Graph Theoretic Investigations of Control in Biological Networks

by

VANDANA RAVINDRAN
201221013

A Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY

to

DHIRUBHAI AMBANI INSTITUTE OF INFORMATION AND COMMUNICATION TECHNOLOGY

January, 2018

Declaration

I hereby declare that

i) the thesis comprises of my original work towards the degree of DOCTOR OF PHILOSOPHY at Dhirubhai Ambani Institute of Information and Communication Technology and has not been submitted elsewhere for a degree,

ii) due acknowledgment has been made in the text to all the reference material used.

Vandana Ravindran

Certificate

This is to certify that the thesis work entitled GRAPH THEORETIC INVESTIGATIONS OF CONTROL IN BIOLOGICAL NETWORKS has been carried out by VANDANA RAVINDRAN for the degree of DOCTOR OF PHILOSOPHY at *Dhirubhai Ambani Institute of Information and Communication Technology* under our supervision.

V Sunitha
Thesis Supervisor

Ganesh Bagler
Thesis Co-Supervisor

Acknowledgments

This was probably the hardest part of the thesis to put in words. So many people have contributed to making this possible- I am indebted to them all (while attempting to acknowledge everyone, my sincere apologies for missing anyone out).

This has been a personal journey for me of discovering my love for mathematics which I always feared for, and none of this would have been possible without my supervisor, Prof. Sunitha. I would like to thank her for giving me the opportunity to work with her after all the hardships faced during the course. I am eternally indebted to her for her support, encouragement and inspiration. I will always be in admiration and awe for her contagious enthusiasm. Thank you Prof. Sunitha for making me the researcher I am. I couldn't have asked for a better place to begin my scientific journey.

My sincere thanks to Prof. Ganesh, who co-supervised me. Thank you for your support, encouragement and ideas. I have learnt an immeasurable amount of science from you and I am eternally grateful for that. Thanks for all the scientific and philosophical insights that kept motivating me to pursue my PhD. I will cherish the endless discussions we have had. You have always been a great support to me at all times. A very special thanks to Prof. Mukesh Tiwari, Prof. Srikrishna Divakaran for being the coolest research progress committee and a source of ideas to tackle the realm of "network controllability". Special thanks to our witty Prof. Rahul for his support and lively discussions. I wish to acknowledge all the professors of DA-IICT who have inspired me directly or indirectly.

I am very grateful to Prof. David Robertson, Dr. Jean-Marc Schwartz, our collaborators on the HIV-1 human molecular interactome project. Thank for the

support and opportunity provided to me in Manchester and Prof Jose Nacher for being a great collaborators and source of ideas to carry out this work.

Special thanks to DST Inspire fellowship for funding my PhD. I am very grateful to Prof. K.S Dasgupta (Director), Prof. N Ramrao (former Director), Mr. Soman Nair (Registrar), Prof. Suman Mitra (Dean-Academics), Prof. Aditya Tatu (former Ph.D.Coordinator), Prof. Anish Mathuria (former Dean-R&D), Prof. L Pillutla (ICT convenor) who have helped me throughout my time at DA-IICT.

This journey would have been incomplete without the support and encouragement of the many fellow researchers. Trying not to leave anyone out- A very special thanks to Milind (LaTeX boy) for his support- with coding and witty humor. A very warm thanks to Ramnaresh, Sarita, Miral, Purvi (PK), Mahipal, Dicksha (IIIT-delhi) and Rahul (IIT-Jodhpur) for help on numerous occasions and for their support and helpful discussions. Special thanks to Nupur, Nidhi, Purvi (PP), Archana, Sumukh, Nirmesh, Parth, Prashant, Kamal and Jainisha for being great party buddies, Visharad and Adrian for being a part of this work. Thanks also to Sneha, Dixita and Trupti for being great hostel buddies and my extended family at DA-IICT. Thanks for being there through all the ups and downs. Thanks to DA-IICT staff for all the support.

I cannot express in words the gratitude I have for the support from my parents. Thank you Amma and Appa for everything and more. Thank you for always supporting me and believing in me. Finally, my sincerest thanks to Apurv. Thanks for your undying love and support and for being there through it all and for making this journey with me.

Abstract

Biological systems exhibit complex phenomena owing to interconnected molecular mechanisms underlying their architecture. Graph is an ideal mathematical abstraction of such systems. The abstract network along with the dynamics of the system forms a complex network. This thesis focuses on modelling and analysis of biological networks from control systems perspective. We first model the system as a directed network and use the technique of maximum matching to identify driver nodes which help in structurally controlling the system. We investigate the cancer signalling network and propose that identification of specific molecules as drivers of regulatory dynamics is a promising step towards targeted cancer therapies. We also capture the role of driver nodes in the HIV-1-human molecular interactome and show the efficiency with which the virus hijacks the host system for effective pathogenesis. In order to achieve a higher level of understanding of control and its implications in biology, we model five different networks ranging from disease to infection to normal regulation. We find that they are characterized with distributed control and with a large fraction of nodes acting as driver nodes. This implies that such networks are difficult to control. We further investigate the structure of driver nodes in these networks to characterize their control profile. Based on these investigations, we propose that, structural controllability applied to networks can lead to novel understanding of disease mechanisms in a more nuanced way compared to other network analysis. Our work provides a snapshot for control in biological systems assuming that the systems operate under homeostasis. We believe that this approach that amalgamates engineering concepts with biological knowledge can provide better insights into cellular mechanisms of a cell.

Keywords: Control theory, network controllability, driver nodes, maximum matching in directed graphs, biological networks, drug targets, viral hijack, control mode, control profile.

Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

1.1 Overview of network biology

Over the last few decades, research in molecular and cell biology has increasingly looked beyond reductionism and moved towards an integrated understanding of molecular and cellular systems. In biological systems the functionality of a living cell is carried out by an intricate network of biochemical, metabolic and information signalling processes, which involves interaction between various proteins and genes which in turn regulate thousands of other genes and proteins forming complexes that regulate the cell, which none of the individual constituents would be able to do. Hence, the true functionality of a cell cannot be determined by just one network but rather a set of interdependent networks ranging from the level of transcription to the process of metabolism. One can divide the cellular function into distinct networks [1, 2]:

1. **Transcription network or gene regulatory network (GRN)** that contains information concerning the control of gene expression in cells.

2. **Protein-protein interaction (PPI) network** that contains information of how different proteins coordinate with each other to enable biological processes within the cell.

3. **Signal transduction network** that represents series of interactions between different bioentities like proteins, macromolecules and chemicals that assist in signal transmission within and outside the cell.

4. **Metabolic and biochemical network** that represents the series of chemical reactions occurring within a cell at different time points.

Given the availability of high throughput techniques, we now have reliable data on various interaction maps. Some examples are gene regulatory networks that contains protein-DNA interaction data [3, 4], PPI networks that are organism specific like human [5, 6, 7, 8], yeast [9] and drosophila [10]. Signal transduction and metabolic data are also successfully mapped [11, 12, 13, 14] with the help of such high throughput experiments.

The approach of representing complex biological systems as networks gives rise to investigating such systems as a whole rather than as individual components and aims at simplification of biological processes for a better understanding of their behaviour.

Complex networks play an important role in various disciplines ranging from computer science, engineering, sociology, physics, epidemiology to molecular biology. Graph theory is a widely diploid and powerful tool for the characterisation of complex systems. The myriad components of a system and their interactions can be represented as graphs where, interacting components are represented by nodes/vertices and the interaction between these components as links/edges. This level of abstraction has made it possible to compare diverse networks and shown that several important properties are shared by them [15, 16]. Various models and methods from graph theory have been applied to study biological networks [2, 17, 18, 19] and has offered possibilities to understand the internal organization and evolution of a cell, fundamentally altering our view of cell biology [1]. The application of graph theory can go beyond mapping of the concrete network systems found within the cell and be used as means of organizing biological information in ways that could lead to new insights. The applications of network biology include identifying and determining some functions of genes/proteins [1, 20, 21], identifying disease genes and drug targets [22, 23, 24] thereby offering various strategies for treatment of diseases and better understanding of the mechanisms underlying a cell [25].

Diseases are not just due to abnormality of a single gene or protein but rather reflects an interplay of multiple molecular processes that are encoded as a network that integrates all physical interactions within a cell, from protein-protein to regulatory protein-DNA and metabolic interactions [26]. Despite the advances in interactome mapping and disease gene identification, our knowledge about the dynamics of the disease remains incomplete. Understanding how cellular systems are controlled is an important theme in biology and medicine. Given that biological networks are dynamical where the functional states at any given time point can be characterized by concentration of molecules like proteins, metabolites or genes. For instance, consider two genes A and B where the activation of gene A inhibits gene B. There are various such processes that govern the normal functioning of a cell and one can represent them as a dynamical system. Control theory deals with the design and analysis of dynamic systems that receive inputs and have outputs. The need to control engineered systems ranging from electric circuits, communication systems, manufacturing processes has resulted in a rich set of mathematical tools [27, 28] that offer ways to control the systems. Issues of control and regulation are central to biology as various signals within the cell drives it to a specific functional state suggesting that the need to control biological systems play a potential role in understanding the evolution and regulation of molecular networks.

1.2 Controllability of biological systems

A dynamical system is said to be controllable if with suitable choice of inputs the system can be driven from any initial state to any final state in finite time [27]. Several graph theory based models are available that can be applied to complex networks to identify the nodes and edges that can control the system [29, 30, 31, 32]. Studies have shown that biological networks exhibit properties of control systems such as feed-forward control, feed-back control and proportional action [33, 34, 35]. Recently, the connection between biological networks and structural control theory has been explored, uncovering interesting biological properties like essentiality, cellular localisation, evolutionary conservation and more.

Studies on control theory and biological networks have focused on two aspects, one describing the general network properties required for identifying the control nodes or edges and the other describing the biological properties associated with these control nodes and edges. Application of control theory using the maximum matching model on transcriptional regulatory networks reveal that it is difficult to control these networks due to relatively large number of nodes involved in control which are determined by degree distribution [30, 36]. On the other hand minimum dominating set model shows that heterogeneous networks with power-law degree exponent smaller than two require fewer nodes to control than the ones with power law exponent larger than two [31]. While both these models describe node controllability, dynamical process defined on edges have also been modelled [32] that have attempted to deepen our understanding of the interactions in a network. Other graph theoretic properties like closeness and betweenness centralities have been applied to better characterise control nodes in networks [37, 38, 39]. Few models have also been developed for the control of non-linear systems. For instance boolean models have been applied for biological network motifs to bring about a clear understanding of the interaction structure and non-linear dynamics [40]. Structural control and feedback vertex set control have been applied on dynamic models of gene-regulatory networks and provided a framework to identify nodes whose override can steer a network's dynamics towards any of its natural long term dynamical attractors [41].

Recent studies on control of biological networks have focused on identifying and characterising the minimal number of nodes whose receiving an external signal drives the system into a state of interest [30, 31]. These nodes are called driver nodes. Protein-protein interaction networks have been analysed, and driver nodes have been shown to be enriched with essential, cancer-related and virus targeted genes. Further their enrichment with regulatory functions like transcription factors, phosphorylation events and genetic interactions have also been studied [42, 43]. These studies have focused on static networks and fail to capture the underlying disease dynamics. Structural controllability analysis of human signalling network has shown the importance of driver nodes as upstream of signalling network and crucial for control [44]. A study by Uhart *et.al.* [45] on controllability of PPI phosphorylation network has shown that certain nodes that are critical for control are regulated by post-translational modifications. The same study and also characterised the edges based on these modifications. Controllability analysis on non-coding RNA (ncRNA)-protein interactions have been applied to identify human disease association with ncRNAs [46]. Only few of the studies have attempted to bring out biological significance among control nodes, while most of them have concentrated on applying various network science techniques to elucidate ways of identifying driver nodes using biological networks as merely examples of real-world networks.

1.3 Contributions of the thesis

From control systems perspective, cellular processes can be viewed as an intricately controlled orchestra of regulatory mechanisms that lend the cell its functional repertoire. Diseases, therefore, can be seen as the result of errors in cellular information processing. Beyond systems modelling of diseases, the focus has also been on finding ways of controlling the disease through therapeutic interventions. Further out reach of this work is an extension of the analogy to ’controllability’ of cell’s regulatory network, and its implications for control in disease as well as mechanisms of viral infection. We have bridged the controllability theorems for complex networks in an attempt to get a deeper understanding of control in biological networks. Aligned with these objectives, the following is the organization of the thesis:

Chapter 2: Methodology Description of the maximum matching algorithm implemented in this thesis.

Chapter 3: Identification of critical regulatory genes in cancer signaling network This chapter focuses on the controllability analysis of cancer signalling network.

Chapter 4: Identification of HIV-1 human molecular interactome Here we describe about our investigations on the HIV-1 human molecular network and show how control theory could provide a deeper understanding in viral hijack.

Chapter 5: Investigation of control configuration in biological networks This chapter focuses on the study of different control configurations in biological networks that determine their ease of control.

Chapter 6: Investigation of control profile in biological networks Here we investigate the structure of biological networks that determine why a node acts as a driver node.

Chapter 7: Conclusion In this chapter we give a summary of our findings based on control theory and the scope for extensions based on this thesis.

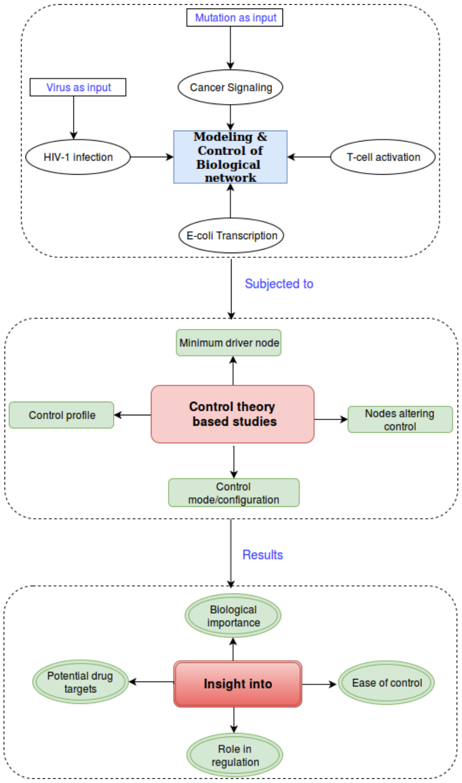


Figure 1.1: Pictorial outline of the thesis

We first model each of the systems as directed networks and use the technique of maximum matching to identify driver nodes [30]. Driver nodes are those nodes that when provided with inputs can steer the system from any initial state to a given final state in finite time [30, 31]. Figure 1 gives a schematic representation of various steps involved in the modelling and analysis of biological systems for investigating their controllability aspects. The following are salient contributions of this thesis:

- We have investigated the cancer signalling network and propose that identification of specific molecules as drivers of regulatory dynamics could be a promising step towards targeted cancer therapies. A special class of nodes identified as indispensable driver nodes could either steer the network from healthy state to disorder by means of mutations, or could be leveraged as drug targets for driving the network into healthy state [47]. We also find that indispensable nodes are preferentially targeted by anti-neoplastic drugs [48].

- We analysed the HIV-1 human molecular interaction as a dynamic network with and without viral interactions in order to understand the intricate and complex nature of viral infection. Existing models have focused on host proteins like hubs that have a direct interactions with viruses in network [49], while proteins that play an indirect yet crucial role may be poorly characterised. Our study investigates these indirect interactions like the receptors, that are not all direct viral targets yet important for viral entry, replication and as efficient drug targets. Further, we also propose the role of indispensable nodes as direct viral targets that could be an alternative to identify host dependent factors. This study attempts to provide insights on mechanisms involved in viral hijack.

- In order to achieve a higher level of understanding of control and their implications in biology, we modelled different networks ranging from disease to infection to normal regulation. The ease with which one can control a network primarily depends on the structure of the network. This determines the number of driver nodes needed for control [36]. We investigated five types of biological networks and find that they are characterized with distributed control and with a large fraction of nodes acting as driver nodes.

This implies that such networks are difficult to control.

- We further investigated the structure of driver nodes in these networks to characterise their control profile [50]. These studies offer insights into high-level organization and function of the networks. Control theory provides us an understanding of why biological networks are closed and help us explain their distributed nature of control.

Our study focuses on modelling and analysis of biological networks from a control systems perspective. We have proposed models of cancer and viral infection by considering how the system dynamics change from normal state to a diseased or infected state by applying the notion of controllability. Further we have also investigated other network controllability properties that determine the various control configurations and structures that determine the ease with which one can control the given biological system. Thus the study of biological networks from a control systems perspective is a novel method of understanding the regulatory mechanisms underlying these networks and provide a systematic way to identify drug targets.

CHAPTER 2

Methodology

2.1 Introduction

Our ultimate understanding of complex systems is reflected by our ability to control its behaviour. This requires a map of the network like the interactions between components, dynamical laws that govern the behaviour of the components and ability to influence the state of the system components. In most real systems like biological systems this information is inadequate. With recent advances in control theory and dynamical systems, notions of control and controllability have taken a new root in the study of complex networks [51]. This has inspired us to get an understanding of the mechanisms governing the behaviour of a complex system by answering questions like: What are the principles that govern the control of complex systems? Which are the components that can alter the state of the system? How do systems organize themselves? To address these questions many mathematical models have been developed some of these models are presented in this chapter.

Graph theoretical methods have been successfully applied to investigate the structural and quantitative properties of dynamical systems modeled as networks using control theory [52]. From a control systems perspective, a system is said to be controllable if we can drive the system from any initial state to any final state in finite time [27]. While earlier studies focused on smaller systems, the main interest lies in control of complex networks such as the Internet, WWW, power grids, gene regulatory networks, protein-protein interaction networks, transportation systems, communication systems. With recent advances in network science we can now characterize the structure of these systems. One of the important advances in this area is that of 'structural controllability' of complex networks, developed in the 1970s by Lin that offered necessary and sufficient conditions to check if any network with linear time invariant (LTI) dynamics is controllable [29].

The starting point for most control theoretical approaches is the LTI control system $(\mathbf{A}, \mathbf{B})$.

$$\dot{x}(t) = Ax(t) + Bu(t) \tag{2.1}$$

Consider the above LTI dynamics on a directed weighted network $G(A)$ of N nodes. Here $x_i(t) \in \mathbb{R}$ is the state variable, which can denote the transcription factor concentration in a gene regulatory network. $A = (a_{ij})_{N \times N} \in \mathbb{R}$ is the state matrix which is nothing but the adjacency matrix or wiring diagram of the network. $u(t) = u_1, \ldots, u_M (M \leq N)$ is the input signals or control signals applied to the network. $B = (b_{iM})_{N \times M}$ is the input matrix or the nodes that are controlled by u. According to Kalman [27], an LTI system (A, B) is controllable if and only if the controllability matrix

$$C \equiv [B, AB, A^2B, \ldots, A^{N-1}B] \tag{2.2}$$

has full rank, i.e.,

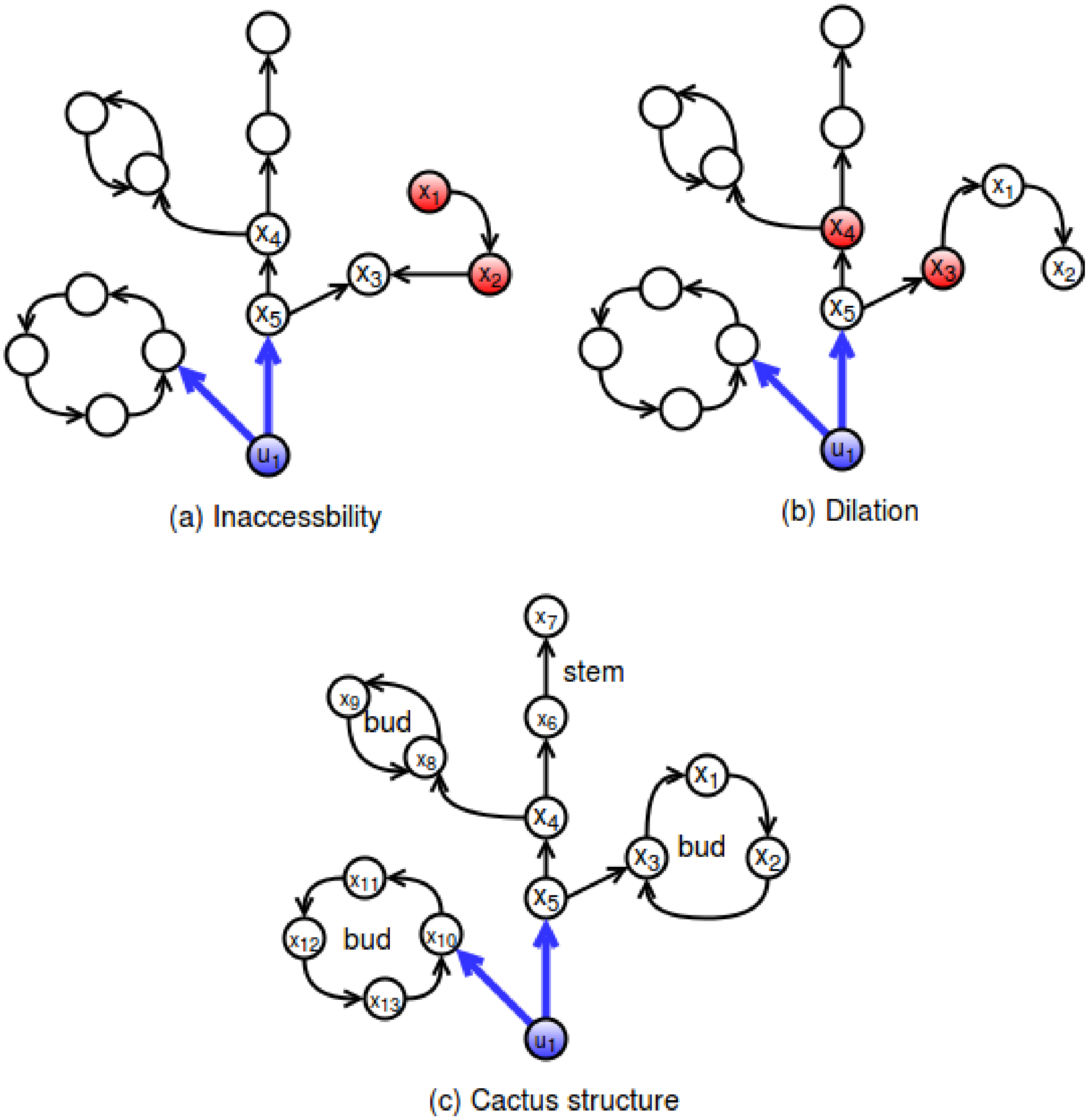
$$rank C = N \tag{2.3}$$

Kalman's criterion for controllability can be tested for systems with small dimensionality of controllability matrix. For larger real-world networks, it is harder to check this condition. Hence, for large systems we need to determine the controllability without calculating the rank of the controllability matrix. Further, the link weights **A** are usually unknown in most networks. We only know if there is a link or not like in biochemical reactions and so it is not possible to verify the Kalman's rank condition using fixed weights [30, 51]. To overcome a few of these limitations Lin provided a graph theoretic framework to check the controllability of a network based on its topology [29]. This is called 'structural controllability'.

