Dense vector representations of proteins based on Gene Ontology annotations using word and sentence embedding tools

JAMES ZHANG[1], Chelsea Ju[2], Muhao Chen[2], Dat Duong[2], Yunsheng Bai[2], Wei Wang[2]

1 BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
2 Department of Computer Science, UCLA

Recent advances of machine learning algorithms make possible more accurate predictions of protein-protein interactions; however, proteins must first be embedded into numerical vectors. Recently, the Gene Ontology (GO), a database of biological terms arranged in a hierarchical graph structure, has been considered a potential source of extracting vector representations of proteins. Here, we utilize two natural language processing methods to embed proteins into dense vector representations. In the first method, the structure of the GO graph, with accompanying protein annotations, is described in a series of sentences. Word embeddings of each protein are generated using Word2Vec. In the second method, sentence embeddings of GO term definitions are inputted into a graph convolutional network trained on entailment relationships of GO terms. Protein embeddings are calculated from GO term embeddings taken from the embedding layer. We find the dense vectors perform well in binary protein-protein interaction and multiclass enzymatic function classification tasks.