proteins, follows the principle that structure determines properties. In recent years, protein structure predicting models, like AlphaFold, have made significant progress in predicting protein higher order structures from protein primary sequencing. However, predicting protein biological activity remains a challenge. Fluorescent proteins (FPs) are a great starting point for deep learning models to connect these two kinds of information due to the availability of datasets and various properties related to biological function. In this project, we would like to leverage deep learning network to link FPs sequence information with theoretical physical properties related to biological activity, specifically focusing on optimizing FPs' properties for specific microscopy experiments.

We plan to use amino acid sequences and their associated physical parameters, such as mutations, fluorescence properties and structural data as training dataset for deep learning models, which could be collected from FPbase containing detailed information of 1,073 FPs. Our goal is to predict specific physical parameters of FPs, such as fluorescence colors, intensities and lifetimes. From our model, we would like to screen for brighter or acid-resistant FPs, which have significant potential in single-molecule imaging or correlative light and electron microscopy.

We will normalize the FPbase dataset, drop missing values and encode categorical variables. The model will start with embedding algorithm to convert protein sequences into numerical representations, followed by ESM-2 for extracting global sequence patterns and macro-level features. We propose using convolutional neural networks (CNNs) to process one-dimensional protein sequences, which has shown strong performance with amino acid sequences. CNNs will refine the extracted global and non-local features through local feature extraction, capturing more detailed sequence patterns. An adaptive weighting mechanism is then applied to balance and integrate the features obtained from previous steps, optimizing the model's focus on different characteristics. Lastly, the model concludes with a multi-task prediction Layer, which uses the processed features to predict multiple protein parameters.

We plan to use different loss functions to predict different FPs' physical parameters, such as MSE to fluorescence intensity and KL divergence to fluorescence spectrum. Adam optimizer will be used during the training process. Then model evaluation and fine-tuning hyperparameters will be conducted, in order to improve accuracy. Meanwhile, we will implement feature regression to identify key protein characteristics associated with specific physical properties to improve the model's interpretability. This will not only help us understand the model's predictions but may also suggest new approaches in FPs design.