2.2 Structural controllability

The graphical interpretation of structural controllability was formulated by Lin [29]. Consider an LTI system $(\mathbf{A}, \mathbf{B})$ represented by a digraph $G(\mathbf{A}, \mathbf{B}) = (V, E)$ where, the vertex set $V = V_A \cup V_B$ includes both the state vertices $V_A = \{x_1, x_2, \ldots, x_N\} \equiv \{v_1, v_2, \ldots, v_N\}$, corresponding to N nodes in the network and the input vertices $V_B = \{u_1, u_2, \ldots, u_M\} \equiv \{v_{N+1}, v_{N+2}, \ldots, v_{N+M}\}$ corresponding to M input signals called origins or roots of the digraph $G(\mathbf{A}, \mathbf{B})$. The edge set $E = E_A \cup E_B$ includes edges among the state vertices $E_A = \{(x_j, x_i) | a_{i,j} \neq 0\}$- the links of network $\mathbf{A}$, and the edges connecting input vertices to state vertices $E_B = \{(u_m, x_i) | a_{i,m} \neq 0\}$.

Based on this, Lin's structural controllability theorem states that an LTI system $(\mathbf{A}, \mathbf{B})$ is structurally controllable if and only if the digraph $G(\mathbf{A}, \mathbf{B})$ has no inaccessible nodes or dilations. A vertex is inaccessible if there are no directed paths reaching that node from input vertices (Figure 2.1(a)). This node cannot be influenced by input signals applied to driver nodes making the network uncontrollable. A digraph $G(\mathbf{A}, \mathbf{B})$ contains a dilation if there is a subset of nodes $S \subseteq V_A$ such that the neighbourhood set of S denoted as $T(S)$, has fewer nodes than S itself. In other words there are more subordinates than superiors (Figure 2.1(b)). If such a structure is present then it is uncontrollable. An alternative graph theoretical formulation of Lin's structural controllability theorem is the presence of cacti structure. A cactus is a structure having (a) stem: an elementary path originating from an input vertex and (b) bud: an elementary cycle with an additional edge that ends but does not begin at a vertex of the cycle. The cactus is a minimal structure that contains neither inaccessible nodes nor dilations (Figure 2.1(c)). Thus the Lin's structural controllability theorem is formulated as follows; An LTI system $(\mathbf{A}, \mathbf{B})$ is structurally controllable if and only if $G(\mathbf{A}, \mathbf{B})$ is spanned by cacti. These conditions can be checked by looking at the topology of the network thus bypassing the numerical issues involved in Kalman's controllability rank test and the difficulties posed by our incomplete knowledge of the edge weights of $G(\mathbf{A}, \mathbf{B})$. This method has helped formulate efficient methods to



(a) Inaccessbility

(b) Dilation

(c) Cactus structure

Figure 2.1: Inaccessibility, dilations and cactus.(a) The red nodes x_1, x_2 are inaccessible from the input node u_1 in blue. (b) The nodes shaded red in the set $S = x_3, x_4$ cause dilation as their neighbourhood set $T(S) = x_5$ contains only one node that aims to control two nodes in S (c) A cactus contains no inaccessible nodes nor dilations. There is one stem $(u_1, x_5, x_4, x_6, x_7)$ and 3 buds $(u_1, x_5, x_3, x_1, x_2, x_3)$. From Liu *et al.*, 2011.

identify a minimum set of inputs that guarantee structural controllability.

In order to control a networked system, we need to identify a set of driver nodes that, when provided with external inputs can offer full control over the network. One can control all the nodes in the network and attain full control, but this is costly and impractical. Hence we are interested in identifying minimum a number of driver nodes, whose control is sufficient to obtain full control of the system. Two well established algorithms to identify minimum inputs are based on maximum matching set [30] and minimum dominating set [31].

2.3 Maximum matching model to identify minimum driver nodes

Kalman's controllability condition cannot identify minimum number of driver nodes, it tells us if we can control a system through a given set of potential driver nodes which we need to guess. Further, one must know all the entries in **A** and **B** which are often unknown in most real systems. Even if this information is available then a brute-force search for the minimum dominating set requires us to compute the rank of almost 2^N distinct controllability matrices, a combinatorially prohibitive task for any network of reasonable size. To overcome this, many studies have mapped the control problem to a purely graph-theoretical problem called maximum matching [30, 52, 53].

Matching is a well studied problem in graph theory, with ample applications [54]. For an undirected graph, matching is a set of edges without common vertices. A maximum matching is a matching of the largest size. The end vertices of a matching edge are called matched and the rest unmatched. Most graphs can have multiple maximum matching (Figure 2.2 (b)). If all vertices are matched, then it is called a perfect matching (Figure 2.2(d)).

In structural control theory, matching was well studied and originally defined in the bipartite representation of a digraph [52, 53]. The extended definition of matching on a digraph naturally connects to the cactus structure. Directed paths and cycles are part of a cactus and hence matchings in digraphs connect to cactus structures. A matching in a directed graph is defined as set of directed edges that do not share common start or end vertices [30]. A vertex is matched if it is the end vertex of a matching edge, otherwise it is unmatched (Figure 2.3(a)). A matching of the largest size is called a maximum matching. If all the vertices are matched then it is called perfect matching (Figure 2.3(f)).

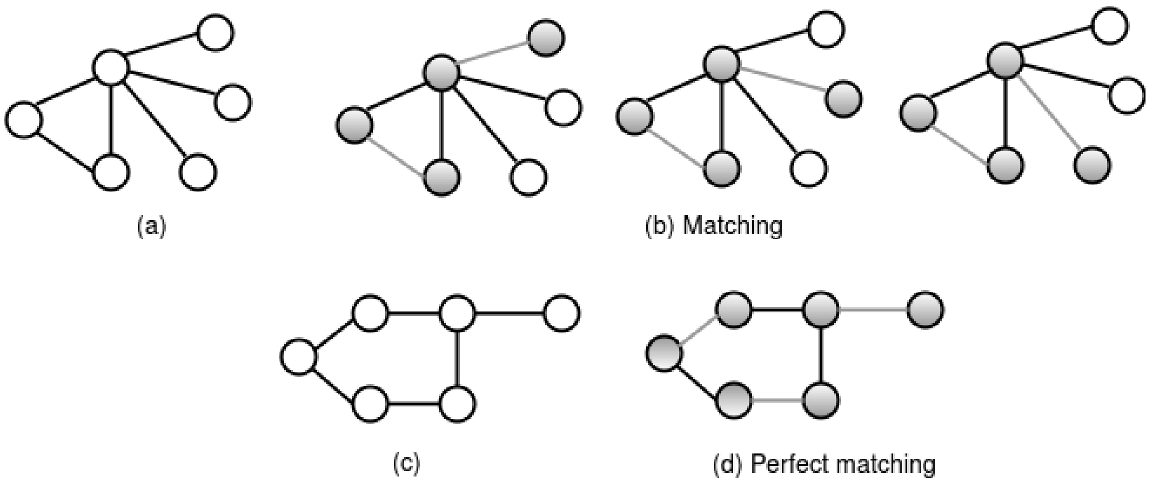


Figure 2.2: Matching in undirected graphs.(a) representation of graph and (b) its matching. (c) and (d) Representation of perfect matching. Edges in matching and matched vertices are coloured in grey.

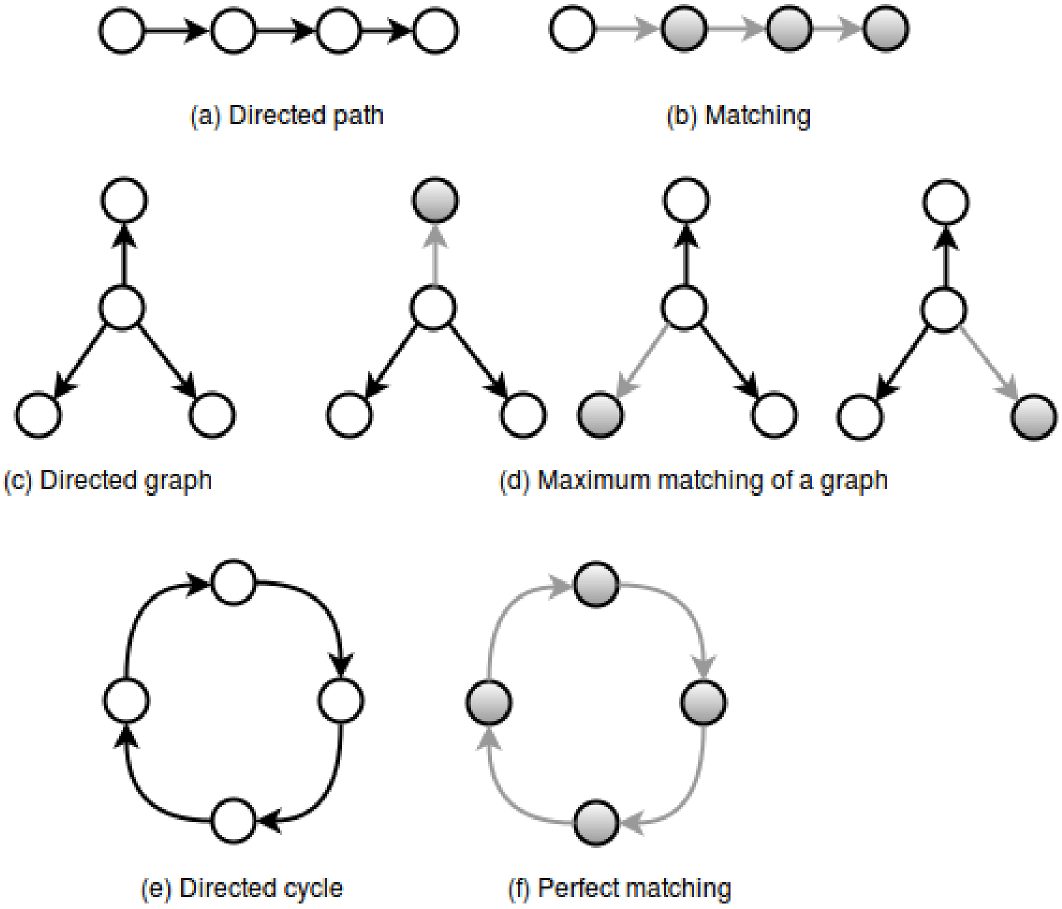
Liu *et.al.* showed that a matching of a digraph can be decomposed into set of directed paths and /or directed cycles that form the basic elements of the cactus structure. They proved a theorem that provides the minimum number of driver nodes or inputs called *Minimum inputs theorem*. To fully control a directed network $G(\mathbf{A})$, the minimum number of inputs, or equivalently the minimum number of driver nodes is,

$$N_D = max\{N - |M^*, 1|\}, \tag{2.4}$$

where, N_D is the size of a minimum set of driver nodes or inputs and $|M^*|$ is the size of the maximum matching in $G(\mathbf{A})$. In other words the driver nodes corresponds to the unmatched nodes. If all nodes are matched i.e. a perfect matching ($|M^*| = N$), then we need at least one input to control the network hence $N_D = 1$. In this case we can choose any node as a driver node.

Algorithmic solution to identify minimum driver nodes

The maximum matching of a graph can be computed in polynomial time. The maximum matching of a digraph can be identified by mapping the graph to its bipartite representation. A digraph $G(\mathbf{A})$ bipartite representation is given as $H(\mathbf{A}) \equiv (V_A^+ \cup V_A^-, \Gamma)$. We create two disjoint sets of *in* (V^-) and *out* (V^+). Here $V_A^+ = \{x_1^+, \ldots, x_N^+\}$ and $V_A^- = \{x_1^-, \ldots, x_N^-\}$ are the sets of vertices corresponding to the


(a) Directed path

(b) Matching

(c) Directed graph

(d) Maximum matching of a graph

(e) Directed cycle

(f) Perfect matching

Figure 2.3: Matching in directed graphs. (a)-(d) Representation of graph and its matching. (f) is a perfect matching in a directed cycle. Edges in matching and matched vertices are coloured in grey.

N columns and rows of the state matrix $\mathbf{A}$. The edge set $\Gamma = \{(x_j^+, x_i^-) | a_{ij} \neq 0\}$. Each node x_i is split into two x_i^+ and x_i^-. A directed link from node j to i corresponds to a connection between node j in the *out* set and node i in the *in* set (Figure 2.4). Self loops are allowed. A maximum matching of the bipartite graph can then be efficiently found using the Hopcroft-Karp algorithm [55]. The unmatched nodes in the *in* set are the driver nodes. This can then be mapped back to the nodes of the original digraph. As there could be multiple maximum matchings for a digraph with the same size N_D, nodes are further classified as critical, ordinary and redundant based on their presence in driver node set [36]. A node is critical if it is never matched, ordinary if occasionally matched and redundant/non-driver if always matched in the *in* set.

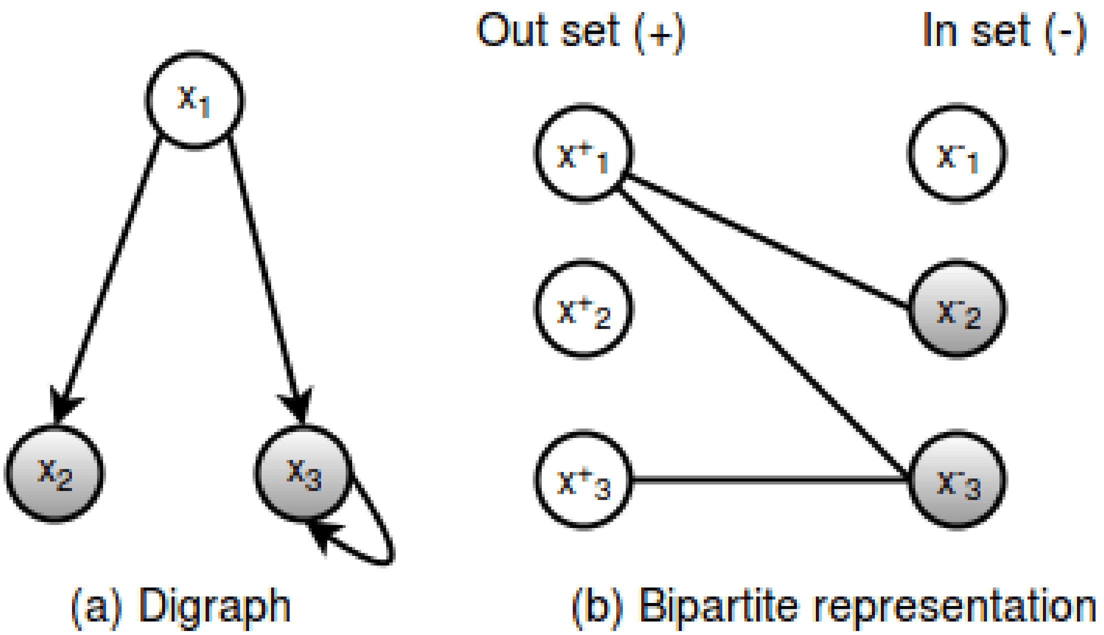


Figure 2.4: Maximum matching calculation. Nodes in grey are matched nodes.

Algorithmic solution to identify driver node class

1. **Identification of critical driver nodes**: A node is critical if it is never matched. In other words, always present in all driver node sets. A node is critical if and only if it has no incoming link [36]. A node n in the *in* set can never be matched if and only there is no link to n.

2. **Identification of redundant and ordinary driver nodes** : Redundant/non-driver nodes are always matched in the bipartite graph. If this is forcibly unmatched then there would be no alternative matching and the number of unmatched nodes will decrease. Obtain a set of matched nodes M in the *in* by finding the maximum matching of bipartite graph using Hopcroft-Karp algorithm. Pick an element i from M. Identify the node j from the *out* set that matched node i. While keeping the current matching, delete the node i and all its links. Check for an augmenting path that starts with node j and ends at an unmatched node and alternates between unmatched and matched links on a path. If there is no augmenting path found, node i needs to be always matched and therefore is redundant. If there is an augmenting path, node i is replaceable and hence is ordinary.

2.4 General properties of driver nodes identified by maximum matching

Studies on various networks [30] and our analysis have shown the following properties of driver nodes (Table 2.1).

1. The fraction of driver nodes is higher among low degree nodes than among hubs. Hub nodes are rarely driver nodes.

2. The size of the driver node set N_D is mainly determined by degree distribution. When the fraction of driver nodes n_D^{real} is compared with $n_D^{rand\text{-}Degree}$ and $n_D^{rand\text{-}ER}$, then $n_D^{rand\text{-}Degree}$ is close to n_D^{real}. Here $n_D^{rand\text{-}Degree}$ is degree preserved randomization, which keeps in-degree and out-degree of each node unchanged but randomly selects the nodes that link each other. $n_D^{rand\text{-}ER}$ is complete randomization and turns the network into a directed Erdos-Renyi random network with N and L unchanged.

3. Sparse and heterogeneous networks are most difficult to control.

Network	**Nodes**	**Edges**	n_D^{real}	$n_D^{rand\text{-}Degree}$	$n_D^{rand\text{-}ER}$	$n_D^{rand\text{-}Joint}$
Cancer Signalling	1232	3060	0.4574	0.4323	0.1387	0.4323
Directed Human PPI	6339	34813	0.3601	0.2654	0.0048	0.2659
HIV-human molecular	6361	40625	0.356	0.2469	0.0019	0.2466
T-cell activation	121	255	0.2892	0.2641	0.194	0.2642
HIV-T-cell activation	137	367	0.2481	0.241	0.1112	0.2361
{E. coli} transcription	424	578	0.7287	0.7295	0.3526	0.73

Table 2.1: Controllability properties of biological networks analysed in this thesis. Here $n_D^{rand\text{-}Joint}$ is joint degree preserved randomization.

2.5 Classifying nodes based on their effect on driver node set size

Another way to assess a node's importance for control is to quantify the impact of its removal on controllability [43, 45, 51]. Consider a network with minimum number of driver nodes N_D. After a node is deleted or removed, denote the minimum number of driver nodes for this network as N'_D. We can classify a node into any one of the following categories.

