Machine learning for predicting chemical-protein relations using graph embeddings

Daniel Viana, Raquel Melo Minardi, Adriano Alonso Veloso

UFMG

Abstract

The volume of biological data available in many repositories is vast and increases almost exponentially. Among this universe of data, we can highlight the relation between proteins and ligands. These relations may be the key to understanding biological processes, drug metabolism, drug design and repositioning, and industrial protein optimization. However, the large amount of data makes the process of analyzing and obtaining them manually a hard and toilsome process, hence making the use of in silico methods indispensable. One of the computational approaches that can be used in this case is graph modeling, where it is possible to represent proteins and ligands as nodes and the relationship between them as edges. So, this work aims to propose a new database of chemical compounds and genomic products graph-based from an unstructured corpus manually curated and to suggest a method capable of predicting a relationship between them through machine learning techniques. For this, we use the corpus chemical-protein interactions available on BioCreative VI Challenge. In order to predict relationships, we first use the Neighborhood Based Node Embeddings (NBNE) algorithm, an unsupervised method capable of generating node embeddings for graphs. Thus we produce a dataset containing the embeddings that represent nodes of the graph that contain a real relation, class one. For class zero, we generate false relationships randomly. To create prediction models, we use the machine learning algorithms: Decision Tree, Random Forest, and SVM as classifiers. Among the experiments performed, the method obtained the best result was SVM. Given the above, the base created is a way to provide and organize unstructured data of relationships between genomic products and chemical compounds and can be used as a query for possible relationships between them. Finally, the proposed method demonstrated to be efficient to suggest new relationships amid graph components through machine learning.

Funding: CAPES