Protein Protein Interaction Networks

Abstract

Proteins are the large complex bio-molecules that perform most of the essential tasks in an organism and are required for the structure, function and regulation of the body's tissues and organs. Proteins do not function in isolation. They interact with one another to perform various cellular and metabolic processes in an organism. These interactions are essential for almost all the processes in a cell and thus understanding these interactions is very crucial.
In this project we visualized these interactions as graph networks known as Protein-Protein interaction networks and studied various properties of these networks. We took the protein interaction data of human bone tissue under normal state and cancerous state and modeled them as graph networks. We calculated various graph related parameters on these networks and found those parameters which can be used to differentiate normal bone network from cancerous bone networks. We also took the data set of Human-EBV (Epstein-Barr Virus) interaction and processed it to visualize the virus pathogenesis at the molecular level.

1 Introduction

Proteins are any of a class of nitrogenous organic compounds that consist of large molecules composed of one or more long chains of amino acids and are an essential part of all living organisms, especially as structural components of body tissues such as muscle, hair, collagen, etc., and as enzymes and antibodies. There are 20 different types of amino acids that can be combined to make a protein. Proteins can be described according to their large range of functions in the body, for example -

- Antibodies bind to specific foreign particles, such as viruses and bacteria, to help protect the body. Eg. ImmunoglobulinG.
- Enzymes carry out almost all of the thousands of chemical reactions that take place in cells. They also assist with the formation of new molecules by reading the genetic information stored in DNA. Eg. Phenylalanine hydroxylase.
- Messenger proteins, such as some types of hormones, transmit signals to coordinate biological processes between different cells, tissues, and organs. Eg. Growth Hormones
- These proteins provide structure and support for cells. On a larger scale, they also allow the body to move. Eg. Actin
- These proteins bind and carry atoms and small molecules within cells and throughout the body. Eg. Ferritin

Proteins interact with one another and also with other molecules that plays a significant role in biological function; it is these interactions that control mechanisms leading to healthy and diseased states. Protein-Protein Interactions (PPIs) are physical interactions between proteins which could be as a result of electrostatic forces, chemical reactions, covalent bonding or even interaction with water.

2 Background

2.1 Protein-Protein Interaction Networks

The mathematical representations of the interactions taking place between various proteins in a cell are collectively known as Protein-Protein Interaction Networks(PPINs). These networks are represented as undirected graphs where the nodes represent the interacting proteins and the edges represent the interaction between these proteins. The totality of protein interactions that happen in

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

an organism is known as an Interactome. For Eg: Human Interactome, Yeast Interactome etc. Any aberrations in the interactome might lead to diseases, for eg. Congenital Heart Disease, Alzheimer's Disease.
The development of large scale PPI screening techniques has resulted in enormous increase in PPI data resulting in development of large and complex interactomes.

2.2 Experimental Identification of Protein-Protein Interactions

The edges in PPINs represent the interaction between two proteins. These interactions are actually identified by various Bio-Physical and Automated methods performed in laboratories.The most widely used method is Yeast Two-Hybrid(Y2H) which is one of the high throughput method in this field.We describe this method in detail below:

Yeast Two-Hybrid (Y2H): It is one of the most commonly used molecular biology technique to identify PPIs by testing for physical interactions between 2 proteins or a protein molecule and a DNA molecule. If the proteins interact, there is a transcription on the reporter gene which can be detected. The premise behind the experiment is that binding of a transcription factor to an upstream activating sequence leading to activation of downstream reporter gene. Transcription factor is the protein controlling the rate of copying of a particular DNA sequence to RNA. Upstream and downstream are relative positioning of genetic code in DNA or RNA. Upstream Activating Sequence (UAS) is the DNA sequence regulating the transcription rate in neighbouring genes. Reporter Gene is the gene attached to another regulatory gene under study (Eg. bacteria, plants etc.). In Figure 1, Gal4 transcription factor gene produces a 2 domain protein, Active Domain (AD) and Binding Domain (BD) essential for transcription on reporter gene (LacZ). The BD is the domain responsible for binding to the UAS and the AD is the domain responsible for the activation of transcription. When Gal4 interacts with the bait and prey separately, there is no activation on the reporter gene (Lacz). However, when Gal4 interacts with the fusion of bait-prey, there is transcription on the reporter gene, indicating there is interaction between the bait protein and the prey protein.