1. Indispensable: A node is indispensable if upon its deletion, we have to control more number of driver nodes, i.e, $N'_D > N_D$. Example removing a node in middle of a path will increase N_D

2. Dispensable: A node is dispensable if upon its removal, we have to control less number of driver nodes, i.e, $N'_D < N_D$. For example, removing a leaf node in a star will decrease N_D by 1.

3. Neutral: A node is neutral if its removal has no effect on N_D, i.e, $N'_D = N_D$. For instance, removal of central hub node in a star will not change N_D.

For a given network N, we delete a node at a time and all its links to create the new network N'. Compute the minimum number of driver nodes N_D using the maximum matching algorithm for each of these networks and compare N_D. Repeat this until all nodes in the network are deleted.

2.6 Graph theoretic characterisation of driver nodes

We partially characterised driver nodes based on the topology of the network. Given a directed network N, we try to characterize the nodes as critical driver nodes (CDN), ordinary driver nodes (ODN) or redundant/non-driver nodes (NDN) based on the degrees of the node and its neighbours.

The case when every node has $indegree = outdegree = 1$ will be the case of the network being a directed cycle, in which case the nodes will all be ODNs. We therefore focus on directed networks which are not cycles. In such networks, those nodes which have in-degree equal to zero will be nodes without any parent and will remain unsaturated in any matching. So such nodes will be CDNs. Amongst those nodes with in-degree larger than zero, consider those which have atleast one parent whose out-degree is one. Such nodes will be saturated by every maximum matching and will consequently be NDNs. This is so because, if the node is unsaturated by some maximum matching say M then by inducing the arc/edge that connects the node with its parent we obtain a matching containing exactly one more number of edges than those in M; a contradiction to M being maximum.

The case that now remains uncharacterised is the case of a node with in-degree larger than one and no parent with out-degree equal to one. Such nodes cannot be CDNs. However, at present, we are not able to identify the cases under which these will be ODNs or NDNs.

2.7 Conclusion

In this chapter we have described the maximum matching method for identifying minimum driver node sets in directed graphs which has been implemented in our study and partially characterised the driver nodes based on its topology. An alternative method is the minimum dominating set approach for undirected networks which can be extended for directed networks as well [31]. For a graph $G(V, E)$, where V is set of vertices and E is set of edges, a subset $S \subseteq V$ is called dominating set (DS) if every node in V is either an element of S or is adjacent to an element of S. That is for any node $v \in V$, $v \in S$ holds or there is a node $u \in S$ such that $(u, v) \in E$ then we say that v is dominated by u. Then S is dominating set if each node in V is either in S or dominated by some node in S. A minimum dominating set (MDS) is a dominating set with minimum number of vertices. The MDS forms the driver node set [31, 56]. We have implemented this method on the HIV-1 human molecular interactome (Appendix 2). While each of these methods identify minimum driver nodes, the choice of method is determined by the kind of control one wants and the type of network. In some cases, maximum matching yields fewer driver nodes than the minimum dominating set. For instance in a directed path, there is only one driver node by the matching method, but the MDS would give more driver nodes. Thus the kind of network would determine the type of method to be implemented.

Chapter 3

Identification of Critical Regulatory Genes in Cancer Signaling Network

3.1 Introduction and motivation

Cancer is a complex disease which is characterized by subtle interplay of regulatory mechanisms underlying its phenotype [57, 58]. Dysregulation of multiple pathways governing fundamental cell processes (such as death, proliferation, migration and differentiation) is known to be a key cause for emergence of cancer. The crosstalk between signalling pathways reflecting salient aspects of disease have been used to model this pathology [22, 59]. The focus behind building such integrative models has been to create a meaningful molecular picture of cancer so as to find ways for controlling the disease [60, 61, 62].

Systems that exhibit complex phenomena owing to interconnected mechanisms underlying its architecture can be studied using graph theoretical paradigm [15, 63, 64]. Study of such networked systems of social, technological and biological origin has added to the understanding of their structure, function and evolution. Availability of rich data mapping biological processes in the postgenomic era has facilitated creation of molecular interaction catalogues (protein interactomes, gene regulatory networks, metabolic pathways and co-expression networks), and better understanding of cellular functions [65, 66, 67]. Using these graph theoretical approaches, various studies have attempted to identify regulatory mechanisms which are central to the disease [1, 68, 69, 70]. Rapid advances in network biology have provided a new conceptual framework revolutionizing the view of biology and disease pathology.

From control systems perspective, cellular processes can be viewed as an intricately controlled orchestra of regulatory mechanisms that lend the cell its functional repertoire. Diseases, therefore, can be seen as the result of errors in cellular information processing. Beyond systems modeling of diseases, the focus has also been on finding ways of controlling the disease through therapeutic interventions. Recent excitement in controllability of networks and identification of driver nodes as agents for steering the state of the network has added much needed impetus in this direction [29, 30, 31, 36, 71]. Control systems theory proposes that the state of complex networks can be controlled with the help of a set of 'driver nodes'. Driver nodes may then be fed with external inputs to steer the state of the network.

Analysis of systems biological models of diseases has provided crucial insights into their mechanisms and potential drug targets [22, 72]. The understanding garnered through such studies has often taken a static perspective of disease interactomes, ignoring dynamical aspects. Study of controllability of diseases and search for driver nodes as potential therapeutic targets provides a new dimension. Such integration of control theory with disease interactomes could pave way for better strategies to assist drug discovery process as well as improve our understanding of the disease [43, 44, 73].

Cancer systems biology has proved to be helpful in implementing new therapeutic strategies [60, 74]. While most studies [72, 75, 76] focus on identification of epigenetic changes and mutations in genes, and target them as means of controlling the disease, control theory offers a framework for arriving at driver genes that could be used for steering the state of the regulatory network. We propose that identification of specific molecules as drivers of regulatory dynamics could be a promising step towards targeted cancer therapies.

Here, we modeled the human cancer signaling network as a directed graph and probed for its critical regulators using structural controllability. We implemented the maximum matching algorithm to identify driver nodes. The driver nodes were divided into backbone, peripheral and ordinary based on their role in regulatory interactions and control of the network. Based on node deletion studies that enumerate impact of a node on ease of network control, indispensable nodes were identified. These indispensable backbone driver nodes were found to be critical for driving the regulatory network into cancer phenotype (via mutations) as well as for steering into healthy phenotype (as drug targets). Thus they emerged as central to control, both as causal elements by virtue of mutations and also as therapeutic agents in the form of cancer drug targets. This study illustrates an application of control theory for investigation of regulatory mechanisms underlying complex diseases.

3.2 Data procurement and network analysis

3.2.1 Human Cancer Signaling Network

We created the Human Cancer Signaling Network (HCSN) starting from the data of cancer signaling network [77] comprising of 1634 genes and 5089 regulatory relations by integrating genetically and epigenetically altered cancer associated genes and signaling pathways for cancer. Thus HCSN embeds molecular correlates associated to cancer. The nodes in HCSN are signaling molecules (such as genes, proteins and other small molecules) and links represent effector actions such as activation or inhibition and protein-protein interactions. As a first step towards construction of HCSN the data was purged to remove all undirected protein-protein interactions, thereby creating a weighted network comprising of 1240 nodes and 3144 directed edges. Consistent with strategies used earlier [44, 43], the directed graph was obtained for performing controllability analysis. Further, multiple directed edges were merged leaving behind 3065 edges in the unweighted graph. In the next step, the nodes data were curated for ambiguities in gene names using HUGO symbol and Entrez ID as gene identifiers. The final HCSN thus constructed had 1232 nodes and 3060 edges (Figure 3.1).

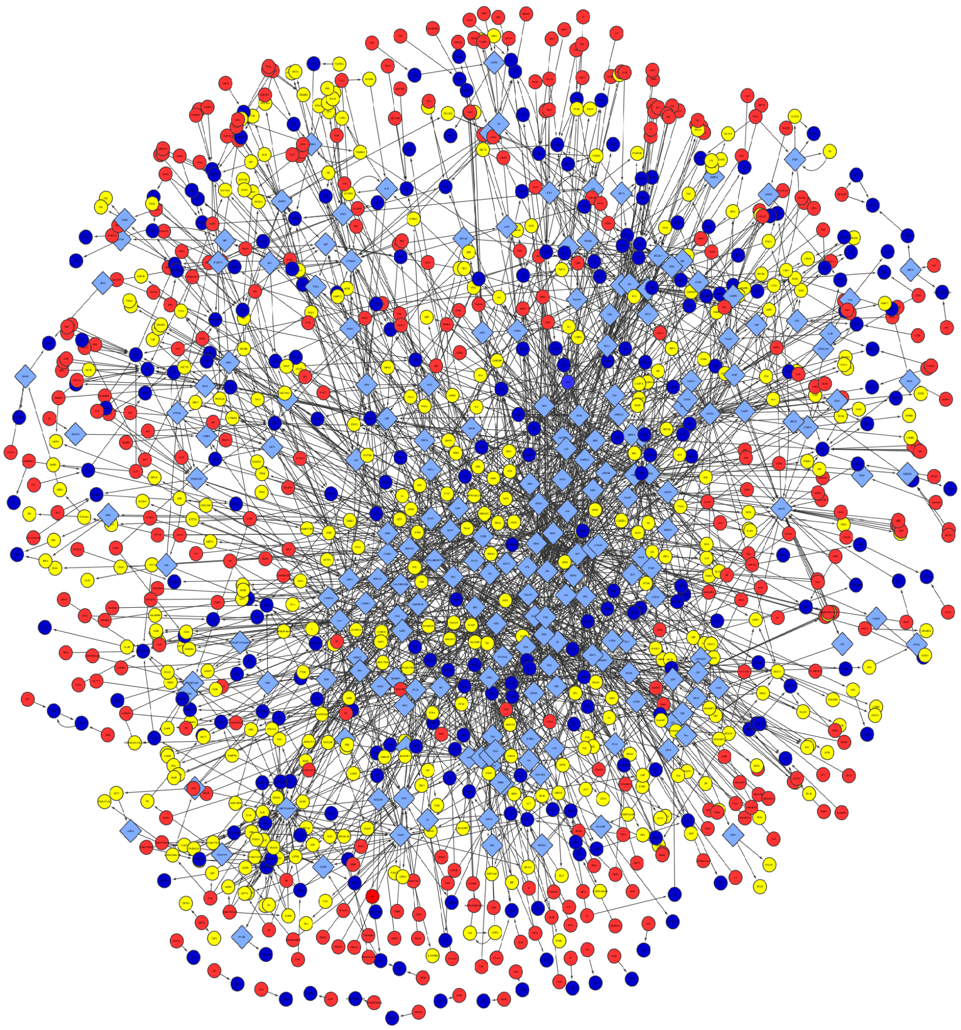


Figure 3.1: Human cancer signaling network (HCSN). Using controllability analysis, driver nodes were identified and characterized based on topology. Driver nodes were classified as peripheral (PDN, shown in red), ordinary (ODN, shown yellow) and backbone (BDN, shown in blue). The indispensable BDNs, that emerged as central to regulatory mechanisms of cancer, are depicted as light blue diamond shaped nodes. PDN: Peripheral driver node, ODN: Ordinary driver node and BDN: Backbone driver node.

3.2.2 Identification and classification of driver nodes

Driver nodes provide means for controlling the state of the network [30]. These nodes when given a certain input can drive the system from any initial state to any desired final state in finite time. The maximum matching algorithm for digraph was used to identify driver nodes in the form of the minimum driver node set (MDNS). MDNS identification was implemented using Controllability Analysis package by Liu.*et.al* [30]. The details of method is explained in chapter 2.

As there could be multiple maximum matchings for a digraph, multiple MDNS exist with the same size N_D. Each node was categorized based on their role as driver node. Based on presence or absence of driver nodes in MDNS, these were classified into three categories: Peripheral Driver Nodes (PDN), Backbone Driver Nodes (BDN) and Ordinary Driver Nodes (ODN). PDNs appeared consistently in all MDNS and had predominantly low degrees thus were primarily located at the periphery in the network. BDNs on the other hand did not appear in any of the MDNS. They were predominantly presented with high degrees and were concentrated in the core of the network. ODNs fell under neither of these categories, inconsistently appeared in MDNS and had no distinct degree characteristics.

Thus every molecule in HCSN was classified into either of the three driver node categories. Figure 2 depicts the segregation of PDNs, ODNs and BDNs based on degree centralities (degree (k): total links incident on a node; in-degree (k_{in}): number of incoming links on a node; and out-degree (k_{out}): number of out going links from a node). HCSN nodes were divided into four classes (low, medium-low, medium-high and high) of approximately equal size such that nodes with equal centrality value were binned into the same class (Table 3.1).

	Low	**Medium-low**	**Medium-high**	**High**
k	1	2	$3-5$	$6-60$
k_{in}	0	1	$2-4$	$5-32$
k_{out}	0	1	$2-3$	$4-49$

Table 3.1: The classification of the nodes into bins based on their degree.

3.3 Biological characterization of HCSN nodes

Molecules in HCSN were characterized for their cellular localisation, gene ontology and biological essentiality. While cellular localisation study was performed for all the molecules (genes, proteins and other small molecules), the remaining two characterization studies were done for only genes. For the latter analyses, non-gene entities were removed from HCSN leaving behind 1178 genes. Among these 342 were PDNs, 375 were ODNs and 461 were BDNs.

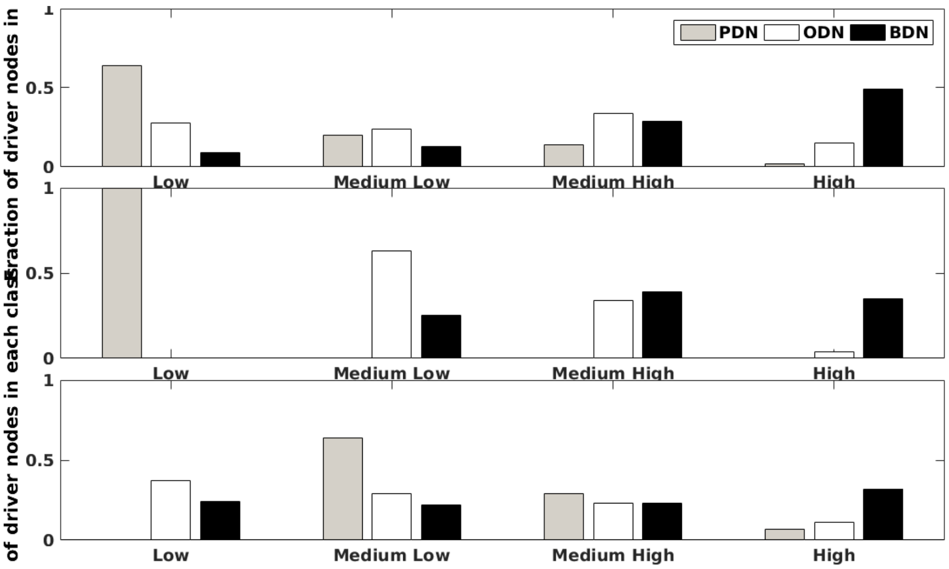


Figure 3.2: Degree characterization of driver nodes. Height of bars corresponds to fraction of driver nodes in each class of degree.

Cellular localization Molecules in HCSN were categorized into five different classes depending upon their participation in different stages of the signalling pathway [77]: Ex, extracellular molecules that are ligands; Cm, cell surface receptors located in cell membrane; Cy, intracellular molecules found in the cytoplasm and cell organelles; Nu, molecules found in nuclear membrane as well as nucleus; and NA, molecules whose cellular localization are indeterminate. Certain molecules had multiple cellular localizations as they were reported to be involved in more than one stage of signalling. For each class of cellular localisation, representation of driver nodes across all three categories (PDNs, ODNs and BDNs) was computed. The significance of any over- and under-representation was evaluated using chi-square statistics (See Appendix 1). The statistics were computed for significance level of 0.05 and 0.01 with degree of freedom 2 (Table 3.2). The cellular characterization of driver nodes reflects its importance or role in signalling pathway. It was observed that the PDNs were significantly present among extracellular proteins which follows from the fact that PDNs are mostly ligands. BDNs, on the other hand, were found predominantly among the intracellular signalling proteins. The results were statistically significant at p-value 0.01 (Figure 3.3). ODNs were found to have no specific characteristic cellular localization.

	PDN	ODN	BDN	χ^2(0.05)5.99/0.01(9.21)
HCSN (1232)	29%	31.21%	39.44%	
Ex (88)	53.41	15.91	30.68	25.53
Cm (237)	23.21	39.66	37.13	9.17
Cy (768)	26.04	28.26	45.70	11.95
Nu (252)	25.00	32.14	42.86	2.43
NA (32)	50.00	25.00	25.00	6.73

Table 3.2: Chi-square analysis of cellular localisation of driver nodes. Ex: Extracellular, Cm: Cell membrane, Cy: Cytoplasm and organelles, Nu: Nucleus and NA-Indeterminate.

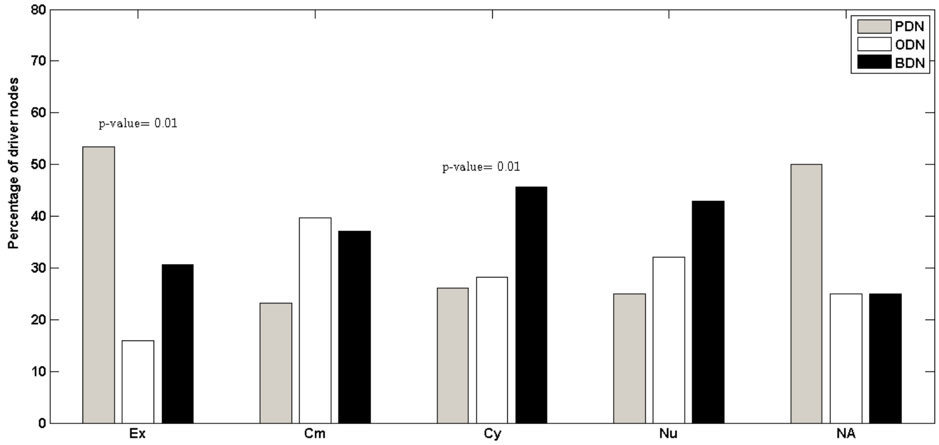


Figure 3.3: Percentage of driver nodes across the different localisation. Ex: Extracellular, Cm: Cell membrane, Cy: Cytoplasm and organelles, Nu: Nucleus and NA-Indeterminate.

Gene Ontology enrichment analysis The gene ontology (GO) enrichment analysis facilitates identification of biological processes, molecular functions and cellular categories that are significantly overrepresented in a set of target genes relative to a background. We performed GO enrichment analysis for each class of driver nodes (target), in the background of HCSN genes, in order to identify characteristic biological attributes among the PDNs, ODNs and BDNs, using WebGestalt [78].

The set of BDNs were presented with more significantly enriched GO terms than the PDNs and ODNs pointing out their distinct ontological character. PDNs comprised of genes with diverse ontologies and thus did not yield any significant enrichment in biological processes or molecular functions. Enrichment in extracellular localization further established characterization of PDNs. ODNs showed weak molecular enrichment for terms such as Calmodulin binding, actin binding, neurotransmitter receptor activity and G-protein coupled receptor activity. This indicates that ODNs function mainly as intermediate messenger proteins. On the other hand, BDNs showed significant enrichment for all the three ontological categories. BDNs participate in apoptosis related processes such as cell death, regulation of cell death, regulation of apoptosis and response to stress, chemical response and cellular responses. In terms of molecular functions, BDNs participate in enzyme binding activity and were localized within the cytoplasm and organelles.

Essentiality Essential genes are the genes that are critical for survival of the organism. Using the DEG (Database of Essential Genes) [79], which contains a list of essential genes for both prokaryotes and eukaryotes, HSCN genes were divided into essential and non-essential. Driver nodes were further characterized for biological essentiality. We observed that the BDNs were significantly overrepresented by essential genes ($Z - score = 4.34$) while they did not have significant representation in PDNs ($Z - score = -2.15$) and ODNs ($Z - score = -2.49$). This indicates to the biological relevance of BDNs and highlights their indispensable nature.

