Network-based Protein Function Prediction via Deep Graph Convolutional Neural Networks

MSc. Vicente Enrique Machaca Arceda

August 13, 2020

1 Introduction

Protein assemblies are modules of Protein-Protein Interactions (PPI), they coordinate the execution of all biochemical, signaling, and functional processes in cells (Alberts, 1998). Moreover, protein assemblies are in order of hundreds, for example, the single-cell *Saccharomyces cerevisiae* has more than 400 assemblies (Srihari et al., 2017). Also, it is known that proteins interact physically with other proteins and biomolecules; for instance, over 80% of proteins not function alone, they work as macromolecular assemblies (Berggård et al., 2007). Furthermore, with new low-cost protein sequencing technologies, there is a massive growth of sequences available, for example, UniProt has 100 million sequences and only 0.5 million are manually annotated (the rest are unlikely to be experimentally characterized) (Gligorijevic et al., 2020). Nowadays, the knowledge of functional roles of proteins is one of the most important problems in bioinformatic.

In this research proposal, we present a deep graph convolutional network to predict network-based protein functions. In Section 2, we present the most relevant works based on protein function prediction. Moreover, in Section 3, we present our proposal based on graph convolutional networks.

2 State of art

Network-based protein function prediction, used networks information. For instance, network-based protein structural, represent nodes are amino acids and edges link amino acids that are spatially close. For example, Newaz et al. (2018) and Ghalehnovi (2019), used graphlets as features along with deep learning.

Another approach takes into account PPI networks; they consider proteins as nodes and edges represent the interaction between proteins. For example, Freschi (2007) annotated functions of protein interactions with a random walk algorithm. Also, Vascon et al. (2020), used a graph-transduction game. Moreover, Vazquez et al. (2003) used PPI networks to assigned the protein's function, according to the network physical interaction; they used their method to analyze the yeast *Saccharomyces cerevisiae* PPI. Finally, there are some platforms, like NetGO, this tool used massive protein-protein network information to automated function prediction (You et al., 2019). There is also, FunCoup, a web framework to infer genome-wide functional coupling (Ogris et al., 2018).

Furthermore, several studies have been proposed in the field of PPI networks. Kovács et al. (2019) studied protein interactions, they offered structural and evolutionary evidence that proteins interact despite their dissimilarity. Moreover, Jia et al. (2019), used Chaos Game Representations (CGR) to predict protein-protein interactions. Furthermore, the use of deep learning has been applied extensively to predict protein interactions Wang et al. (2019); Zhang et al. (2019); Zeng et al. (2020).

3 Proposal

In this research, a Graph Convolutional Neural Network (GCN) is proposed for protein function prediction based on protein-protein interactions. I will apply the method proposed by Susha (2019), where a PPI network is represented as a weighted attributed graph, where each protein is represented as a node and protein-protein interactions are edges. Then, we will apply a series of convolutional layers followed by pooling layers. Finally, we will classify the network/graph with its corresponding function. We will use the databases proposed in You et al. (2019).

The convolution operation is the main challenge in this research. In this case, we need to find a new representation for each vertex by aggregating the attributes of its neighbors. Susha (2019) update each vertex with information of the vertices in k-order proximity. For example in Figure 1 (right), we represent the vertices in 1-order proximity of red vertex.

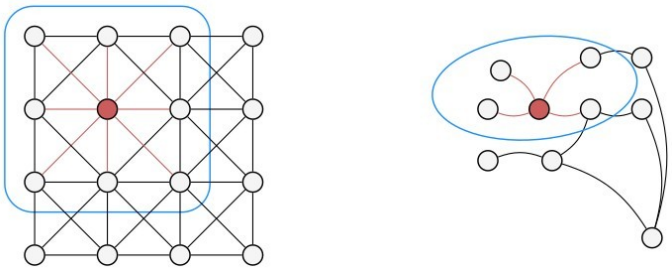

Figure 1: Analogous of convolution on images and graphs. Left: CNN operator for images. Right: CNN operator for graphs. Source: Susha (2019)

3.1 Why am I interested for the project?

The knowledge of protein function and protein-protein interactions are the key point to understand the nature of biology. Classical methods are too expensive, nevertheless, computer science could help. I'm a computer science researcher passionate in the understanding of protein/gene functions and the future applications of this field.