2.3 Properties of PPINs

The protein interaction networks exhibit certain important properties which have crucial biological consequences.These properties are defined as follows:

2.3.1 Small World Effect

The PPINs show small world effect i.e the proteins are highly connected. This also means that the graph's diameter is small no matter how large the graph is in terms of number of edges and nodes. This high connectivity has an important biological implication -: It results in highly efficient and quick flow of biological signals within the network necessary for various metabolic processes in the cell. But this high level of connectivity also raises an important question that if the network is so highly connected then why don't perturbations in a single protein or gene has dramatic consequences on the network.The answer is that our biological systems are robust enough to cope up with perturbations in few proteins in the network. This is due to another important property of PPINs known as Scale Free Networks, described in the next section.

2.3.2 Scale Free Networks

The PPINs are scale free networks which means that the majority of the nodes(proteins) in these networks are connected to few other nodes while some few proteins are connected to large number of other proteins. These high degree nodes are known as hubs and the corresponding proteins are known as hub proteins. The scale free property of the PPINs has some important biological consequences.It provides stability to these networks.If some failure occurs at random then the probability that a hub protein will be affected is less as the majority of proteins are those having small degrees. Also if a hub protein is affected then the connectivity of the graph be still maintained by the presence of other proteins.However the scale free property also makes the network vulnerable to targeted networks. If we lose some more hubs then the graph is turned into smaller isolated subgraphs.The hub proteins contain the essential genes. Most of the cancer linked proteins are the hub proteins. We observed the degree distribution of the PPIN of normal human bone and observed that the degree distribution of graph followed the Power Law indicating scale free property (Figure 2).