3.4 Controllability analysis of HCSN

After characterizing driver nodes for their topological and biologically relevant features, we investigated for possible means by which control of HCSN could be achieved. We surmise that the network could be controlled from two contrasting perspectives: (1) to drive the molecular regulatory network into cancer state, or (2) starting from cancerous state, to drive the network into a healthy state. Here we take the view that, by virtue of its dynamical state, this network of cancer-associated regulatory mechanisms could broadly be in either in the state of 'cancer phenotype' or 'healthy phenotype'. Cancer associated gene mutations are a reflection on the functionality of these genes in the absence of which the network

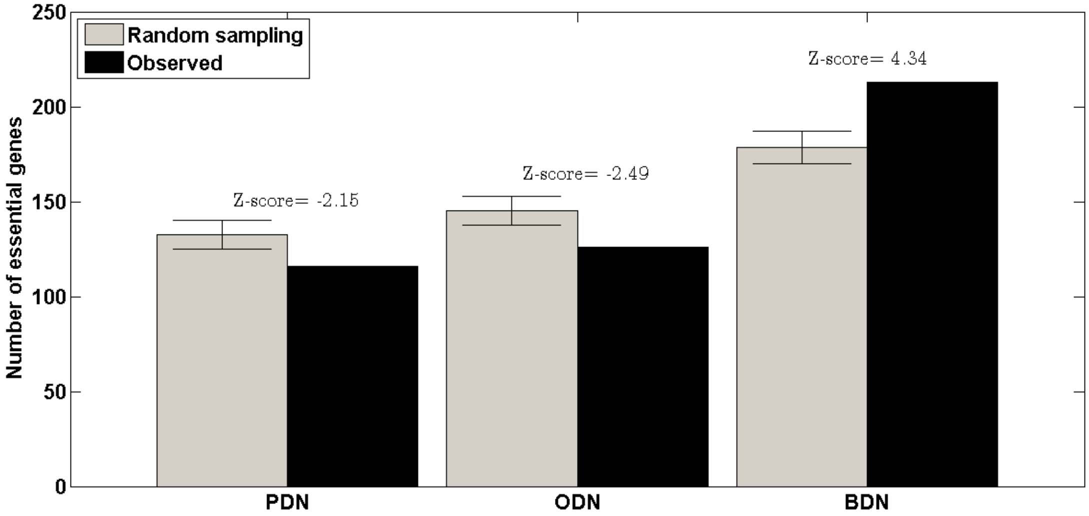


Figure 3.4: Representation of essential genes among driver nodes. BDNs were significantly over-represented by essential genes.

is driven into cancer state. On the other hand, gene targets are drivers utilized to steer the cancer into healthy phenotype via drug interventions. We investigated the association of driver nodes with empirically reported cancer genes as well as that with genes that are known to be specifically targeted by anticancer (antineoplastic) drugs.

Driving into cancer state: Association of driver nodes with cancer genes HCSN genes were grouped into two classes based on their association with cancer. The cancer association was adjudged on the basis of cancer causing mutations as well as cancer genes ascertained from literature and primary datasets [77]. The cancer-associated gene list contained 2128 genes in total out of which 440 genes were part of the HCSN. BDNs were significantly overrepresented (47% BDNs as compared to 27% of PDNs and 26% of ODNs; $p = 0.01$) among the cancer-associated genes (Figure 3.5). Among the remaining 738 genes not linked to cancer, all three types of driver nodes were almost equally represented (223 PDNs, 260 ODNs and 255 BDNs; $p = 0.01$). This highlights the key role played by BDNs in the regulatory mechanisms, in the absence of which the network is steered into cancerous phenotype.

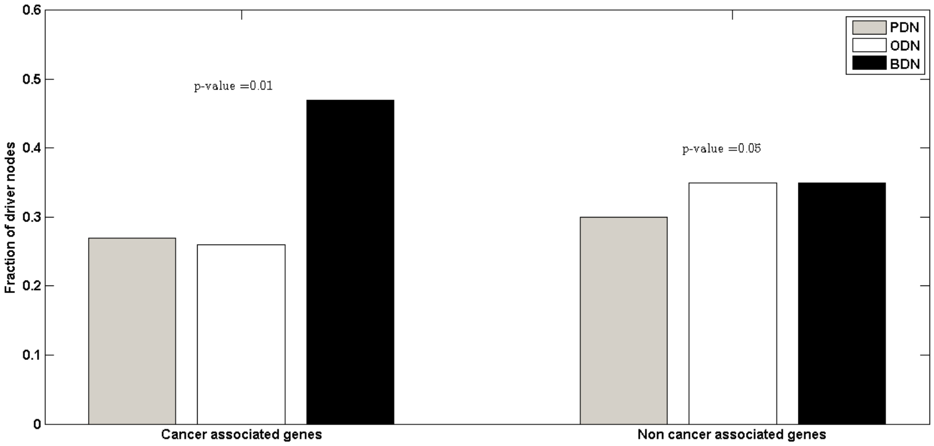


Figure 3.5: Driver node representation among cancer associated and non-cancer associated genes of HCSN. BDNs were significantly enriched with cancer associated genes.

From Cancer state to healthy state: Drug association studies We considered antineoplastic drugs as molecular agents that drive the cancer phenotype into healthy state by controlling/tweaking the mechanisms of target genes. Drug-target associations linking antineoplastic drugs with HCSN genes were compiled with the help of DGIdb (The drug gene interaction database) [80]. To ignore spurious drug-target interactions, only highly curated associations were obtained by using ‘source trust curation level’ as ‘expert curated’. We obtained 298 ANPDs that target 156 genes in HCSN via 893 drug-targets interactions. It was observed that the nature of distribution of both ‘number of targets a particular ANPD regulates’ as well as ‘number of ANPDs with which a HCSN gene is regulated by’ were heterogeneous, indicating dominance of few promiscuous ANPD regulators as well as few key HCSN targets, respectively (Figure 3.6, in and out degree distributions).

First, we compared number of HCSN genes that are targeted by ANPDs (Figure 3.7a). We observed that, compared to random sampling, while the PDNs ($Z-score = -2.54$) and ODNs ($Z-score = -1.93$) were under-represented among the targets of ANPDs, BDNs were significantly over-represented ($Z-score = 4.18$). This implies that ANPDs tend to avoid PDNs and ODNs as their targets, they preferentially target BDNs. This points at the relevance of BDNs as drivers of

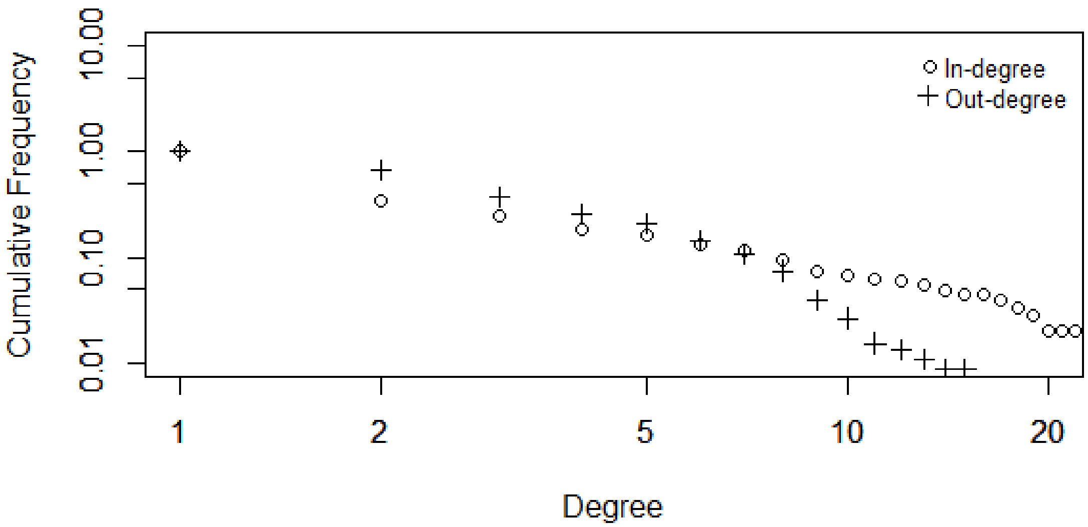


Figure 3.6: Distribution of number of ANPDs a HSCN gene is regulated by (in-degree) and number of targets a particular ANPD regulates (out-degree).

the state of regulatory network and their importance as points of control in steering the cancer into healthy phenotype through ANPDs. Next, we compared the number of ANPDs that target genes in each of the three driver node categories (Figure 3.7b). We found that number of PDN genes that were targeted by ANPDs were comparable to that by chance. ODNs were avoided as their targets by ANPDs ($Z - score = -2.69$). BDNs, on the other hand, were targeted with higher preference ($Z - score = 0.91$). Together, these results highlight the importance of BDNs as means of control of cancer and that they are used as key regulators by ANPDs to drive the state of regulatory network into a healthy phenotype.

3.4.1 Effect of node deletion on controllability of HCSN

Controllability of a network is dictated by its degree heterogeneity as well as its edge density [30]. The denser a network, fewer the number of driver nodes needed for its control. In comparison, sparse and heterogeneous networks have larger number of driver nodes and hence are more difficult to control. Every node in a network thus contributes towards controllability to varying degree signifying its relevance.

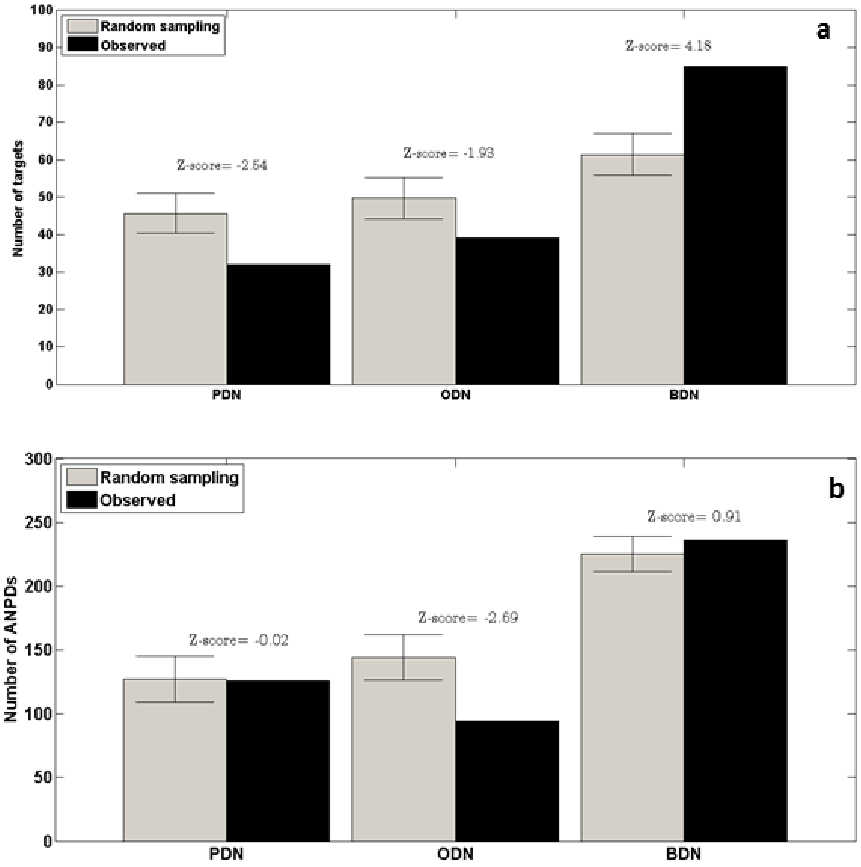


Figure 3.7: Statistics depicting the key role of BDNs in driving the HCSN from cancer state into healthy phenotype as measured in terms of (a) number of genes targeted by ANPDs, and (b) number of ANPDs that act through a gene target in the network.

In order to assess and enumerate the contribution of a node to controllability of the network, we conducted node deletion studies to measure the altered size of MDNS (N'_D) as compared to that of the original network (N_D). Increase in number of driver nodes ($N'_D > N_D$) upon removal of a node implies its importance to control of the network, and such nodes were named 'indispensable'. With the same logic, nodes were named 'dispensable' or 'neutral' if their removal caused decrease ($N'_D < N_D$) in number of driver nodes or no change ($N'_D = N_D$), respectively (See Chapter 2 for method). Table 3.3 depicts statistics of HCSN nodes characterized for their role as driver node (PDNs, ODNs and BDNs) as well as their impact on control (Indispensable, Dispensable and Neutral).

BDNs, that were observed to be critical for control of HCSN (Figure 3.5 and Figure 3.7), comprised of only indispensable and neutral nodes. Interestingly, all

	PDN	BDN	ODN	Total
Indispensable	-	209	-	209 (16.96%)
Dispensable	237	-	286	523 (42.45%)
Neutral	125	279	96	500 (40.58%)
Total	362 (29.38%)	488 (39.61%)	382 (31%)	1232

Table 3.3: Statistics of HCSN nodes characterized for their role as driver node (PDNs, ODNs and BDNs) as well as their contribution to control (Indispensable, Dispensable and Neutral).

indispensable nodes were BDNs. We surmise that the central role of BDNs in control mechanisms of HCSN is plausibly lent by indispensable nodes. We scrutinized this proposition by further dissecting the BDNs to dissociate contribution of indispensable nodes vis-á-vis neutral nodes.

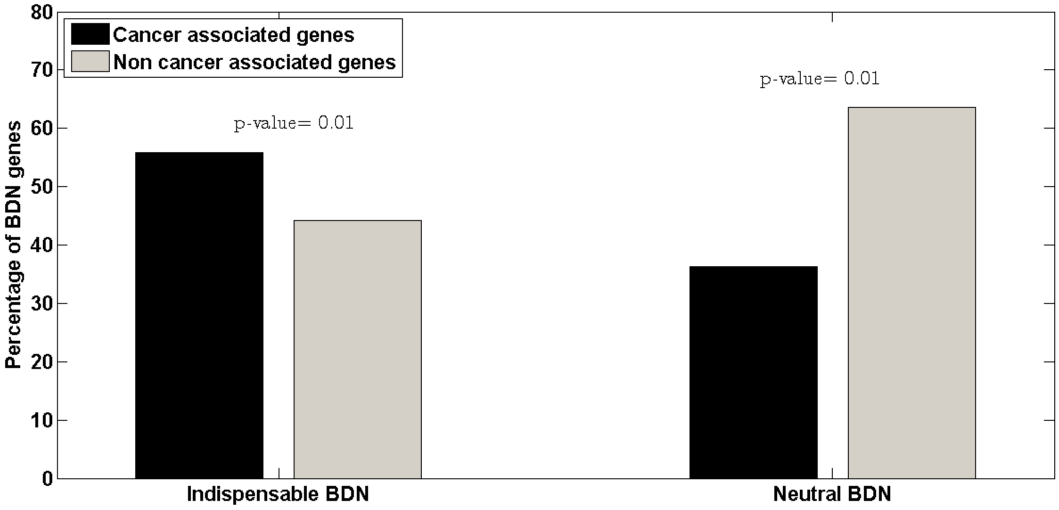


Figure 3.8: Association of cancer genes among Indispensable and neutral BDNS. Indispensable BDNs were associated with cancer than neutral BDNs.

Indispensable BDNs are central to control of cancer We investigated for the possible differential role of indispensable and neutral BDNs for their association with cancer as well as for involvement in drug mechanisms. We observed that mutations in Indispensable BDNs was strongly associated with cancer (Figure 3.8). Further, the ANPDs were divided into two categories based on the type of BDN genes they target to alter the cancer phenotype into healthy one (Figure 3.9). We, also studied the role of BDN gene types in altering the cancer phenotype into healthy one by virtue of action of ANPDs.

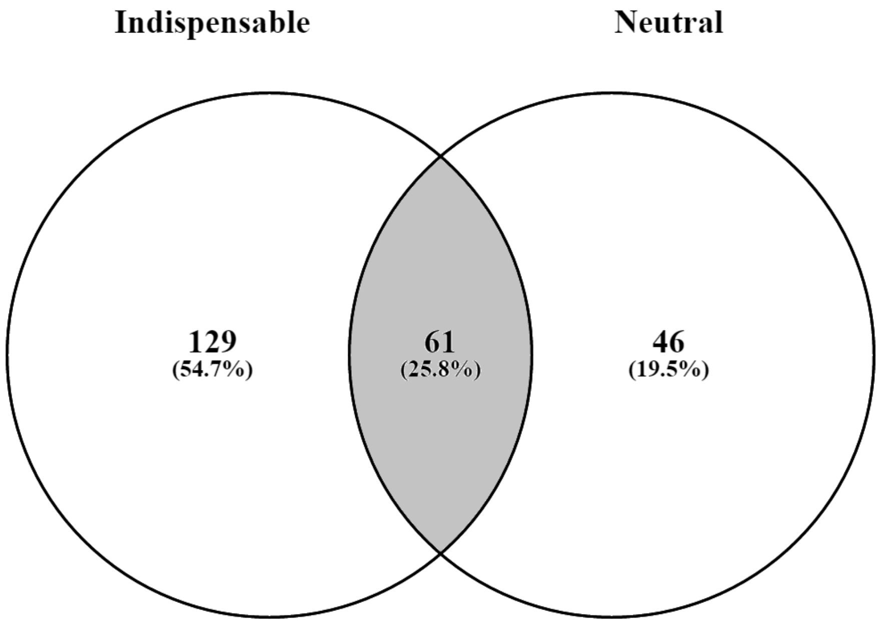


Figure 3.9: Distribution of ANPDs among the indispensable and neutral BDNs.

Among all ANPDs that are known to target BDNs, the mechanism of drug action of around two third of these anticancer drugs were found to be mediated through indispensable BDNs (Figure 3.10). This pattern was consistent and in fact more accentuated among ANPDs that are specific to either class of BDNs. This brings out the central role of indispensable BDNs in control of cancer. They have more unique ANPDs targeting them and also are crucial in maintaining the number of driver nodes in the network. Thus indispensable BDNs quantify to be important in maintaining the state of the system. They can also be viewed as molecules for targeted cancer therapeutics.

3.5 Drug-gene interaction in HCSN based on controllability analysis

We also looked at the interaction of drugs among the nodes classified based on their impact on driver node set size alone. We mapped the concept of controllability of HCSN to empirically tested Antineoplastic drugs that potentially drive

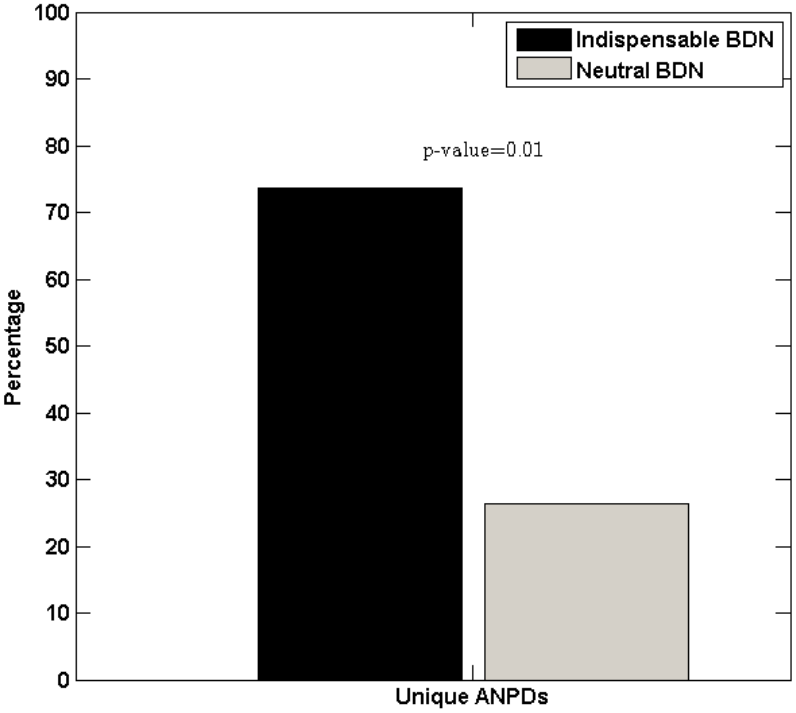


Figure 3.10: Distribution of unique ANPDs among the indispensable and neutral BDNs. Indispensable BDNs had significantly more unique ANPDs acting as cancer regulators through them.

the state of the regulatory network. Using the earlier database DGIdb [80] we complied the drug-target interactions for the genes in HCSN with ANPDs (Figure. 3.11). The data obtained involves 298 ANPDs that target 156 genes in HCSN through 893 drug-target interactions. The ANPDs were segregated further to identify drugs that exclusively target genes in one class. The fraction of genes exclusively targeting indispensable genes was compared to those targeting Dispensable and neutral genes. We focused on ANPDs that exclusively target in one of the driver genes class, ignoring those targeting genes belonging to more than once class. Among these 171 specific ANPDs, we found that 85 of them target indispensable genes (49.7%), as compared 42 (24.6%) that target dispensable genes and 44 (25.7%) that target neutral genes (Figure. 3.12). Thus we observed that indispensable genes are preferentially targeted by ANPDs, indicating their therapeutic relevance for control of cancer. We suggest that anti-cancer drugs potentially act as external signals, primarily through indispensable genes, to drive the

