Computational Analysis of Large and Heterogeneous Biological Networks

Thesis submitted by

Soumyendu Sekhar Bandyopadhyay

Doctor of Philosophy(Engineering)

Department of Computer Science and Engineering
Faculty Council of Engineering & Technology
Jadavpur University
Kolkata-700032, India
2023

Index No. 99/19/E

Abstract

Most cellular functions are performed by proteins, which are linear polymers of amino acids. Proteins may consist of several hundred to several thousand individual amino acids held together by peptide bonds. The physical relationship of two or more proteins is referred to as protein-protein interaction (PPI) and plays a critical role in the regulation of cellular activities, signaling pathways, and disease mechanisms. Protein domains, motifs, and surfaces all play important roles in PPIs by allowing proteins to recognize and interact with one another. The volume and variety of heterogeneous data in biology are increasing exponentially. With the increase of protein interaction, the mining techniques of key protein clusters, homology analysis between functionally similar proteins, and assessment of interaction affinity between proteins are causing an immense surge of information which is becoming increasingly difficult to process due to limited computational resources. Most of the information of life is encoded inside the DNA of a cell, from which proteins are synthesized. The main challenge in dealing with this huge biological dataset is the way we store and process them. Big Data analysis will enable us to process this data with its different analytical techniques. Thus, the need of the hour is to harness technologies, like Big Data framework, which will help distribute computations over a group of nodes and hasten the process of data analysis. The research work embodied in this thesis has addressed the problem of computational analysis of large and heterogeneous biological networks that mainly focuses on five major verticals: *In-silico* analysis of large-scale human protein sequences using Big Data framework, analysis of large-scale human proteome at fuzzy semantic space, computational modeling of human-nCoV PPIN, developing *in-silico* methods of drug repurposing for COVID-19, and prediction of PTM sites in protein sequences.