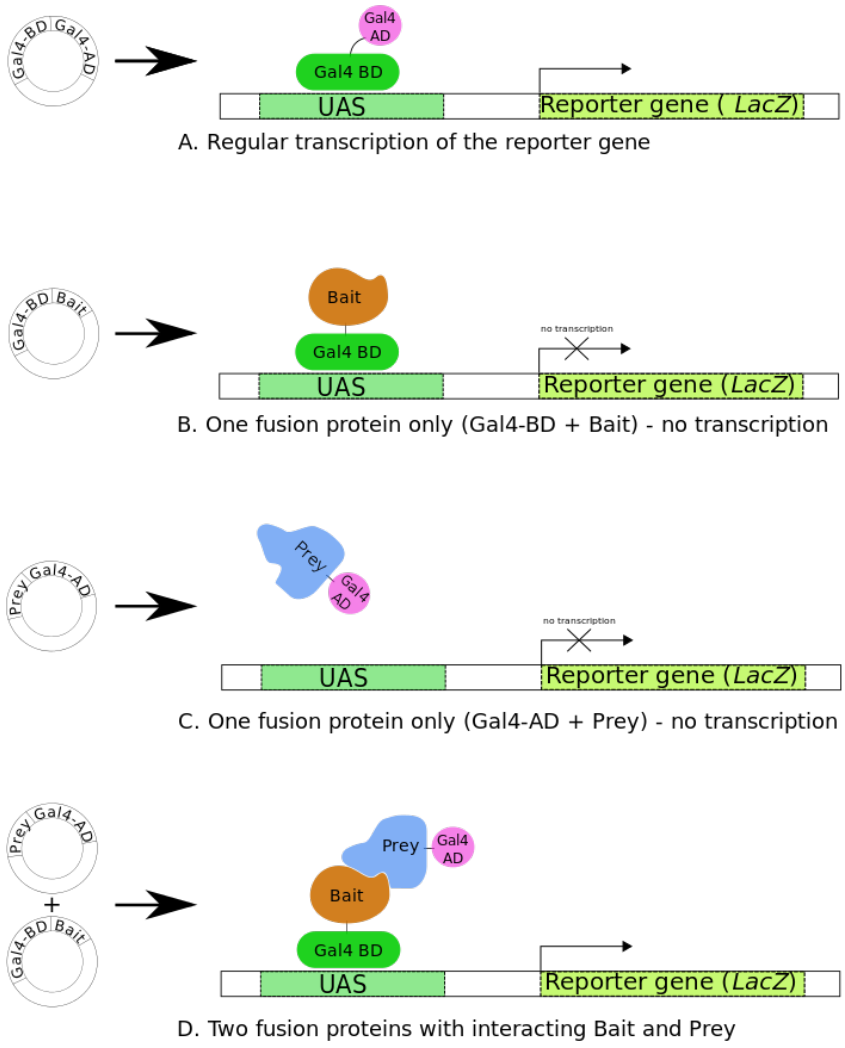


Figure 1: Y2H method for identification of PPINS

2.3.3 Transitivity

The transitivity or the Clustering Coefficient of a graph is defined as the measure of the tendency of nodes to cluster together.High transitivity means that the network contains communities or groups of nodes that are densely connected internally. In Biology these communities or clusters reflect functional modules and protein complexes. Protein complexes can be considered a type of module in which proteins are interacting with each other in a stable manner. These modules may represent protein communities of specific organs. For Example : The proteins found in one organ may be clustered together more densely than the proteins found in some another organs. There are some connecting or bridge proteins between these modules.

2.4 Network Parameters

In this project we calculated various graph parameters on the PPINs of normal bone tissue and cancerous bone tissue to identify those parameters which are significantly different in both of these networks and thus can be used to distinguish normal state from cancerous state. We computed the following parameters on the PPINs :

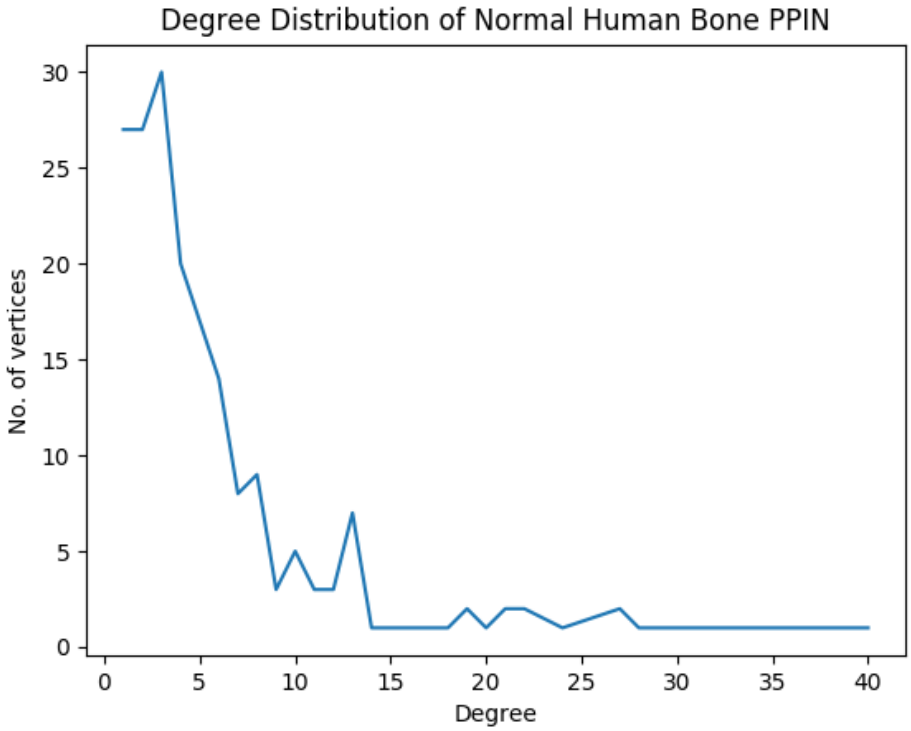


Figure 2: Degree Distribution of Normal Human Bone PPIN

2.4.1 Graph Diameter

The diameter of a graph is defined as the shortest distance between the two most distant nodes in the network. Once the shortest path length from every node to all other nodes is calculated, the diameter is the longest of all the calculated path lengths.