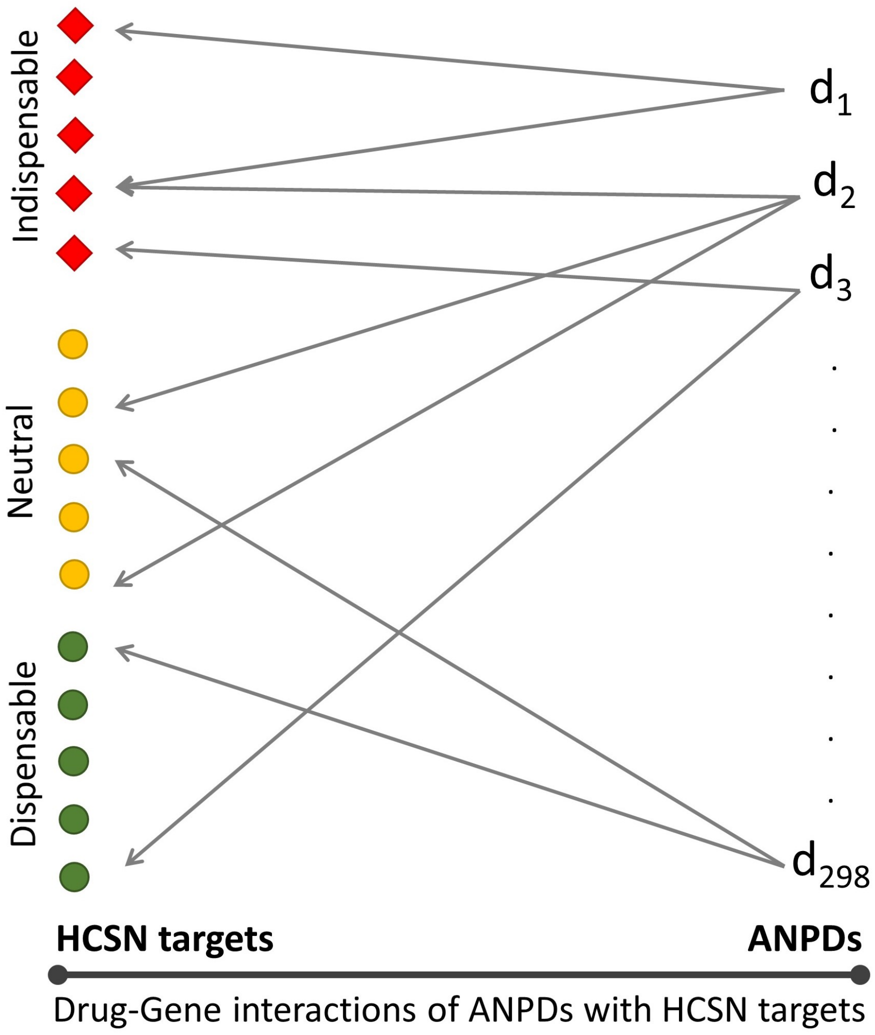


Figure 3.11: Illustration depicting drug-gene interactions between genes of Human Cancer Signalling Network and antineoplastic drugs. HCSN genes were classified into Indispensable, Dispensable and Neutral, based on their contribution to controllability of the network.

state of the underlying signalling mechanism into the healthy phenotype [48]. Application of the engineering concept of structural control in a signalling network to a complex biological signalling disease-associated network, thus reveals empirically reported anti-cancer drugs to be analogous to external input that drives the state of the network out of pathological expression pattern.

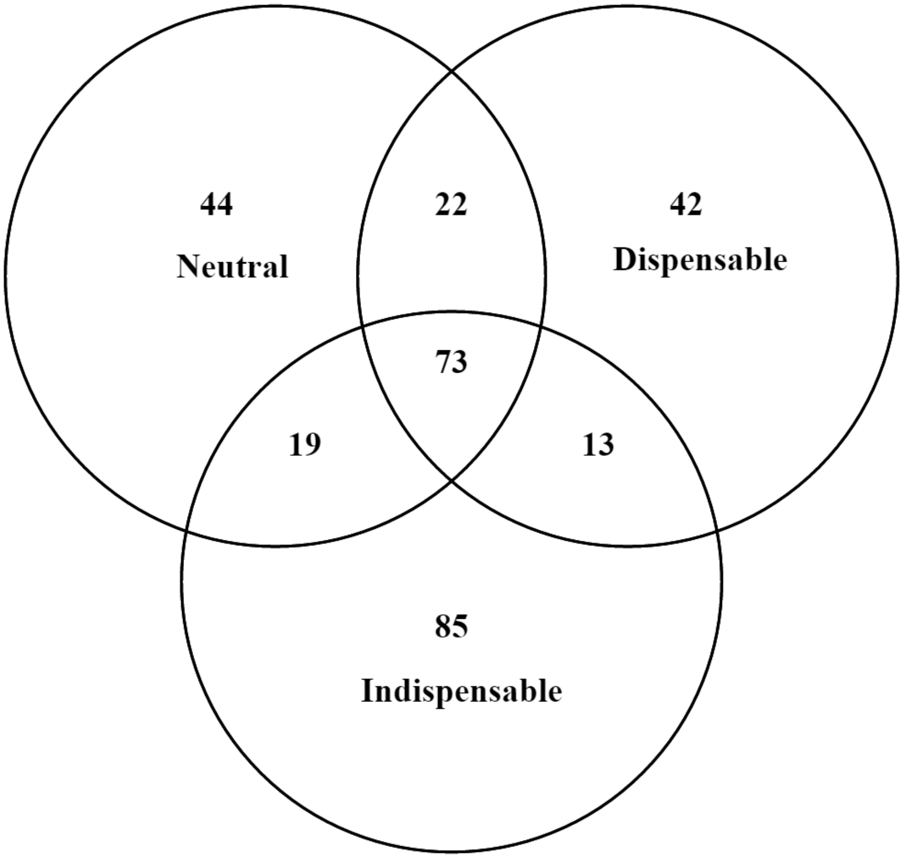


Figure 3.12: Venn diagram representing the number of ANPDs that target different types of driver genes in HCSN. Among the 298 ANPDs, 73 indiscriminately targeted all three types of driver nodes, and 54 of them had ambiguous drug-gene interactions cross-category binding. Among the drugs that exclusively targeted to either of the three classes of nodes, Indispensable genes were prominent targets (85), as compared to Dispensable (42) and Neutral (44) genes

3.6 Discussion and Conclusions

Our study integrates the systems biological approach to cancer regulatory mechanisms with control theory to identify biological implications of 'driver nodes'. We propose the notion of regulatory network as an underlying molecular framework that is subject to control through indispensable backbone driver nodes. These nodes could either steer the network from healthy state to disorder by means of mutations, or could be leveraged as drug targets for driving the network into healthy state. These results are based on empirical data of cancer causing mutations and known drug targets. Our finding of indispensable genes to be key in steering the state of the system is consistent with observations made from other biological networks [43]. Further, this observation associating disease control to nodes in the exclusion sets of MDNS is aligned with Liu *et al.* [44]. Beyond identification of nodes central to control of a network, our study involves a systematic analysis of biological mechanisms underlying cancer. Consistent with earlier studies, our observations indicate that cancer mutations occur in signalling proteins that are enriched as BDNs [77].

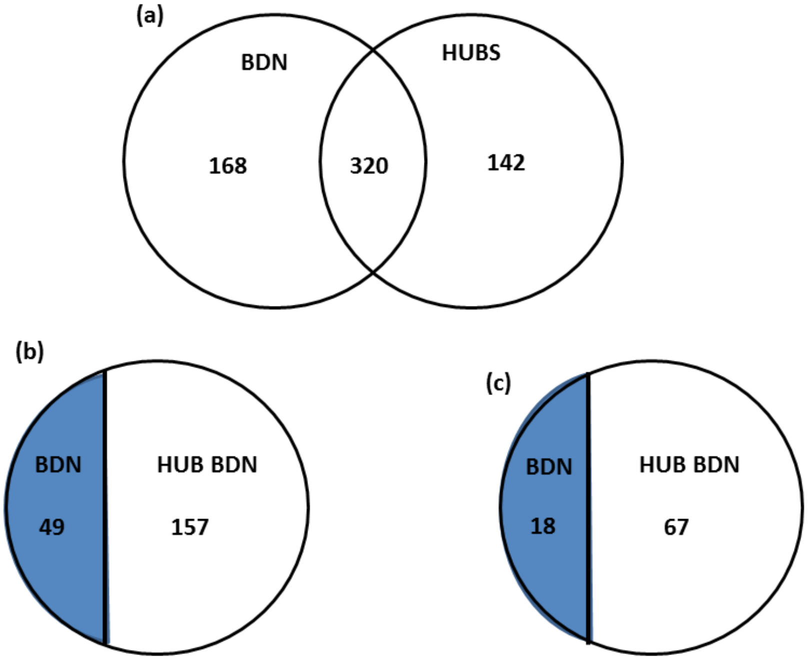


Figure 3.13: Comparison of BDNs and degree hubs highlighting that controllability analysis adds to disease network analysis from the perspective of hubs. (a) While 320 BDNs were also hubs, the former yielded 168 driver genes that were not identified as hubs. (b) BDNs also returned 49 non-hub genes associated to cancer. (c) Similarly, BDNs yielded 18 ANPD targets that were not present in the hub set.

Earlier studies have pointed out relevance of hubs (nodes with high degree and betweenness) in disease networks [22]. Though the backbone driver nodes identified from our study are characterized with relatively high degrees when compared to that of peripheral and ordinary driver nodes, the range of variation is wide (Figure. 3.2). Between BDNs (487) and comparable number of degree hubs

(464 with $k \geq 4$), we found that 34% of BDNs were unique (non-hubs) control elements. BDNs also comprise of cancer associated genes (49) and targets of ANPDs (18) that are unaccounted by degree hubs (Figure. 3.13). These observations imply that structural controllability offers a complementary method for identification of genes critical for control of disease. Further, when we probed for genes that develop resistance to ANPDs from COSMIC database [66], out of 9 genes from HCSN that were reported to undergo drug based resistance to 8 ANPDs, majority were BDNs. These included 5 Indispensable BDNs and 3 Neutral BDN and 1 PDN (Table 3.4).

Drug	Gene
Imatinib	**ABL1**, PDGFRA, **KIT**
Gefitinib, Afatinib	**EGFR**
Bosutinib, Dasatinib, Nilotinib	**ABL1**
Sunitinib	PDGFRA, **KIT**
Vismodegib	**SMO**
Selumetinib	**MAP2K1**
Endocrine therapy	*ESR1*
PD0325901	MAP2K1, *MAP2K2*
Dabrafenib	*BRAF*

Table 3.4: The genes that undergo resistance to drugs. The drugs highlighted are part of ANPDs in our study. The genes highlighted by bold are Indispensable BDN, italicized are Neutral BDN and rest is PDN.

HCSN shows a small world and scale free architecture, similar to many other biological regulatory networks. While this implies ease of information flow across the network, it has been reported that many regulatory networks (such as Yeast, *E. coli.*) present large number of driver nodes indicating that achieving ‘structural control’ in such networks is not easy [30]. The scale free nature of degree distribution is one of the key aspects of network topology that has been shown to be linked to this feature. Large number of driver nodes indicate that control of information flow so as to drive the state of the network to a desired state is arduous as one would need to provide large number of input signals through these driver nodes. On one side, consistent with evolutionary arguments, this would mean that the signalling network is robust to spurious random mutations and resists being driven into a cancerous state. On the other hand, from the perspective of finding control mechanisms (drugs) through which the network could be driven from cancerous to healthy state, this implies that such a drug-mediated control may not be easy to achieve through a small number of 'targets'. Not far from reality, cancerous state does come across as a not-so-easy state to control from therapeutic angles. One of the implications of this conclusion is that one needs to find alternative ways for achieving therapeutic solutions by mediating through a large number of genes. Studying network controllability of biological networks is difficult as the true dynamics of the systems is often unavailable. Here we apply linear control theory for studying a non-linear system assuming that most biological systems operate under homeostasis. Identifying driver nodes in a system could provide a global impact but when to target them in a perturbed system requires knowledge of underlying dynamics, which is often unknown in most biological networks [73]. We believe that this approach that amalgamates engineering concepts with biological knowledge could provide better insights into mechanisms of cancer.

CHAPTER 4

Investigation of HIV-1 Human Molecular Interactome

4.1 Introduction and motivation

The rapid emergence of infectious diseases like Ebola, HIV-1, Hepatitis calls for immediate attention to determine practical solutions and strategies to combat them. HIV-1 continues to be a serious health issue world wide. HIV-1 is a retrovirus comprising of 9 genes that code for 21 proteins. HIV-1 like all other viruses needs to exploit the cellular machinery of the host to replicate. To achieve this, the viral molecules target many host molecules through a complex network of mostly protein-protein interactions [81]. In reality, the virus with its minimalistic genome needs to interact with the host system in a highly complicated manner, thereby regulate the molecules and the functions to efficiently hijack the host and replicate successfully.

With the advancement in proteomics and genomics there is an ever-increasing volume of data generated on the host genes that are modulated by HIV-1 during infection. These data have helped in a better understanding of the various interactions between the host and viral proteins. Further, the data has helped to construct several viral-host molecular interaction networks. Several databases like VirusMint [82], VirHostNet [83], JNets [84] and HHPID [85] catalogue these viral-host interactions and provide a global snapshot of the host-cellular processes perturbed by HIV-1 infection. Given the availability of this host-pathogen interactions, it is now possible to study such systems as networks. For instance, MacPherson *et.al.* [86] have modelled the HIV-1-host interaction as a network to identify core processes that are active during infection and also cellular subsystems that are affected by HIV-1. Another study [49] has integrated different viral and bacterial human protein-protein interactions as networks and provided a holistic view of strategies used by pathogens to subvert human cellular processes and infect cells.

Graph theory has been widely used as a model to describe and visualize host cellular systems at molecular level. Many studies have highlighted that the 'hubs'-highly connecting proteins and high centrality proteins,are targeted by viruses [49, 87, 88].Majority of these studies analyse and visualise the viral-host interactions together. To verify these and to get a better understanding of the network parameters central to HIV-1 human molecular interactome, we performed network analysis on HIV-1 human interaction network. The results are presented in (Appendix 2). These studies have however ignored the dynamics of infection while in reality HIV-1 perturbs the host cellular system in an indiscriminate manner. So an interesting question is how does HIV-1 orchestrate its control on the host system in a order to efficiently hijack the host cell.

Control theory has emerged as a mathematical framework for understanding the dynamics and how best to control an engineered system. Control theory has been widely applied to the study of complex networks and to identify ways for controlling its behaviour [29, 30, 31, 51]. The goal is to identify the minimum number of inputs, termed 'driver nodes', that can alter the state of the system. Past studies have employed control theory to identify minimum number of proteins required to control diseases like cancer [44, 46, 47, 89] and other essential biological functionalities [42, 43, 90, 91].

In this chapter, we use control theory as a paradigm for investigation of HIV-1 infection and hijack. While past studies have explored the use of control theory and shown that only a few molecules are associated as viral targets [42, 43], they do not talk about cause of hijack. Here, we have modelled HIV-1 host interactions as dynamic network both with and without the HIV-1-host interactions included. We investigated this network from controllability perspective to explore the utility of principles of control theory to dissect mechanisms of HIV-1 infection. By constructing complex networks of molecular events mediated by virus during its exploitation of host machinery we can improve our current model of HIV-1 infection and host cell perturbation. This information aids in antiviral treatments.

4.2 Network compilation and characterization

The directed human protein-protein interaction (PPI) network was obtained from Vinayagam *et.al.* [43]. This consisted 6339 proteins and 43813 interactions. Interaction direction represents potential signal flow between proteins, which was predicted using a Naive Bayesian classifier [92]. In order to generate an integrated HIV-1 human molecular network, we first collected HIV-1 interactions with human from HHPID [85]. A total of 15230 interactions were retrieved and were further curated by ignoring the number of publications, counting each reaction type only once and finally selecting only those nodes that had shared nodes with the directed human signalling network. Among 6339 human proteins, 2529 human proteins within the PPI network interact with HIV-1, resulting in a total of 5811 additional interactions. The directions in the HIV-1 human molecular network were assigned using the method provided by MacPherson *et.al.*, where each HHPID interaction was assigned a direction based on the type of interaction. Direction represents whether the viral protein acts upon the host or vice versa. For instance, Nef inhibits ACHE would be given a forward direction as the viral protein acts upon the host, whereas Nef is inhibited by ACHE would be attributed a backward direction, since it is the host protein that acts on the virus one [86]. Thus the HIV-1 human molecular network consisted of 6361 proteins and 40625 interactions (Figure 4.1).

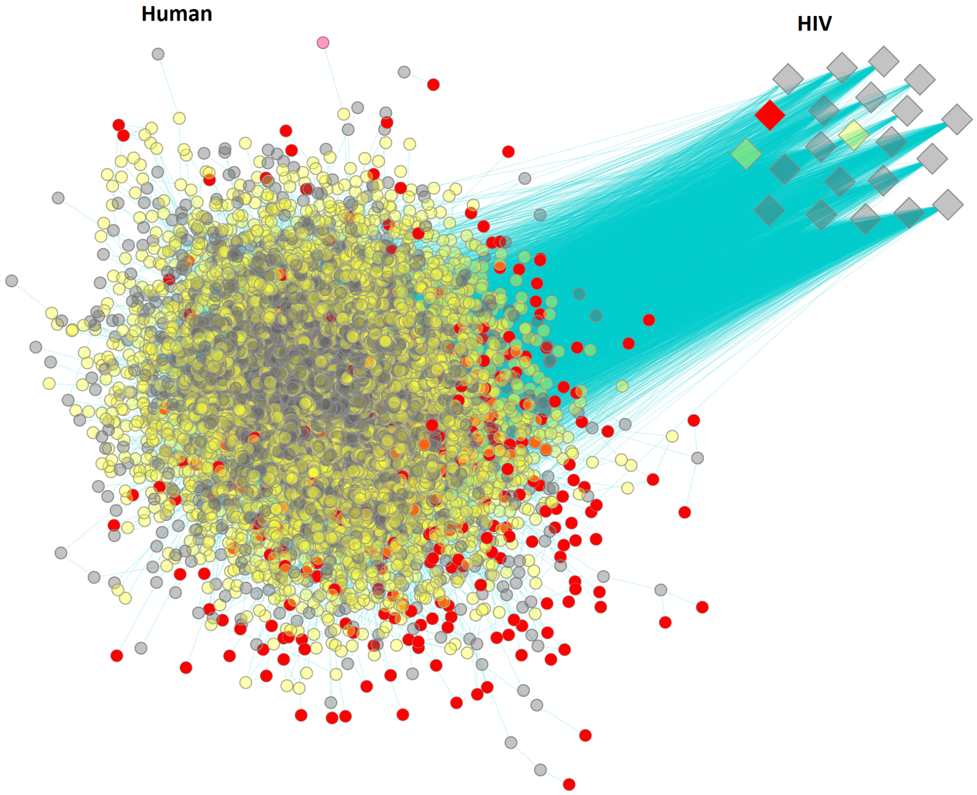


Figure 4.1: Human HIV-1 molecular network. Using controllability analysis, driver nodes were identified and characterized as critical (shown in red), ordinary (shown in yellow) and redundant (shown in grey). The HIV-1 proteins are represented as diamonds.

4.3 Characterization of nodes in directed human PPI network based on controllability analysis

We analysed the human PPI network for its controllability and characterized nodes based on their role in control. The maximum matching model was used to identify the minimum number of driver nodes N_D [30] (See Chapter 2). 36% of the nodes were classified as driver nodes. Though the size of N_D is unique, the set is different so we further classified the nodes based on their presence in the minimum drive node set (MDNS). 6% of the nodes were critical meaning they are present in all MDNS and have to be controlled. 53% of the nodes were ordinary i.e present in some MDNS and 42% were redundant i.e never part of MDNS are not required for control. Based on this classification, we identified the controllers of the network primarily critical nodes i.e nodes where inputs need to be provided in order to gain control over the system.

In order to assess the role of human proteins in context to cell signalling, we characterized them as either signalling proteins, kinases, receptors and transcription factors [43]. In total, the proteins were classified into 1006 signalling proteins, 545 receptors, 366 kinases and 1150 transcription factors (Table 4.1).

	Critical			
	Observed	Random mean	Z-score	p-value
Signaling proteins	51	60.32	-1.33	0.1834
Receptors	351	32.25	62.5	p<.001
Kinases	29	21.62	1.7	0.0891
Transcription factor	1	68.57	-9.14	6.25e-20
	Ordinary			
	Observed	Random mean	Z-score	p-value
Signaling proteins	388	528.76	-9.69	3.33e-22
Receptors	112	286.43	-15.29	8.92e-53
Kinases	160	192.18	-3.42	0.0006
Transcription factor	367	603.15	-16.09	3.00e-58
	Redundant			
	Observed	Random mean	Z-score	p-value
Signaling proteins	567	418.19	10.69	1.13e-26
Receptors	82	226.79	-12.89	5.12e-38
Kinases	177	152.2	2.7	0.0069
Transcription factor	782	477.92	20.14	3.29e-90

Table 4.1: Classification of driver nodes based on cellular localization.

