Protein-Protein Interaction Prediction Using Distributed Representation and Graph Convolutional Networks

Hayden Gemeinhardt

Research Advisors: Adam Bess, Dr. Supratik Mukhopadhyay

Abstract

We introduce a link prediction method for protein-protein interaction. Previous studies have created distributed representations of biological sequences to be applied for bioinformatic investigations. In this study, we focus on the pretrained model ProtVec, a distributed representation of protein FASTAs. We pull data from three databases, namely hu.MAP, IntAct, and STRING, for a total of 94,650 proteins and 6,324,918 distinct interactions between them with weights above a negligible threshold of 0.01. Then, we use the proteins' amino acid sequences in FASTA format to feed them into ProtVec for vectorized representations. Using the protein vectors and weighted interactions, we utilize GraphSAGE, a GCN framework designed for large graphs, to train a link prediction model. To evaluate the model, we use test sets of both randomly selected proteins and specific notable proteins in COVID and cancer research where an accuracy upwards of 90% was achieved. These results indicate that accurate predictions regarding a protein's interaction with other proteins can be made. This model needs to only be trained once and can then be implemented into software such as LSU's DeepDrug, where an AI algorithm can utilize the model to accurately predict the importance of a protein in drug discovery use cases. Furthermore, the protein embeddings derived from the model can be used to run classification tasks, determining protein families or protein pathways a protein interacts with and the protein's involvement on a biological-systems level.