2.4.2 Connected Components

A connected component of an undirected graph is a maximal set of nodes such that each pair of nodes is connected by a path.

2.4.3 Clustering Coefficient

Clustering coefficient is a measure of the degree to which nodes in a graph tend to cluster together.The local clustering coefficient C_i for a vertex v_i is given by the proportion of links between the vertices within its neighbourhood divided by the number of links that could possibly exist between them.Network average clustering coefficient is defined as the average of the local clustering coefficients of all the vertices n :

$$\bar{C} = \frac{1}{n}\sum_{i=1}^{n} C_i. \tag{1}$$

2.4.4 Connectivity

Connectivity of a graph can be expressed as follows :

$$C = \frac{A}{B} \tag{2}$$

where A is the total number of edges realized in a given graph and B represents the maximum number of possible edges.

2.4.5 Betweenness Centrality

Betweenness centrality in a graph is based on the shortest paths. It represents the number of time a vertex acts as a bridge along the shortest paths between two other vertices. The betweenness centrality of a node v is given by the expression:

$$g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{3}$$

where σ_{st} is the total number of shortest paths from node s to node t and $\sigma_{st}(v)$is the number of those paths that pass through v.

2.4.6 Closeness Centrality

The closeness centrality of a node in a connected graph defined as the reciprocal of the sum of the length of the shortest paths between the node and all other nodes in the graph. The more central a node is, the more close it is to all other nodes. Mathematically the closeness centrality of a node x is defined as :

$$C(x) = \frac{1}{\sum_y d(y, x)} \tag{4}$$

where $d(y, x)$ is the distance between vertices x and y.

2.4.7 Degree Centrality

In an undirected graph, the degree centrality of a node is defined as the number of edges incident on it. The degree centrality of a vertex v, is defined as:

$$C_D(v) = \deg(v) \tag{5}$$

The proteins having high degree (greater than a threshold) are known as hub proteins.

3 Experiments

3.1 Comparison of PPINs of Normal and Cancerous Human Bone Tissue

3.1.1 Data

The data set for the protein interactions in Normal and Cancerous bone tissues was obtained from [1]. The original data consisted of 8246 rows. Each row consisted of a pair of protein and their corresponding expression values.

3.1.2 Pre-processing

In the data set the expression data for proteins was expressed in Digital Expression Unit (DEU). We considered the DEU greater than zero as 1 and DEU equal to zero as 0. For a pair of proteins where both the proteins had DEU equal to 1, we considered it to be a valid interaction and retained that pair. Otherwise if one or both the proteins in a pair had DEU as 0 meaning either one or both were unexpressed, we discarded that pair. In this way we pre-processed the raw data set to consider the valid protein interactions for both normal as well as cancerous bone tissues.

