Name: Arghamitra Talukder

UIN: 726007850

Expected Graduation Date: May 2021

Project Title: Multimodal Data Fusion and Machine Learning for Deciphering Protein-Protein Interactions

Research Area: Computer Engineering/ Biomedical

Faculty Name: Yang Shen

Option: I

Arghamitra

Student Signature

Faculty Signature

- Project summary:

The structural interactions between proteins can be considered as the core of cellular processes and the partnership provides a lot of answers to decode Molecular Biology. As PPI is crucial to most cellular functions, their interactions in 3 dimensional space must be understood. Due to the huge quantity of unknown biological interactions, it is a next to impossible task to identify each physical protein interaction in person; not only because the identification presents an enormous quantity but also it is a costly and resource intensive job. Though from time to time various experimental and computational methods have been applied to predict the PPI, a knowledge gap is there to understand their 3 dimensional interactions. This research project aims to use the existing data and available tools of machine learning resulting in an algorithm to predict protein protein interaction. The algorithm will use data science to match the existing patterns with a model in the light of physics and biology. The success matric of the algorithm will be the accuracy of the testing and validation group of data; it also aims to cover a broad range of scope making it more versatile.

- Introduction:

A functional human body is made of a lot of active organs, different macro and micro molecules. One of the most important cellular molecules is protein. Protein contributes in most biological processes including genetic expression, intercellular communication, morphology, nutrition absorption and so on (Thomas et al). Proteins are made of a character string where the characters represent amino acid. The amino acids bond together in different configurations to express the specific functionalities as proteins. As the mechanisms of the human body are unrevealed, one context was very clear that the majority of the proteins interact with each other and to understand their behaviors they should be analyzed from the perspective of protein protein interaction. There are experimental methods as well as computational methods. Given the short scale application of the experimental techniques like affinity purification, yeast two hybrid, co-immunoprecipitation, computational methods are more suitable approaches to follow.

The computational method adopted homology based approaches like interolog search. Interolog search is based on the principle that interactions are conserved and interlogs are homologous pairs of protein interactions across different species. The homology based method also includes phylogenetic similarities which relates to the common ancestor proteins among species (Abbasi et al). The simulation based methods include protein docking. Protein docking is molecular modeling which predicts the mutual orientation (Tradigo et al). A lot of machine learning techniques have been also applied based on protein sequence, structure and function. The limitations with these approaches are the difficulties to model any conformational changes and lack of thorough understanding of the binding mechanism (Abbasi et al).

This project offers a computational method to predict the overview how the proteins interact in 3-dimensional space. An accurate PPI prediction model will serve a number of objectives

including: pathways for unknown proteins, different binding modes, specificity of protein based multiple targets, effectiveness of drugs, design of new protein etc.

- Goals:

The goal of this project is to develop an algorithm using machine learning to predict the physical protein protein interactions expressing them in different modalities like text, graph and image.

- Methodology:

We are proposing a novel perspective to approach the PPI with different modalities. Though a large amount of data is available on protein protein interaction, there is a gap to know how different proteins interact with each other in 3 dimensional space. This project is focused to develop a new algorithm to predict PPI interaction in 3d space using the present tools of artificial intelligence and data science. Our methodology involves protein representation in different modalities like text, graph and image and their non-covalent interactions in each form. The ultimate goal of the project is to extend the state of art with a balanced combination of data driven and physics constrained modelling. The success of the algorithm will be demonstrated by its prediction accuracy for the testing dataset and how well it can replicate the PPI in comparison to other methods. The success metric of the algorithm is also dependent on the scope of its applicability or how widely it applies to PPIs.

- Design Component:

Identify the data sources and acquire the data: I will make a list of the data I need. I will evaluate if I can use all of the data. It also needs to be determined which dataset can prove the proposed algorithm and which dataset can contradict the theory (Tyagi)

Develop a baseline model and then explore other models to shortlist the best ones: Next step would be trying the dataset with the very common models like linear regression alongside with the successful papers and procedures on PPI like ComplexContact, FilterDCA and Ouroboros.With the trial I can shortlist the state of art depending on the performance (Tyagi).

Fine-tune shortlisted models and check for ensemble methods: I can hyperparameter tuning using cross-validation. I will be developing the method and the model with individual proof of concept. In case the model is working as expected, I need to investigate the reason or parameters which make a difference or make the results good (Tyagi).

Document Code and Communicate the solution: I will be documenting each and every step in an organized way to keep track of the project (Tyagi).

- Reference:

Tradigo, Giuseppe, et al. “Algorithms for Structure Comparison and Analysis: Docking.” *Encyclopedia of Bioinformatics and Computational Biology*, Academic Press, 6 Sept. 2018, www.sciencedirect.com/science/article/pii/B9780128096338204858.

Abbasi, Wajid Arshad & Minhas, Fayyaz ul Amir Afsar. (2018). Problems in Protein-Protein Interactions (A Literature Review).

Thomas, Neil, et al. “Can We Learn the Language of Proteins?” *The Berkeley Artificial Intelligence Research Blog*, 4 Nov. 2019, bair.berkeley.edu/blog/2019/11/04/proteins/.

Tyagi, Harshit. “Task Cheat Sheet for Almost Every Machine Learning Project” 4 July, 2020, https://towardsdatascience.com/task-cheatsheet-for-almost-every-machine-learning-project-d0946861c6d0