3.2 How will I complete the project?

I have scheduled my activities. Also, I have defined milestones in order to accomplish the project (Table 1). Now, I'm in the literature review and the implementation of the first method that is based on graph convolutional networks. Moreover, I planned to use autoencoders as the second method.

Table 1: Activities for the research proposal.

Activities	trimester 1	trimester 2	trimester 3	trimester 4
Literature review	x	x		
Method 1 implementation	x	x		
Evaluation of results		x		
Paper redaction and submition		x		
Method 2 implementation		x	x	
Evaluation of results				x
Paper redaction and submition				x

3.3 What makes me suitable to complete the project?

I have computer science aptitudes to accomplish the project. Moreover, I have experience in research projects. I consider myself, very passionate about the field of PPI networks and its applications, even if I were not selected, I am going to continue the research because is the field I choose to research.

References

Alberts, B. (1998). The cell as a collection of protein machines: preparing the next generation of molecular biologists. *cell*, 92(3):291–294.

Berggård, T., Linse, S., and James, P. (2007). Methods for the detection and analysis of protein–protein interactions. *Proteomics*, 7(16):2833–2842.

Freschi, V. (2007). Protein function prediction from interaction networks using a random walk ranking algorithm. In *2007 IEEE 7th International Symposium on BioInformatics and BioEngineering*, pages 42–48. IEEE.

Ghalehnovi, M. (2019). *Novel Computational Approaches for Network-Based Protein Structural Classification*. PhD thesis, University Of Notre Dame.

Gligorijevic, V., Renfrew, P. D., Kosciolek, T., Leman, J. K., Berenberg, D., Vatanen, T., Chandler, C., Taylor, B. C., Fisk, I. M., Vlamakis, H., et al. (2020). Structure-based function prediction using graph convolutional networks. *bioRxiv*, page 786236.

Jia, J., Li, X., Qiu, W., Xiao, X., and Chou, K.-C. (2019). ippi-pseaac (cgr): Identify protein-protein interactions by incorporating chaos game representation into pseaac. *Journal of theoretical biology*, 460:195–203.

Kovács, I. A., Luck, K., Spirohn, K., Wang, Y., Pollis, C., Schlabach, S., Bian, W., Kim, D.-K., Kishore, N., Hao, T., et al. (2019). Network-based prediction of protein interactions. *Nature communications*, 10(1):1–8.

Newaz, K., Ghalehnovi, M., Rahnama, A., Antsaklis, P. J., and Milenković, T. (2018). Network-based protein structural classification. *Royal Society Open Science*, 7(6):191461.

Ogris, C., Guala, D., Kaduk, M., and Sonnhammer, E. L. (2018). Funcoup 4: new species, data, and visualization. *Nucleic acids research*, 46(D1):D601–D607.

Srihari, S., Yong, C. H., and Wong, L. (2017). *Computational prediction of protein complexes from protein interaction networks*. Morgan & Claypool.

Susha, P. (2019). *Attributed Graph Classification via Deep Graph Convolutional Neural Networks*. PhD thesis, University Of Windsor.

Vascon, S., Frasca, M., Tripodi, R., Valentini, G., and Pelillo, M. (2020). Protein function prediction as a graph-transduction game. *Pattern Recognition Letters*, 134:96–105.

Vazquez, A., Flammini, A., Maritan, A., and Vespignani, A. (2003). Global protein function prediction from protein-protein interaction networks. *Nature biotechnology*, 21(6):697–700.

Wang, L., Wang, H.-F., Liu, S.-R., Yan, X., and Song, K.-J. (2019). Predicting protein-protein interactions from matrix-based protein sequence using convolution neural network and feature-selective rotation forest. *Scientific reports*, 9(1):1–12.

You, R., Yao, S., Xiong, Y., Huang, X., Sun, F., Mamitsuka, H., and Zhu, S. (2019). Netgo: improving large-scale protein function prediction with massive network information. *Nucleic acids research*, 47(W1):W379–W387.

Zeng, M., Zhang, F., Wu, F.-X., Li, Y., Wang, J., and Li, M. (2020). Protein–protein interaction site prediction through combining local and global features with deep neural networks. *Bioinformatics*, 36(4):1114–1120.

Zhang, L., Yu, G., Xia, D., and Wang, J. (2019). Protein–protein interactions prediction based on ensemble deep neural networks. *Neurocomputing*, 324:10–19.