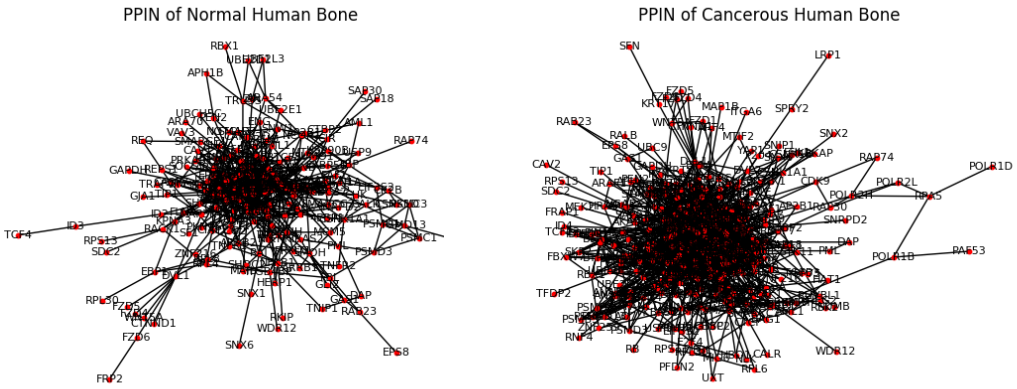
3.1.3 Experimental Setup

The experiment was carried out using Python 2.7.10 and libraries - Numpy, Pandas, Matplotlib, Networkx, Operator, Collections and coded on MacOS - Mojave (version 10.14).

3.2 Epstein-Barr Virus (EBV) Pathogenesis

3.2.1 Data

The data set for the protein interactions for EBV Pathogenesis is obtained from [2]. The data set contains an interactome of nodes representing EBV and Human proteins and the interaction between them.



(a) Normal Human Bone (b) Cancerous Human Bone

Figure 3: PPINs of Normal and Cancerous Human Bone Tissue

Network Parameter	Normal Bone PPIN	Cancerous Bone PPIN
Number of Nodes	192	351
Number of Edges	619	1783
Connectivity	0.03375	0.02902
Average Clustering Coefficient	0.20992	0.229
Diameter	7	7
Connected Components	1	1
Average number of neighbors	6.44791	10.15954
Number of Hub Proteins (degree>=12)	29	95

Figure 4: Network Parameters for Normal and Cancerous Bone PPINs

3.2.2 Experimental Setup

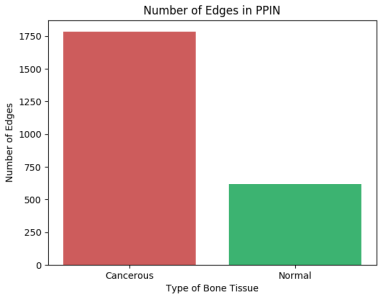
The experiment was carried out using Cytoscape 2.8.0 for visualizing the Human - EBV protein interactions and analyzing the network.

4 Results

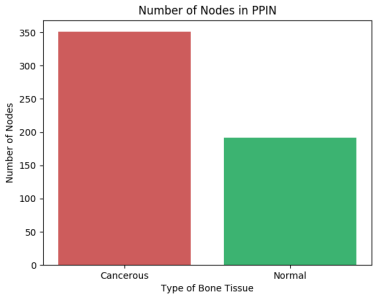
The results of the above two experiments were obtained as follows:

4.1 Comparison of PPINs of Normal and Cancerous Human Bone Tissue

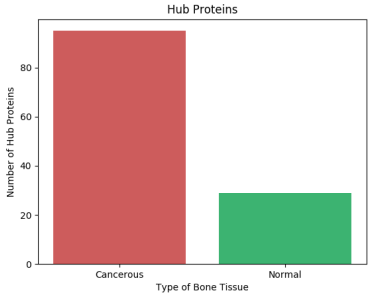
We constructed and visualized the PPINs for Normal bone tissue and Cancerous bone tissue (Figure 3). The proteins having degree greater or equal to 12 were considered as hub proteins. The parameters considered in the study were clustering coefficient, connected components, network diameter, number of nodes, number of hub proteins, number of edges, connectivity, average number of neighbors, degree centrality, betweenness centrality and closeness centrality. The figure 4 shows the values observed for these parameters in both of the networks. Figure 5 shows the comparative view of these parameters for normal and cancerous bone networks. Figure 6 shows some important proteins found found in both PPINs based on degree centrality, closeness centrality and betweenness centrality.