The critical nodes were highly enriched as receptors being predominantly in the upstream of signalling processes, while underrepresented as transcription factors. On the other hand redundant nodes were enriched as transcription factors and signalling proteins mainly being in the downstream of signalling processes and under represented as receptors. Ordinary nodes did not show any specific cellular characterization.

4.4 Role of receptors in viral entry

Viruses replicate within living cells and use the host cellular machinery for the synthesis of their genome and other components. To gain access into the cell, they have evolved different mechanisms to deliver their genes and accessory proteins into host cells. Viruses take advantage of receptors to gain entry into the cell [93]. The first step of HIV-1 replication is binding and entry into the host cell with the help of primary cellular receptors that plays a major role in determining viral tropism and ability of HIV-1 to degrade the host immune system. To infect cells, the HIV envelope protein (Env) binds to CD4 receptors and then other coreceptors to trigger fusion of viral and host cell membranes, initiating infection [94].

Here we analogise virus as an external input to the cell, that gains entry into the cell through receptors that were observed to be predominantly critical nodes. In order to access the role of receptors for viral entry, we looked at it from control perspective and calculated control centrality. It is a measure that tells us how powerful a node is in controlling the network. Mathematically, control centrality of a node captures the dimension of controllable subspace or the size of controllable subsystem when we control node i only [38]. Out of 6339 proteins from the signalling network, about 40% of the nodes are targets of HIV-1 in which 114 are receptors. We compared the control centrality among the different driver node class. The mean control centrality for receptors was higher compared to signalling proteins, kinases and transcription factors (Table 4.2). We validated the difference in mean using one way ANOVA at significance level p<0.0001 and F score of 442.68. Further post-hoc test also showed significant difference in mean control centrality between receptors and other class of proteins but no significant difference between the mean control centrality of signalling proteins and kinases was seen. Control centrality measure reveals the importance of receptors for viral entry that has profound implications for viral tropism, transmission, pathogenesis and therapeutic interventions.

	Mean	Standard deviation
Signalling proteins (1006)	0.239	0.049
Receptors (545)	**0.257**	**0.006**
Kinases (366)	0.242	0.041
Transcription factors (1150)	0.173	0.113

Table 4.2: Mean control centrality measure among different class of signalling proteins.

4.5 Controllability analysis of HIV-1 human molecular network

To better understand the pathogenesis of HIV-1, we modelled it as a dynamic network with viral interactions involved in forming the HIV-1 human molecular network and without viral interactions in the directed human PPI network. We termed these networks as infected and uninfected to perform controllability analysis in order to investigate how viruses alter the state of the network and drive into infection.

Driver nodes were classified and compared in the uninfected and infected network (Table 4.3). In the HIV-1 infected network the number of driver nodes $N_D = 2264$ whereas the same was $N_D = 2283$ prior to HIV-1 infection. The number of driver nodes did not change significantly. We further characterized all the nodes in the network and identified the preserved critical nodes that could be leveraged as potential drug targets for antiviral therapy.

Network	**N**	**E**	**N_D**	**Critical**		**Ordinary**		**Redundant**	
Uninfected	6339	34813	2283	377		3330		2632	
				HIV	**Human**	**HIV**	**Human**	**HIV**	**Human**
Infected	6361	40625	2264	1	266	2	3443	19	2630

Table 4.3: Comparison of driver nodes in uninfected and infected network.

Among the HIV-1 proteins, p51 was the only node classified as critical, nucleocapsid and p1 as ordinary and remaining 19 proteins as redundant. One can observe a difference in critical nodes upon infection among the human proteins. Initially 377 proteins were characterized as critical which subsequently reduced to 266 upon HIV-1 interaction. Among the critical proteins 111 changed to ordinary nodes in the infected network. ICAM-1 which was initially classified as ordinary changes to redundant upon infection. Three proteins namely MAD2L1, SCNN1A and TRIM6 which were initially redundant changes to ordinary in the infected network.

The critical nodes consisted of 127 proteins that were exclusively targeted by HIV-1. Upon infection, 15 proteins preserved their critical node status. We propose that these 15 proteins could be possible targets for anti-viral drugs (Table 4.4). Some of these have been experimentally identified as RNAi screens and druggable genes [43]. By comparing the two networks and looking at the nodes that control the system provides us with an better understanding of HIV-1 pathogenesis.

Critical node	Cellular localization	RNAi screen	Druggable
ADORA2A	Receptor		Yes
KIR3DL1	Receptor		
SIGMAR1	Receptor		
NRXN1	Receptor	Yes	
PTH1R	Receptor	Yes	Yes
KLRC2	Receptor	Yes	
PFKM	Signalling protein	Yes	
NTRK3	Receptor		Yes
TFR2	Receptor	Yes	
CD97	Receptor	Yes	Yes
LPAR2	Receptor	Yes	Yes
PTPRN2	Receptor	Yes	Yes
IL22RA1	Receptor		
LPAR3	Receptor	Yes	Yes
OPN4	Receptor	Yes	Yes

Table 4.4: Preserved critical nodes that are HIV-1 targets.

4.6 HIV-1 target indispensable nodes

We also investigated the network by looking at the importance of a protein upon its removal on controllability. The proteins/nodes in the network were classified based on their impact on size of driver nodes N_D. In order to efficiently control a system, we need to steer inputs to minimum number of driver nodes. If the number of driver nodes increases, controlling such systems is difficult. In order to enumerate the importance of a node on N_D, we classified each nodes as indispensable, dispensable and neutral as mentioned in chapter 2. A node is indispensable if the number of driver nodes increases, dispensable if there is decrease in number of driver nodes and neutral if there is no change in the number of driver nodes. 21% of the proteins were classified as indispensable, 37% dispensable and 42% as neutral in the uninfected network. Further, we characterized these nodes based on their degree distribution. We observed that the indispensable nodes had high in, out and total degree. Dispensable nodes were predominantly low degree nodes while neutral nodes had no distinct degree characterization.

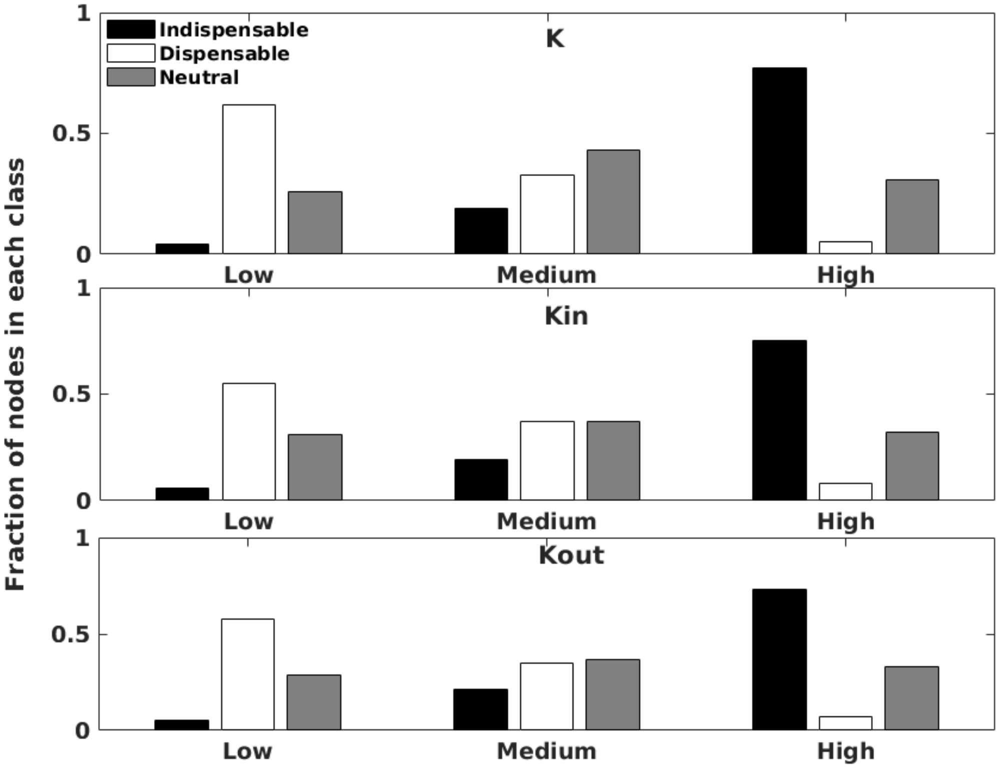


Figure 4.2: Degree characterization of nodes. Height of bars corresponds to fraction of driver nodes in each class of degree. K= Total degree, Kin=In-degree, Kout=Out-degree.

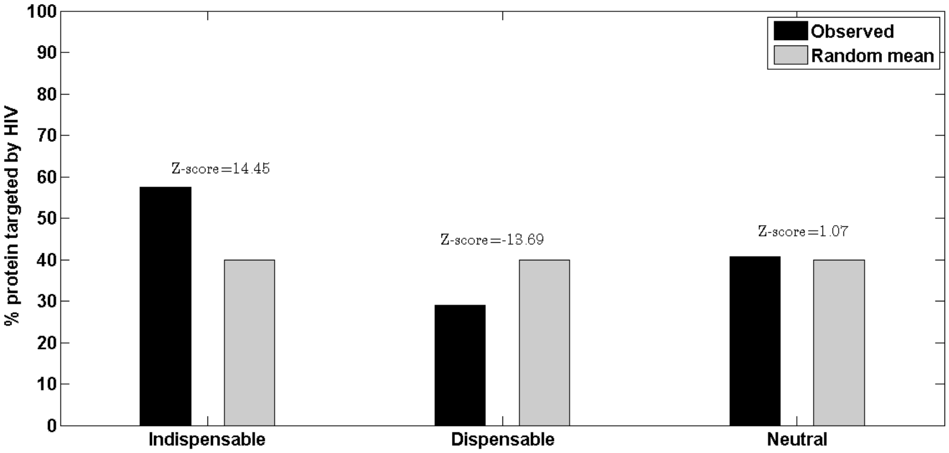


Figure 4.3: HIV-1 targets among different class of proteins based on their impact on N_D.

We also categorized the proteins in HIV-1 infected network (Table 4.5) and observed no change in the classification of nodes in the infected and uninfected network. While this comparison does not reflect the change in state from healthy to normal, it does highlight the importance of indispensable nodes for HIV-1 targets. Among the HIV-1 targeted proteins, 58% were indispensable, 29% were dispensable and 41% were neutral (Figure: 4.3). This shows that HIV-1 target indispensable nodes that have high degree and are viral targets that have direct interactions with virus proteins. This is in accordance with other studies which have shown that viruses target high degree nodes [49, 81, 95].

Network	**Indispensable**		**Dispensable**		**Neutral**	
Uninfected	1330		2347		2662	
	HIV	**Human**	**HIV**	**Human**	**HIV**	**Human**
Infected	19	1331	1	2346	2	2662

Table 4.5: Classification of proteins based on their impact on N_D.

4.7 Controllability analysis of T-cell activation network

Of all immune cells, CD4+ T lymphocytes are the most important cell type involved in HIV-1 infection. HIV-1 enters into these cells through the interaction of its envelope protein with CD4 receptor and CXCR4 or CCR5 co-receptors [96]. Apart from viral entry, this interaction of HIV-1 and receptors triggers signals that activate multiple pathways stimulating cellular responses within the host cell. We have analysed the signalling pathways essential for T-cell activation through the T-cell receptor complex (TCR), its co-stimulators and co-inhibitors by constructing a networks prior to HIV-1 interaction and post interaction from data of Oyeyemi *et.al* [97] that contains 137 nodes (16 HIV and 121 human proteins) and 336 interactions. The T-cell activation network comprised of 121 proteins and 255 edges, while the HIV-1 T-cell activation network has 137 proteins and 367 interactions. All the interaction type either inhibit or activate a protein and hence represent state change of nodes in the network.

We classified the proteins/nodes in both the networks based on their role as driver node and as well as their role in altering the number of driver nodes N_D.

Upon activation with HIV-1, the critical and dispensable proteins reduced while there was an increase in ordinary and neutral proteins (Table 4.6). We

	Before Infection (N=121)		After Infection (N=137)	
	N	%	N	%
Critical	16	13.22	14	10.22
Ordinary	42	34.71	55	40.15
redundant	63	52.07	68	49.64
Indispensable	36	29.75	41	29.93
Dispensable	30	24.79	26	18.98
Neutral	55	45.45	70	51.09

Table 4.6: Classification of proteins based on controllability analysis.

also looked at the classification of HIV-1 proteins. Env gene glycoprotein namely gp120 was classified as indispensable. This could possible answer its importance in viral entry. Vpu and gp41 were classified critical, meaning they act as driver nodes. gp41 serves to anchor gp120 to which it remains attached and to effect membrane fusion. On the other hand, vpu accessory gene contribute virus replication by facilitating the assembly, binding and release of mature virus particles. Thus one assists in viral entry and the other in replication that leads to infection.

	Critical	Ordinary	Redundant
Indispensable			gp120
Dispensable		nucleocapsid,rev,p6,rt	
Neutral	vpu,gp41	protease, nef, vif, vpr,tat,gp160	capsid, matrix,in

Table 4.7: Classification of HIV proteins based on controllability analysis.

When compared the driver nodes in both the networks, 12 nodes preserved their critical node status. Out of these CD45 and DLGH1 interact with gp120, gp160 and nucleopcapsid. Interestingly, logical steady state dynamic studies have shown that switching off DLGH1 activated gp160 and deactivates nucleocapsid thus showing that viral cells may control host signal transduction dynamics by inhibiting or activating host proteins to favour viral replication needs at various stages of infection [97]. CD4, loses its critical node status upon interaction with gp160, gp120, nef, vpr, vpu, vif, B-arrestin and protease.

Since the type of interaction determines the course of signal transduction, we also looked at the role of the edges/interactions in control. We classified links as critical: If on its removal the number of driver nodes increases, redundant: It can be removed without affecting the current set of diver nodes and ordinary: If it is neither critical nor redundant [30].

	T-cell activation E=287		HIV-1 infected E=401	
	Activation	**Inhibition**	**Activation**	**Inhibition**
Critical	31	4	36	7
Ordinary	135	29	155	40
Redundant	78	10	129	34

Table 4.8: Classification of edges based on control.

A difference in number of edges is due to the fact that while identifying driver nodes, the weights of the edges is not considered, i.e, in the presence of activation and inhibition reaction between same proteins simultaneously, the edge was considered only once in driver node identification. While in the case of edge controllability, the reactions are considered different and hence the increase in number.

The critical edges are more of activation type in both the networks compared to inhibition 4.8. A detailed list of the critical edges are provided in Appendix 2. In the HIV-1 T-cell activation network, 7 interactions were critical out of which 5 were activation and 2 inhibition type of reaction A6. Our results suggests that viral-host interactions contribute more towards activation of host proteins and hence aid in activation of other downstream cellular process of these proteins. Also, the critical viral-host interactions, could be useful check points for antiviral therapy. This suggests that virus does not impede signals but rather amplifies them [97].

4.8 Conclusion

Viruses act by exploiting their host cellular machinery in order to replicate. We have applied control theory to show how HIV-1 hijacks the host cells in order to enhance its replication. We investigated two networks, viz., uninfected and infected, from the perspective of control and used different measures of complex networks like degree, centrality to identify nodes critical in control. We found that receptors were part of driver nodes and had higher capacity to reach subsystems based on their control centrality measure. This explains the role of receptors in viral entry and how efficiently they facilitate the virus to activate specific pathways required for successful infection. Further, we propose preserved critical driver nodes as potential drug targets since they do not change their critical driver node status upon HIV-1 infection. These nodes were predominately receptors, few of them being RNAi screens and verified drug targets. By targeting these, one could potentially suppress HIV-1 replication. As mentioned earlier, receptors facilitate viral entry and is the first step in HIV-1 replication thus is target for several antiretroviral agents: attachment inhibitors, chemokine receptor antagonists and fusion inhibitors [98, 99]. The indispensable nodes/protein were found to be crucial for controlling the network as they had higher degree and were HIV-1 targets. Based on this study, we propose that viruses efficiently use their minimalistic genome to replicate in the host cell.

Analysis of the T-cell activation pathways adds to a better understanding of the viral hijack which can be further extended to other pathways involved in HIV-1 infection to better understand signal information flow and to identify drug intervention points. We also analysed the uninfected and infected networks using minimum dominating set model (MDS) [31], to gain further insights on the viral hijack mechanism. These results are provided in Appendix 2. From control systems perspective cellular processes can be viewed as an intricately regulated system and viruses can be seen as ’attacks’. Our work provides a conceptual and technical framework for incorporating signalling pathways to get more insights on mechanism of pathogenesis. This method to analyse complex biological networks has the potential for furthering systems biology research.

CHAPTER 5

Investigation of Control Configuration in Biological Networks

5.1 Introduction and motivation

Biological processes are governed by physical and chemical interactions between proteins. Cells respond to environmental changes that influence the signalling network. For instance, in order to drive the human cancer signalling network from a cancerous state to a healthy state, a natural conversion of a cancer cell into a normal cell is not possible. This can be achieved by perturbing the system using drugs or other external hues.

The maximum matching model developed to identify driver nodes, nodes with which we can achieve full control, predict the existence of multiple control configurations, in turn prompting us to classify each node in the network based on its role in control. In Chapter 2, we classified a node as critical, ordinary, redundant/non-driver according as its presence in all, some, none of the sets of driver nodes. This classification of driver nodes has lead to identification of two distinct control modes/configurations for a network. The control mode of a network can be altered through small structural perturbations [36]. In this chapter we perturb the network configuration, classify the driver nodes in the perturbed network to gain insights into the dynamics of the system represented by the initial network. We use this technique to identify the control modes of some biological networks.

5.2 Control modes in networks

The role of each node in controlling a network is identified by classifying each node into one of the three categories based on its presence in minimum driver node set (MDS). Critical, meaning the node is always a driver node, i.e. it is present in all MDS; redundant if it never acts as a driver node and therefore not a part of any MDS; ordinary if it acts as driver node in some but not all MDS. This classification leads to bifurcation phenomenon, predicting that a bimodal behaviour determines the controllability of many real world networks [36]. This bimodality uncovers two control modes, centralized and distributed. The control modes are determined by fraction of redundant nodes.

Studies on network models have shown that for networks with symmetric in- and out-degree distributions, the fraction of redundant nodes ($n_r = N_R/N$, where N_R is number of redundant nodes and N is total nodes in the network) undergoes a *bifurcation* at a critical mean degree $\langle k \rangle_c$. In particular it has been shown that for low mean degree $\langle k \rangle$ n_r is uniquely determined by $\langle k \rangle$, but when $\langle k \rangle$ exceeds $\langle k \rangle_c$ there exist two different values for n_r, one with very high and the other with low values leading to bimodal behaviour (Figure 5.1). Hence for large $\langle k \rangle$ two control modes coexist [36]. These modes are:

1. **Centralized control** : For networks that follow the upper branch of the bifurcation diagram (Figure 5.1), most nodes are redundant nodes (Figure 5.1(c)) and n_r is very high. In these networks, only a small fraction of nodes are required for control, i.e., $n_c + n_i$ is small, where n_c is the fraction of critical nodes and n_i is the fraction of ordinary nodes in the network. This mode of control is like that of an organisational setting where leadership is concentrated in the hands of a few supervisors and the employees are only executors.

