Name: Arghamitra Talukder

UIN: 726007850

Expected Graduation Date: May 2021

**Research Title:** **Physics-Constrained Modeling of Molecular Texts, Graphs, and Images for Deciphering Protein-Protein Interactions** (!!)

Research Area: Computer Engineering/ Biomedical(!!)

Faculty Name: Yang Shen

Option: I(!!)



|  |  |
| --- | --- |
| Student Signature | Faculty Signature |
|  |  |

* 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. As a very simple example: while someone reads this passage, our heart is pumping blood throughout the body and hemoglobin (a protein) is carrying the oxygen with the blood; transporter proteins are enabling neurons to exchange signals into the brain and photoreceptor proteins are recognizing the boundaries of the letters (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 that is why to understand their behaviors they should be analyzed from the perspective of protein protein interaction.

* Problem statement:

As PPI is crucial to most cellular functions, their interactions in 3 dimensional space must be understood. Every year more and more protein structures are being solved but the growth is exponential. Thus 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. Thus 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 new protein etc.

* Our proposal:

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 attachment in each form. The ultimate goal of the project is to extend the state of art with a balanced combination of data dependency and physics constrained modelling.

The success of the algorithm will be dependent on its prediction accuracy and how well it can replicate the PPI in comparison with the testing dataset. The success matric of the algorithm is also dependent on the wide scope of PPI modelling or how many models it can get right.

* Previous work:

Deciphering PPI has been an upcoming challenge for a while and a couple of approaches were taken to solve the problem. 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. Though experimental methods would result in a more accurate detection, they are often expensive and time consuming.

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. thus they are functionally related and involved in the same biological process. There are some drawbacks of these methodologies; they are dependent on homologs. These methods are also unfit for proteins with diverse targets (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).

* 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](http://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/.