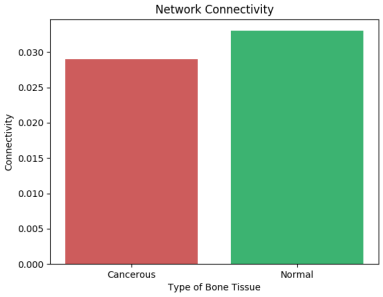
(a) Number of Edges



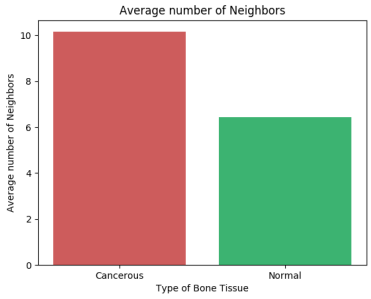
(b) Number of Nodes



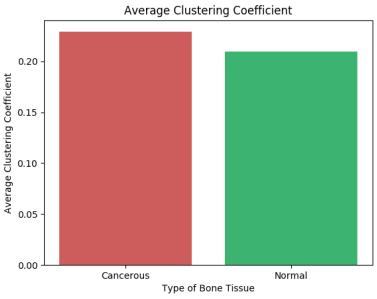
(c) Number of Hub Proteins



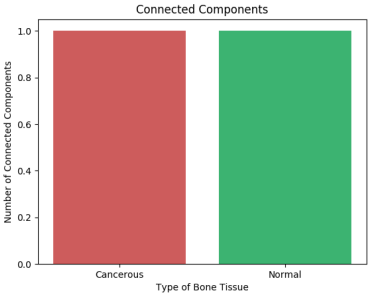
(d) Connectivity



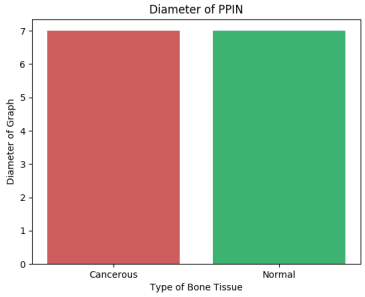
(e) Average Number of Neighbors



(f) Average Clustering Coefficient



(g) Connected Components



(h) Graph Diameter

Figure 5: Comparative view of Normal and Cancerous Bone PPINs

Normal Bone	Cancer Bone
SMAD3 – 40	STAT1 – 66
FYN – 38	FYN – 63
ERK1 – 32	SMAD3- 63
HDAC1 – 35	STAT5B – 50
RELA – 28	ERK1 - 50

(a) Degree Centrality

Normal Bone	Cancer Bone
HDAC1 – 0.1426	SMAD3 – 0.0936
CTNNB1 – 0.1329	FYN – 0.0703
SMAD3 – 0.1323	HDAC1 – 0.0630
FYN – 0.1201	CTNNB1 – 0.0601
ERK1 – 0.0993	STAT5B – 0.0561

(b) Betweenness Centrality

Normal Bone	Cancer Bone
4EBP1 – 0.3026	4EBP1 – 0.3196
AKT1 – 0.3797	AKT1 – 0.3954
CCND1 – 0.3827	CCND1 – 0.3867
14-3-3-zeta – 0.3485	MAPK14 – 0.4156
CSNK1A1 – 0.3046	TFDP1 – 0.2829