2. **Distributed control**: In networks that follow the lower branch of the bifurcation diagram $n_c + n_i$ can exceed 90% of the nodes, meaning most nodes participate as driver nodes in some control configuration. This mode of con-

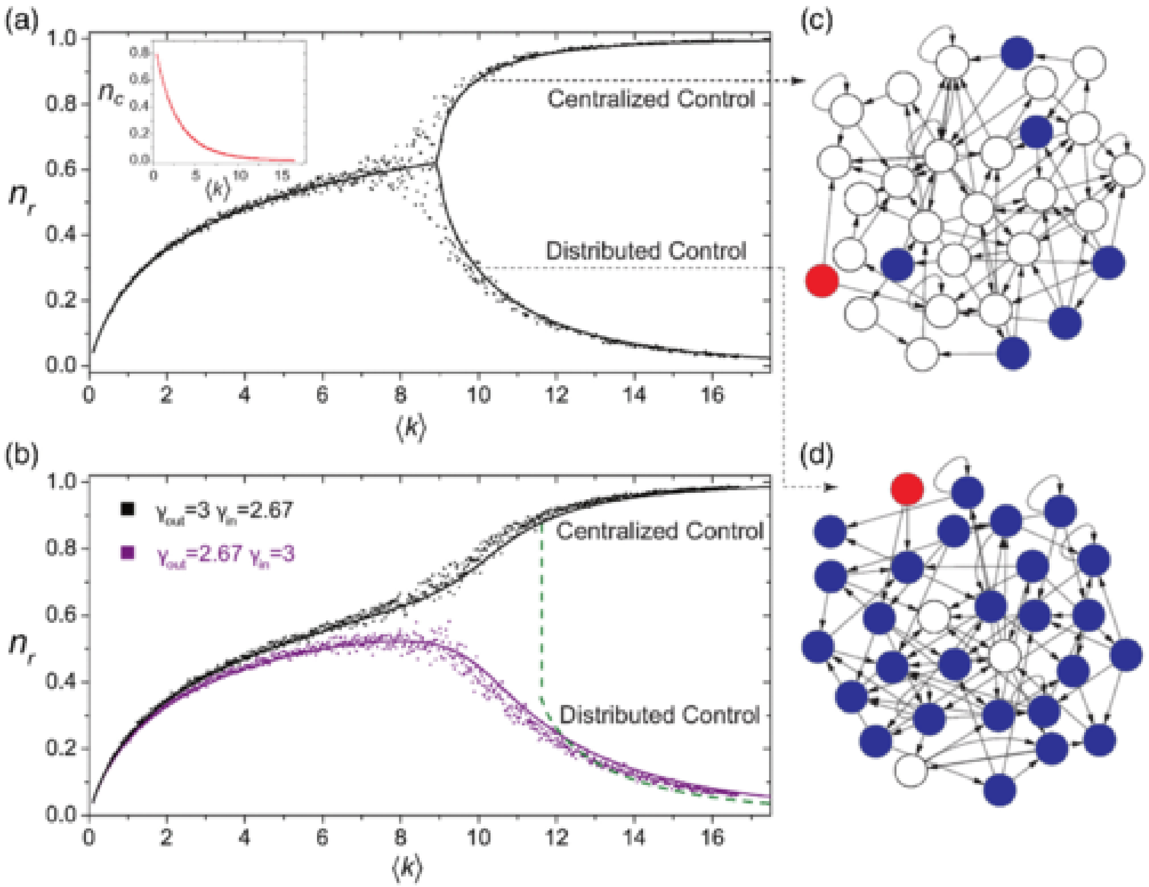


Figure 5.1: Emergence of bimodality in controlling complex networks. (a) Bimodality in network for symmetric in- and out-degree distribution ($\gamma_{in} = \gamma_{out} = 3$) for high $\langle k \rangle$. The plot is of n_r and n_c) vs $\langle k \rangle$ in scale-free networks. (b) n_r in networks with asymmetric degree distribution ($\gamma_{in} = 2.67, \gamma_{out} = 3$, upper branch and $\gamma_{in} = 3, \gamma_{out} = 2.67$ lower branch). the control mode is predetermined by their degree asymmetry. (c) and (d) are networks representing centralized or distributed control. Both networks have $N_D = 4$ and $N_c = 1$ (red node) but different number of redundant nodes (uncoloured node), $N_r = 23$ in (c) and $N_r = 3$ in (d). From Y.Liu *et.al.*, 2016.

trol is like that of an organisation where different employees take leadership roles at different times as in case of change in shift.

5.3 Identification of control modes in biological networks

Identification of control mode and altering the state of a network can be achieved by its transpose. The transposed network of a given network is obtained by reversing the direction of each link of the given network. The control mode is captured by comparing n_r of the network with that of its transposed network n_r^T. If $\Delta n_r = n_r - n_r^T > 0$ then the network is said to be in centralized mode. If $\Delta n_r < 0$ then the network is said to be in distributed and if $\Delta n_r = 0$ then the mode of control of the network cannot be determined. For some biological networks we have determined the control mode as tabulated below (Table 5.1).

Network	Nodes	Edges	n_r	n_r^T	Δn_r	Control mode
Cancer signaling	1232	3060	0.396104	0.350649	0.045455	centralized
Directed human PPI	6339	34813	0.41505	0.425304	-0.010254	undetermined
HIV-human molecular	6361	40625	0.416352	0.427449	-0.011097	undetermined
T-cell activation	121	255	0.520661	0.528926	-0.008265	undetermined
HIV-T-cell activation	137	367	0.49635	0.613139	-0.116789	distributed
E. coli transcription	423	578	0.096926	0.269504	-0.172578	distributed

Table 5.1: Control mode of some biological networks

The human signalling network follows a centralized mode of control. One possible reason could be the nature of the network. In this network, cells receive information from neighbouring cells through receptors which in turn activates or inhibits proteins down stream of it. In this way many proteins coordinate with one another to maintain cellular information processing. Since there is a kind of hierarchy among the proteins, this could be a reason for why such systems follow a centralized mode of control.

The HIV-1 t-cell activation and the *E. coli* transcription networks have distributed control. In the case of HIV-1 t-cell activation network, the HIV-1 proteins interact with human proteins for its replication. Further due to infection, the human proteins independently activate the immune system to combat the virus, thus allowing for a distributed mode of control. So is the case with the *E. coli* transcription network which comprises motifs that have a specific function in determining different gene expressions. In the case of the directed human PPI network, HIV-1 human molecular network and T-cell activation network, the mode of control cannot be determined.

5.4 Conclusion

Identifying the control mode of a network throws light on its dynamics. This tells us the ease with which one can control such systems. Biological networks are in general difficult to control due to their underlying complexity. By identifying the mode, we can perhaps find ways to effectively control the system. Further, identifying the control mode of a biological network will add to better understanding of mechanisms underlying its nature. For most biological networks, many molecular details like function, interactions, kinetic parameters among genes and proteins are unknown. Despite the lack of a thorough knowledge of the entities in the network, the control mode can shed some light on the mechanisms of the network.

The fact that control modes are suited for different task raises an important question as to what is the least modification required to change the mode of a network from centralized to distributed and vice versa? It is to be noted that if the original network is centralised then its transpose is distributed [36]. It is of interest to identify as few edges as possible that when reversed can change the mode of control. Biological processes can now be reversed at a cellular level. For instance tumorigenesis can be reversed through targeted inactivation of oncogenes [100]. Study by Cho *et.al.* have shown that the network can be rewired by inducing reshaping of the attractor landscape (a molecular phase portrait describing the dynamics of a molecular regulatory network) and phenotype landscape (that is determined by the steady states of particular output molecules) in the attractor landscape [101]. One should bear in mind that, the purpose of control determines the control mode best suited for the system.

Chapter 6

Investigation of Control Profile in Biological Networks

6.1 Introduction and motivation

Studying control properties of a complex network provides insights into how the dynamical system represented by the network can be influenced to achieve desired behaviour. Many biological systems which are important for the normal regulation of a cell are dynamic systems. Undesirable behaviour of such systems is observed in the form of diseases and has lead to interest in studying the control of such complex systems through the corresponding networks. In the preceding chapters, we have established the connection between driver nodes which when provided with an external input can trigger the system to a desired state and their biological implications. The mathematical models implemented therein have shown the minimum number of controls required to control the system [30]. While topological properties of a network like its degree distribution is correlated with the minimum number of controls, it does not provide an explanatory detail of each control. For example a financial system and a biological system may require the same number of controls, but the structures giving rise to these controls may be different. Understanding the control properties of a complex network requires more than just knowing the number of controls [102]. For an effective control strategy, it is also important to characterize the functional origin of each control. To know why a node is a driver node, we look at the control profile of the network. It is a statistic that quantifies the different proportions of control-inducing structures present in the networks [50].

Structural controllability formalized the idea that propagation of control is influenced by a backbone of directed paths/stems and cycles/buds that form a structures called cacti [29]. Each stem requires at least one control [53]. A cycle can be controlled by exactly one node lying on it. Ultimately it is the stem that dictates the number of controls. In this chapter, we determine why a node is a driver node based on its position in the network, and identify the control profile. Further, based on the control profile we deduce the control strategy for each of the networks studied in the preceding chapters.

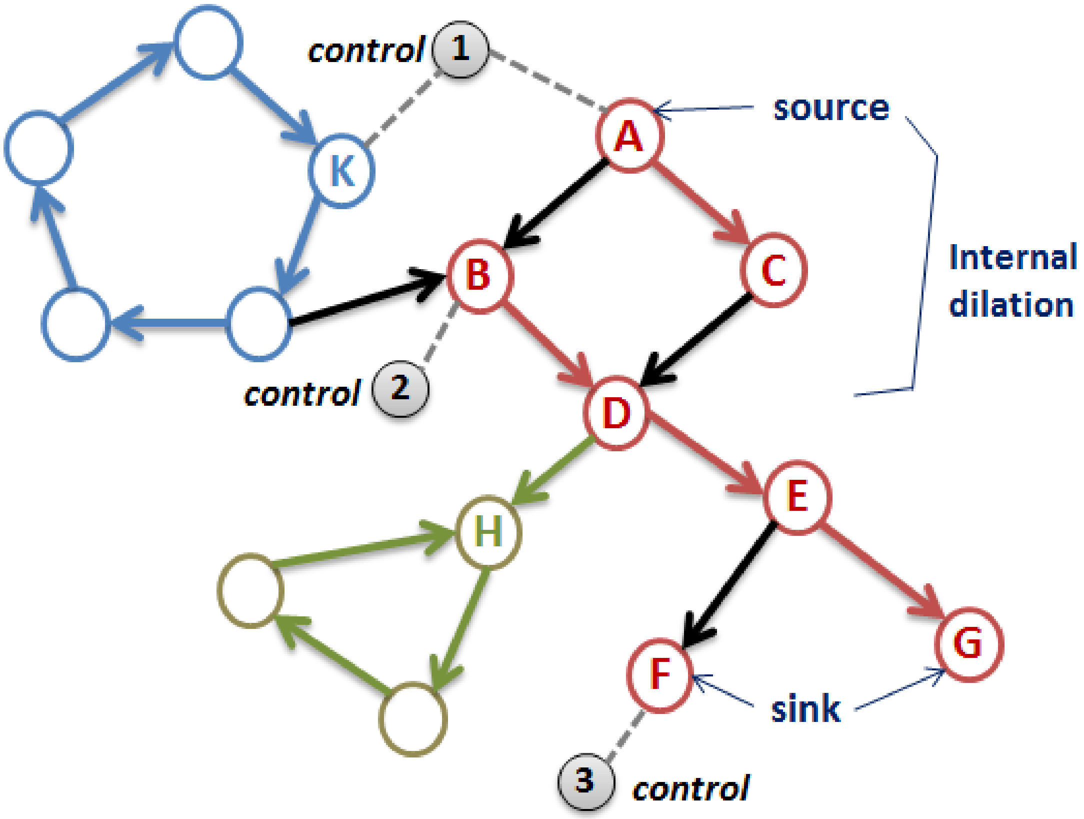


Figure 6.1: Control structure of a network constructed with stems (red) and cycles (blue and green). There is one source node A, and two sink nodes F and G giving rise to one external dilation and one internal dilation due to A-B-C-D topology. There are three controls, control 1 for nodes A and K, control 2 for node B and control 3 for node F. Black lines indicate links not involved in control structure. From Ruths and Ruths, 2014.

6.2 Identification of control profile

Since the controllability of a network is determined by the presence of cacti structure, this would mean that the number of stems in the network would determine the number of controls. The location of a stem is dictated by two topological features: sources which are at the origin of the stem (node A in Figure 6.1) and sinks which are nodes with no outgoing links that lie at end of the stem (node F and G in Figure 6.1). If N_s and N_t denote the number of sources and sinks respectively in a complex system then the number of stems in the system is given by $max(N_s, N_t)$.

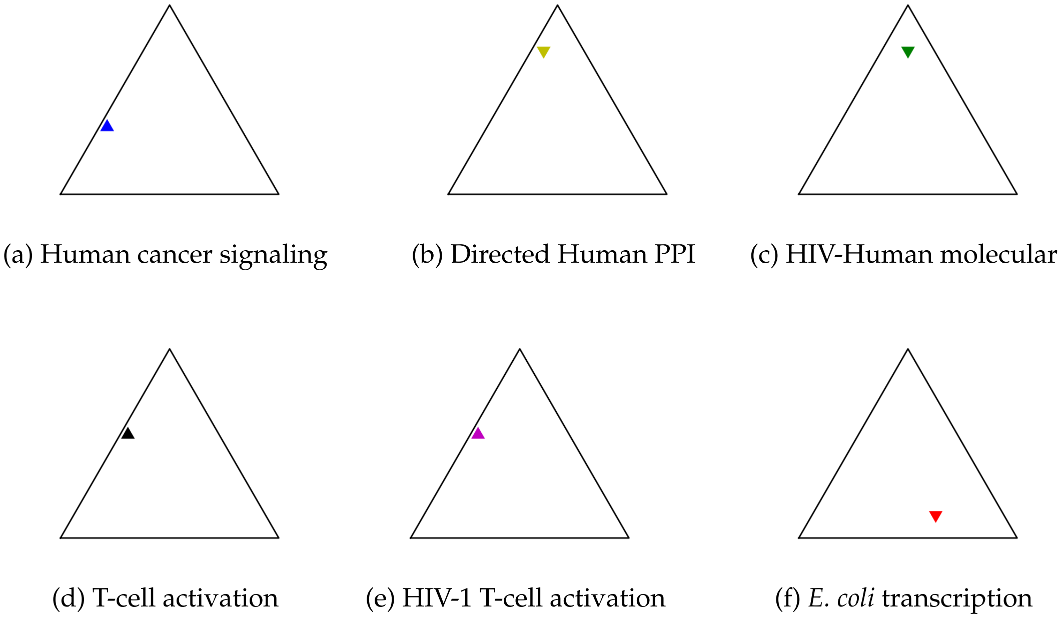
To understand why a node is a driver node, we decompose the driver nodes into three groups [50]: (1) *source nodes*- since these appear at the origin of the stem they must be directly controlled, (2) *external dilations* which arise due to surplus of sink nodes- since each source node can control only one sink node, the number of external dilation N_e is $max(0, N_t - N_s)$, and (3) *internal dilations*- a structure that occurs when a path branches into two or more paths in order to reach other nodes (Figure 6.1); N_i denotes the number of internal dilations in a network. Thus the minimum number of independent controls N_D required to gain full control is the sum of the number source nodes, the external dilations and the internal dilations i.e. $N_D = N_s + N_e + N_i$. The classification described above gives rise to a control profile for a network given by $(\eta_s = N_s/N_D, \eta_e = N_e/N_D, \eta_i = N_i/N_D)$, which quantifies the different proportions of control-inducing structures present in the network.

As in previous chapters, N_D is computed using the maximum matching algorithm. Degree analysis is used for computing N_s and N_e. Now, N_i can be solved as a function of N_s, N_e and N_D. The table below summarizes the control profile for the networks described in the previous chapters (Table 6.1).

The control profile plots for these networks (Figure 6.2) are presented as triangular plots in which the corners correspond to the points (1,0,0), (0,1,0) and (0,0,1) in the three-dimensional (η_s, η_e, η_i) space. These heat maps also capture overlapping data points. These plots were generated using ZEN package in python (`http://zen.networkdynamics.org/`).

Network	Nodes	Edges	Driver nodes	η_s	η_e	η_i
Cancer Signalling	1232	3060	47%	0.29	0	0.18
Directed Human PPI	6339	34813	36%	0.06	0.02	0.28
HIV-human molecular	6361	40625	36%	0.04	0.04	0.28
T-cell activation	121	255	29%	0.13	0	0.16
HIV- T-cell activation	137	367	25%	0.41	0	0.59
E. coli transcription	423	578	73%	0.11	0.62	0.002

Table 6.1: Control profile of some biological networks



(a) Human cancer signaling

(b) Directed Human PPI

(c) HIV-Human molecular

(d) T-cell activation

(e) HIV-1 T-cell activation

(f) *E. coli* transcription

Figure 6.2: Control profile plots of some biological networks

6.3 Interpretation of control profile in biological networks

The control profile results offer insights into high-level organization and function of complex networks. They categorise real networks into three clusters based on whether the network is dominated by source nodes, or by external dilations or by internal dilations. Random networks cannot be dominated by any set of nodes and hence cannot have any control profile [50]. We analyse the control profile of networks in our study and characterise them based on the cluster they belong to. The human cancer signalling network is source dominated (Figure 6.2 (a)). This means the ratio of sinks to sources is less than one i.e., there are fewer sinks than sources. These networks have no external dilation. But this does not mean that there are no sink nodes. It means there is at least one distinct source which reaches a sink through a directed path. Since the source nodes lie at the boundary they are easily accessible and therefore are control targets. For example, receptor proteins responsible for transducing extracellular stimuli into intracellular signals which were in chapter 3, characterised as PDNs (peripheral driver nodes) in the human cancer signalling network are the source nodes [103]. Since these source nodes are readily accessible and influence the protein interactions within the cell, they are used as potential drug targets [104]. This procedure which is in practise is in synchronisation with the results of our theoretical study.

The T-cell activation and HIV-1 T-cell activation networks are internal dilation dominated. They have no external dilation (Figure 6.2 (d) and (e)). In such signalling networks, the source and certain intra-cellular molecules drive the signal transduction within the cell. For instance the HIV-1 virus attacks the CD4 receptor which in turn activates other down-stream proteins, for the release of T-helper cells. Thus the source and some internal dilation nodes are responsible for control.

The Human PPI network and HIV-1 human molecular network are also internal dilation dominated networks (Figure:6.2 (b),(c)). Such networks lack sources which indicate clear input and sinks that indicate clear system output. The top-down architecture generally implies that control at the source cannot drive the system to any desired state. This means that the system is closed or mostly closed [50]. The PPI network and the HIV-1 molecular network also have feedback loops. Like a protein 'a' can activate protein 'b' and excess of protein 'b' could inhibit protein 'a'. This gives rise to feedback loops and forms a closed system.

The *E. coli* transcription network is external- dilation dominated (Figure:6.2 (f)). This implies that sinks outnumber the sources. So, controls applied to sources will yield correlated behaviour within the network. The genes in a transcription network exhibit high degree of correlation expression. For instance a particular gene can co-express or down-regulate the expression of another gene. Now, if we seek to fully control such a system, we need to add controls beyond the sources.

6.4 Conclusion

Control profiles offer a way to capture the origin of control in networks and to understand why a node acts as a driver node. Studying control profile of biological networks helps in improved understanding of the system and in identifying those nodes that can efficiently control the system. For the networks studied in this thesis, we can deduce, based on the control profile that the source dominated networks like human cancer signalling and T-cell signalling network can be efficiently controlled through the source nodes. In the case of internal dilated networks like HIV-human molecular network and the human PPI network which are closed systems and obey certain conservation laws, some non-source nodes are also required to gain control. We conclude that the control profile of a network adds insights about the structure of the network which could then offer strategic ways to control the system. This is helpful particularly in biological networks that are highly constrained and require large sets of control nodes to gain full control.

CHAPTER 7

Conclusions and Future Work

"Our imagination is the only limit to what we can have in the future."

- Charles F. Kettering,

The study of complex networks has promoted the development of systems biology. By building the complex network for a biological system, and through the study of network characteristics, we try to elaborate how molecules, interactions and structures of the network determine biological functions. This provides us insights about the cellular organisation, function and is also helpful in disease diagnosis, treatments and drug design. In recent years, control of complex networks has been revisited, inspiring several fundamental questions like what are the control principles of complex networks and how are they organised to balance control and functionality? To address these question several graph theoretic methods are available in complex networks as discussed in chapter 2. This motivates us to study the control of biological networks as well.

7.1 Salient contributions of the thesis

The objective of this thesis is to model and analyse biological networks from a control systems perspective. While a detailed conclusion is provided at the end of each chapter, here we provide an overall summary of the thesis. We have proposed that the ideas of structural controllability can be used, to capture how mutations caused by external perturbations changes the normal state of a cell into disease state and triggering back to desired normal state by drugs. We have also captured the role of driver nodes in the HIV-1-human molecular interactome and showed the efficiency with which the virus hijacks the host system for effective pathogenesis.