(c) Closeness Centrality

Figure 6: Key proteins with respect to various centrality measures

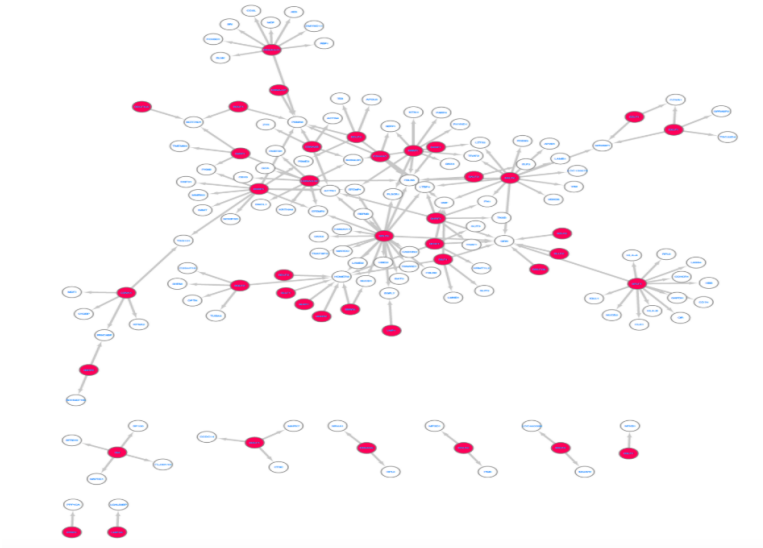


Figure 7: EBV-Human Interactome

4.2 Epstein-Barr Virus (EBV) Pathogenesis

We visualized the network for the EBV-Human interactome as seen in Figure 7 (The red nodes represent the EBV proteins). We observed that there were a total of 152 proteins out of which 112 were Human and 42 were EBV unique proteins. There were a total of 173 interactions which was indicated by the edges in the figure. Out of these interactions, 24 are multi-targeted Human proteins which interact with more than 1 EBV protein.

5 Conclusion

The experimental results showed that the cancerous bone's PPINs are more denser than that of normal bone's. Cancerous networks have significantly more edges and more nodes than normal networks. This concurs with the fact that more and more proteins get expressed in cancerous tissues due to genetic mutations and there is a multifold increase in their interactions. The clustering coefficient was also found higher in cancerous tissues than in normal tissues and so was the average number of neighbours.The high degree proteins(hub) were also significantly higher in cancerous networks. This may be attributed to the growth of cancer tumor where rapid and uncontrollable cell division leads to expression of more proteins which is undesirable. Therefore these parameters can be used as distinguishing factors for cancerous and normal bone tissues and can be used as a way to detect cancer to some degree.The hub proteins contain the essential genes. Most of the cancer linked proteins are the hub proteins. The proteins having higher centrality measures are important from the therapeutic point of view. These proteins lie on signal transduction path and control the information flow.Thus these proteins are important proteins in signalling pathways and can be used as new drug targets for research in advanced cancer treatments.

In EBV pathogenesis, we observed that the Multi Targeted human proteins (which interact with more than one EBV proteins) are crucial for Virus Life Cycle metabolism and thus can be used for further analysis. These proteins are known as EBV-Targetted Human Proteins(ET-HPs) The average degree of ET-HPs(approx. 15) is significantly larger than other human proteins(approx. 6) and thus they are hub proteins. EBV targets ET-HPs as they are key for human metabolism and are highly interconnected to maximize efficiency of biological processes. This choice of proteins is best for virus so that it can hijack host metabolism and use it for its own survival.

We observed that various network science concepts are very much relevant in the field of biology and these concepts can be applied for both diagnosis of diseases and drug discovery.

References

[1] http://www.iaees.org/publications/journals/nb/articles/2013-3(1)/2-Islam-Abstract.asp

[2] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3531279/

[3] Mileidy W.Gonzalez, Maricel G.Kann, Protein Interactions and Disease

[4] K. M. Taufiqur Rahman, Md. Fahmid Islam, Rajat Suvra Banik, Ummay Honi, Farhana Sharmin Diba,Sharmin Sultana Sumi, Shah Md. Tamim Kabir, Md. Shamim Akhter, Changes in protein interaction networks between normal and cancer conditions: Total chaos or ordered disorder?

[5] https://www.wikipedia.org/

[6] Anna, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=2890233

6 Team Members

Aditya Joshi (ajoshi6)
Karthikeyan Vaideswaran (kvaides)

7 GitHub

https://github.ncsu.edu/kvaides/PPIN