The dynamical properties of a cell are hardwired in the genome and influenced by environmental and epigenetic changes. Thus the cell is naturally receptive to external cues and this provides us an opportunity for its manipulation to achieve desired outcomes. In order to take the best advantage of this property we require a deeper understanding of when and where to apply these external influences. By looking at the control configuration and control profile of networks, one can get a better understanding of the ease with which the network can be controlled and the nature of driver nodes. Thus, the application of network controllability can be a systematic reasoning about which nodes to target to achieve a desired outcome in perturbed systems like cancer and viral infection. It also shows the strategy of these networks to circumvent external control and maintain its function even if the number of nodes being influenced is large. The above findings suggest that control theory is a promising approach to analyse complex biological networks composed of data from proteomics to transcriptomics. The integration of signalling data and viral interaction data have led to novel findings. The enrichment observed by driver nodes with disease association suggests them as important target molecules for future development and design of drugs.

7.2 Future research directions

Future work concerns deeper analysis of particular mechanisms, new proposals to try different methods and anything exciting in this field inspired by sheer curiosity. This thesis has been mainly focused on directed networks, leaving the study of undirected networks outside the scope of this thesis. The following ideas are to be explored.

1. It would be interesting to consider undirected networks like PPI networks and bipartitie networks like ncRNA-protein interaction networks relating to diseases and perform controllability analysis using the minimum dominat-

ing set method [105].

2. The current model focuses on nodal dynamics, while this can also be extended to controlling the edges that determine the kind of interaction. We have performed preliminary analysis of edge dynamics using switchboard model [32] on the biological networks considered in this thesis and present it in Appendix 3. The biological implications of this analysis are yet to be investigated.

3. Multi-layer networks may also be considered such that gene-disease-drug interactions are captured to better understand the influence of one network on the other [90]. An extension could be to identify the disease modules and look for association between driver nodes and these modules.

4. In Chapter 4 the focus was on HIV-1. This study can be extended to other viral-host molecular interaction networks to get a global map of the way different pathogens interact with the human cell.

5. Graph theoretic characterisation of driver nodes based on graph parameters like tenacity, maxflow-mincut can be applied to get more efficient control nodes.

References

[1] Albert-Laszlo Barabasi and Zoltan N Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113, 2004.

[2] Georgios A Pavlopoulos, Maria Secrier, Charalampos N Moschopoulos, Theodoros G Soldatos, Sophia Kossida, Jan Aerts, Reinhard Schneider, and Pantelis G Bagos. Using graph theory to analyze biological networks. *BioData mining*, 4(1):10, 2011.

[3] Albin Sandelin, Wynand Alkema, Pär Engström, Wyeth W Wasserman, and Boris Lenhard. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic acids research*, 32(suppl 1):D91–D94, 2004.

[4] Edgar Wingender, Peter Dietze, Holger Karas, and Rainer Knüppel. TRANSFAC: a database on transcription factors and their dna binding sites. *Nucleic acids research*, 24(1):238–241, 1996.

[5] Kyungsook Han, Byungkyu Park, Hyongguen Kim, Jinsun Hong, and Jong Park. HPID: The human protein interaction database. *Bioinformatics*, 20(15):2466–2470, 2004.

[6] TS Keshava Prasad, Kumaran Kandasamy, and Akhilesh Pandey. Human protein reference database and human proteinpedia as discovery tools for systems biology. *Reverse Chemical Genetics: Methods and Protocols*, pages 67–79, 2009.

[7] Chris Stark, Bobby-Joe Breitkreutz, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, and Mike Tyers. BioGRID: a general repository for interaction datasets. *Nucleic acids research*, 34(suppl 1):D535–D539, 2006.

[8] Ioannis Xenarios, Danny W Rice, Lukasz Salwinski, Marisa K Baron, Edward M Marcotte, and David Eisenberg. DIP: the database of interacting proteins. *Nucleic acids research*, 28(1):289–291, 2000.

[9] Peter E Hodges, Andrew HZ McKee, Brian P Davis, William E Payne, and James I Garrels. The yeast proteome database (YPD): a model for the organization and presentation of genome-wide functional data. *Nucleic Acids Research*, 27(1):69–73, 1999.

[10] Jingkai Yu, Svetlana Pacifico, Guozhen Liu, and Russell L Finley. DroID: the drosophila interactions database, a comprehensive resource for annotated gene and protein interactions. *BMC genomics*, 9(1):461, 2008.

[11] Mathias Krull, Nico Voss, Claudia Choi, Susanne Pistor, Anatolij Potapov, and Edgar Wingender. Transpath®: an integrated database on signal transduction and a tool for array analysis. *Nucleic acids research*, 31(1):97–100, 2003.

[12] Luke E Ulrich and Igor B Zhulin. MiST: a microbial signal transduction database. *Nucleic acids research*, 35(suppl 1):D386–D390, 2007.

[13] Peter D Karp, Christos A Ouzounis, Caroline Moore-Kochlacs, Leon Goldovsky, Pallavi Kaipa, Dag Ahrén, Sophia Tsoka, Nikos Darzentas, Victor Kunin, and Núria López-Bigas. Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic acids research*, 33(19):6083–6089, 2005.

[14] Minoru Kanehisa, Susumu Goto, Miho Furumichi, Mao Tanabe, and Mika Hirakawa. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic acids research*, 38(suppl 1):D355–D360, 2010.

[15] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47–97, 2002.

[16] Mark EJ Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.

[17] Eric Alm and Adam P Arkin. Biological networks. *Current opinion in structural biology*, 13(2):193–202, 2003.

[18] Baruch Barzel, Amitabh Sharma, and AL Bababási. Graph theory properties of cellular networks. *Handbook of systems biology: concepts and insights, 1st edn. Academic Press, Cambridge*, pages 177–193, 2012.

[19] Wolfgang Huber, Vincent J Carey, Li Long, Seth Falcon, and Robert Gentleman. Graphs in molecular biology. *BMC bioinformatics*, 8(Suppl 6):S8, 2007.

[20] Victor Spirin and Leonid A Mirny. Protein complexes and functional modules in molecular networks. *Proceedings of the National Academy of Sciences*, 100(21):12123–12128, 2003.

[21] Soumya Jyoti Banerjee, Saptarshi Sinha, and Soumen Roy. Slow poisoning and destruction of networks: Edge proximity and its implications for biological and infrastructure networks. *Physical Review E*, 91(2):022807, 2015.

[22] Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, 12(1):56–68, 2011.

[23] Marc Vidal, Michael E Cusick, and Albert-László Barabási. Interactome networks and human disease. *Cell*, 144(6):986–998, 2011.

[24] Gabriel Östlund, Mats Lindskog, and Erik LL Sonnhammer. Network-based identification of novel cancer genes. *Molecular & Cellular Proteomics*, 9(4):648–655, 2010.

[25] Emre Guney, Jörg Menche, Marc Vidal, and Albert-László Barábasi. Network-based in silico drug efficacy screening. *Nature communications*, 7:10331, 2016.

[26] Jörg Menche, Amitabh Sharma, Maksim Kitsak, Susan Dina Ghiassian, Marc Vidal, Joseph Loscalzo, and Albert-Lászlo Barabási. Uncovering disease-disease relationships through the incomplete interactome. *Science*, 347(6224):1257601, 2015.

[27] Rudolf Emil Kalman. Mathematical description of linear dynamical systems. *Journal of the Society for Industrial and Applied Mathematics, Series A: Control*, 1(2):152–192, 1963.

[28] Jean-Jacques E Slotine and Weiping Li. *Applied nonlinear control*, volume 199. Prentice hall Englewood Cliffs, NJ, 1991.

[29] Ching Tai Lin. Structural controllability. *IEEE Transactions on Automatic Control*, 19(3):201–208, 1974.

[30] Yang-Yu Liu, Jean-Jacques Slotine, and Albert-László Barabási. Controllability of complex networks. *Nature*, 473(7346):167–173, 2011.

[31] Jose C Nacher and Tatsuya Akutsu. Dominating scale-free networks with variable scaling exponent: heterogeneous networks are not difficult to control. *New Journal of Physics*, 14(7):073005, 2012.

[32] Tamás Nepusz and Tamás Vicsek. Controlling edge dynamics in complex networks. *Nature Physics*, 8:568–573, 2012.

[33] Marie E Csete and John C Doyle. Reverse engineering of biological complexity. *science*, 295(5560):1664–1669, 2002.

[34] Shmoolik Mangan and Uri Alon. Structure and function of the feedforward loop network motif. *Proceedings of the National Academy of Sciences*, 100(21):11980–11985, 2003.

[35] Jarred M Callura, Charles R Cantor, and James J Collins. Genetic switchboard for synthetic biology applications. *Proceedings of the National Academy of Sciences*, 109(15):5850–5855, 2012.

[36] Tao Jia, Yang-Yu Liu, Endre Csóka, Márton Pósfai, Jean-Jacques Slotine, and Albert-László Barabási. Emergence of bimodality in controlling complex networks. *Nature Communications*, 4:2002, 2013.

[37] Soumya Jyoti Banerjee and Soumen Roy. Key to network controllability. *arXiv preprint arXiv:1209.3737*, 2012.

[38] Yang-Yu Liu, Jean-Jacques Slotine, and Albert-László Barabási. Control centrality and hierarchical structure in complex networks. *Plos one*, 7(9):e44459, 2012.

[39] Xiao-Fei Zhang, Le Ou-Yang, Yuan Zhu, Meng-Yun Wu, and Dao-Qing Dai. Determining minimum set of driver nodes in protein-protein interaction networks. *BMC Bioinformatics*, 16(1):146, 2015.

[40] Alexander J Gates and Luis M Rocha. Control of complex networks requires both structure and dynamics. *Scientific Reports*, 6:24456, 2016.

[41] Jorge Gomez Tejeda Zañudo, Gang Yang, and Réka Albert. Structure-based control of complex networks with nonlinear dynamics. *Proceedings of the National Academy of Sciences*, 114(28):7234–7239, 2017.

[42] Stefan Wuchty. Controllability in protein interaction networks. *Proceedings of the National Academy of Sciences USA*, 111(19):7156–7160, 2014.

[43] Arunachalam Vinayagam, Travis E Gibson, Ho-Joon Lee, Bahar Yilmazel, Charles Roesel, Yanhui Hu, Young Kwon, Amitabh Sharma, Yang-Yu Liu, Norbert Perrimon, et al. Controllability analysis of the directed human protein interaction network identifies disease genes and drug targets. *Proceedings of the National Academy of Science USA*, 113(18):4976–4981, 2016.

[44] Xueming Liu and Linqiang Pan. Identifying driver nodes in the human signaling network using structural controllability analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 12(2):467–472, 2015.

[45] Marina Uhart, Gabriel Flores, and Diego M Bustos. Controllability of protein-protein interaction phosphorylation-based networks: Participation of the hub 14-3-3 protein family. *Scientific Reports*, 6:26234, 2016.

[46] Haruna Kagami, Tatsuya Akutsu, Shingo Maegawa, Hiroshi Hosokawa, and Jose C Nacher. Determining associations between human diseases and non-coding rnas with critical roles in network control. *Scientific Reports*, 5:14577, 2015.

[47] Vandana Ravindran, V Sunitha, and Ganesh Bagler. Identification of critical regulatory genes in cancer signaling network using controllability analysis. *Physica A: Statistical Mechanics and its Applications*, 474:134–143, 2017.

[48] V. Ravindran, V. Sunitha, and G. Bagler. Controllability of human cancer signaling network. In *2016 International Conference on Signal Processing and Communication (ICSC)*, pages 363–367, Dec 2016.

[49] Matthew D Dyer, TM Murali, and Bruno W Sobral. The landscape of human proteins interacting with viruses and other pathogens. *PLoS Pathog*, 4(2):e32, 2008.

[50] Justin Ruths and Derek Ruths. Control profiles of complex networks. *Science*, 343(6177):1373–1376, 2014.

[51] Yang-Yu Liu and Albert-László Barabási. Control principles of complex systems. *Reviews of Modern Physics*, 88:035006, 2016.

[52] Takeo Yamada and Leslie R Foulds. A graph-theoretic approach to investigate structural and qualitative properties of systems: A survey. *Networks*, 20(4):427–452, 1990.

[53] Christian Commault, Jean-Michel Dion, and Jacob W van der Woude. Characterization of generic properties of linear structured systems for efficient computations. *Kybernetika*, 38(5):503–520, 2002.

[54] Douglas Brent West et al. *Introduction to graph theory*, volume 2. Prentice hall Upper Saddle River, 2001.

[55] John E Hopkroft and Richard M Karp. An n5= 2 algorithm for maximum matching in bipartite graphs. *SIAM J. Comput*, 2:225–231, 1973.

[56] Jose C Nacher and Tatsuya Akutsu. Structurally robust control of complex networks. *Physical Review E*, 91(1):012826, 2015.

[57] Douglas Hanahan and Robert A Weinberg. The hallmarks of cancer. *Cell*, 100(1):57–70, 2000.

[58] Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *Cell*, 144(5):646–674, 2011.

[59] Tony Pawson and Rune Linding. Network medicine. *FEBS Letters*, 582(8):1266–1270, 2008.

[60] Rexxi D Prasasya, Dan Tian, and Pamela K Kreeger. Analysis of cancer signaling networks by systems biology to develop therapies. *Seminars in Cancer Biology*, 21(3):200–206, 2011.

[61] Janine T Erler and Rune Linding. Network medicine strikes a blow against breast cancer. *Cell*, 149(4):731–733, 2012.

[62] Evangelia Koutsogiannouli, Athanasios G Papavassiliou, and Nikolaos A Papanikolaou. Complexity in cancer biology: is systems biology the answer? *Cancer Medicine*, 2(2):164–177, 2013.

[63] MEJ Newman. *Networks: An introduction*. Oxford University Press, Oxford, 2010.

[64] Sergei N Dorogovtsev and José FF Mendes. *Evolution of networks: From biological nets to the Internet and WWW*. Oxford University Press, Oxford, 2013.

[65] Minoru Kanehisa, Susumu Goto, Yoko Sato, Miho Furumichi, and Mao Tanabe. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research*, 40(D1):D109–D114, 2011.

[66] Simon A Forbes, David Beare, Prasad Gunasekaran, Kenric Leung, Nidhi Bindal, Harry Boutselakis, Minjie Ding, Sally Bamford, Charlotte Cole, Sari

Ward, et al. Cosmic: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Research*, 43(D1):D805–D811, 2015.

[67] Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. The cancer genome atlas (TCGA): an immeasurable source of knowledge. *Contemporary Oncology*, 19(1A):A68–A77, 2015.

[68] Vinay Randhawa and Ganesh Bagler. Identification of src as a potent drug target for asthma, using an integrative approach of protein interactome analysis and in silico drug discovery. *OMICS: A Journal of Integrative Biology*, 16(10):513–526, 2012.

[69] Shikha Vashisht and Ganesh Bagler. An approach for the identification of targets specific to bone metastasis using cancer genes interactome and gene ontology analysis. *PLoS ONE*, 7(11):e49401, 2012.

[70] Vinay Randhawa, Purnima Sharma, Shashi Bhushan, and Ganesh Bagler. Identification of key nodes of type 2 diabetes mellitus protein interactome and study of their interactions with phloridzin. *OMICS: A Journal of Integrative Biology*, 17(6):302–317, 2013.

[71] Anna Lombardi and Michael Hörnquist. Controllability analysis of networks. *Physical Review E*, 75(5):1–5, 2007.

[72] Pamela K Kreeger and Douglas A Lauffenburger. Cancer systems biology: a network modeling perspective. *Carcinogenesis*, 31(1):2–8, 2010.

[73] Indika Rajapakse, Mark Groudine, and Mehran Mesbahi. What can systems theory of networks offer to biology? *PLoS Computational Biology*, 8(6):e1002543, 2012.

[74] Ting-Ting Zhou. Network systems biology for targeted cancer therapies. *Chinese Journal of Cancer*, 31(3):134–141, 2012.

[75] Hiroaki Kitano. Cancer as a robust system: implications for anticancer therapy. *Nature Reviews Cancer*, 4(3):227–235, 2004.

[76] Florian Gnad, Sophia Doll, Gerard Manning, David Arnott, and Zemin Zhang. Bioinformatics analysis of thousands of TCGA tumors to determine the involvement of epigenetic regulators in human cancer. *BMC Genomics*, 16(8):1–15, 2015.

[77] Qinghua Cui, Yun Ma, Maria Jaramillo, Hamza Bari, Arif Awan, Song Yang, Simo Zhang, Lixue Liu, Meng Lu, Maureen O'Connor-McCourt, et al. A map of human cancer signaling. *Molecular Systems Biology*, 3(1):1–13, 2007.

[78] Jing Wang, Dexter Duncan, Zhiao Shi, and Bing Zhang. Web-based gene set analysis toolkit (WebGestalt): Update 2013. *Nucleic Acids Research*, 41(W1):W77–W83, 2013.

[79] Hao Luo, Yan Lin, Feng Gao, Chun-Ting Zhang, and Ren Zhang. DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements. *Nucleic Acids Research*, 42(D1):D574–D580, 2014.

[80] Malachi Griffith, Obi L Griffith, Adam C Coffman, James V Weible, Josh F McMichael, Nicholas C Spies, James Koval, Indraniel Das, Matthew B Callaway, James M Eldred, et al. DGIdb: mining the druggable genome. *Nature Methods*, 10(12):1209–1210, 2013.

[81] Lu-Lu Zheng, Chunyan Li, Jie Ping, Yanhong Zhou, Yixue Li, and Pei Hao. The domain landscape of virus-host interactomes. *BioMed research international*, 2014:867235, 2014.

[82] Andrew Chatr-aryamontri, Arnaud Ceol, Daniele Peluso, Aurelio Nardozza, Simona Panni, Francesca Sacco, Michele Tinti, Alex Smolyar, Luisa Castagnoli, Marc Vidal, Michael E. Cusick, and Gianni Cesareni. VirusMINT: a viral protein interaction database. *Nucleic Acids Research*, 37(suppl 1):D669–D673, 2009.

[83] Thibaut Guirimand, Stéphane Delmotte, and Vincent Navratil. Virhostnet 2.0: surfing on the web of virus/host molecular interactions data. *Nucleic acids research*, 43(D1):D583–D587, 2014.

[84] Jamie I MacPherson, John W Pinney, and David L Robertson. Jnets: exploring networks by integrating annotation. *BMC bioinformatics*, 10(1):95, 2009.

[85] Roger G Ptak, William Fu, Brigitte E Sanders-Beer, Jonathan E Dickerson, John W Pinney, David L Robertson, Mikhail N Rozanov, Kenneth S Katz, Donna R Maglott, Kim D Pruitt, et al. Short communication: Cataloguing the hiv type 1 human protein interaction network. *AIDS research and human retroviruses*, 24(12):1497–1502, 2008.

[86] Jamie I MacPherson, Jonathan E Dickerson, John W Pinney, and David L Robertson. Patterns of hiv-1 protein interaction identify perturbed host-cellular subsystems. *PLoS Comput Biol*, 6(7):e1000863, 2010.

[87] Jason E McDermott, Ronald C Taylor, Hyunjin Yoon, and Fred Heffron. Bottlenecks and hubs in inferred networks are important for virulence in salmonella typhimurium. *Journal of Computational Biology*, 16(2):169–180, 2009.

[88] Saliha Durmuş Tekir, Tunahan Çakir, and Kutlu Ö Ülgen. Infection strategies of bacterial and viral pathogens through pathogen–human protein–protein interactions. *Frontiers in microbiology*, 3:46, 2012.

[89] Yazdan Asgari, Ali Salehzadeh-Yazdi, Falk Schreiber, and Ali Masoudi-Nejad. Controllability in cancer metabolic networks according to drug targets as driver nodes. *PloS one*, 8(11):e79397, 2013.

[90] Peng Gang Sun. Co-controllability of drug-disease-gene network. *New Journal of Physics*, 17(8):085009, 2015.

[91] Xiao-Fei Zhang, Le Ou-Yang, Dao-Qing Dai, Meng-Yun Wu, Yuan Zhu, and Hong Yan. Comparative analysis of housekeeping and tissue-specific driver nodes in human protein interaction networks. *BMC bioinformatics*, 17(1):358, 2016.

[92] Arunachalam Vinayagam, Ulrich Stelzl, Raphaele Foulle, Stephanie Plassmann, Martina Zenkner, Jan Timm, Heike E Assmus, Miguel A Andrade-

Navarro, and Erich E Wanker. A directed protein interaction network for investigating intracellular signal transduction. *Sci. Signal.*, 4(189):rs8–rs8, 2011.

[93] Alicia E Smith and Ari Helenius. How viruses enter animal cells. *Science*, 304(5668):237–242, 2004.

[94] Craig B Wilen, John C Tilton, and Robert W Doms. HIV: cell binding and entry. *Cold Spring Harbor perspectives in medicine*, 2(8):a006866, 2012.

[95] Xionglei He and Jianzhi Zhang. Why do hubs tend to be essential in protein networks? *PLoS genetics*, 2(6):e88, 2006.

[96] Edward A Berger, Philip M Murphy, and Joshua M Farber. Chemokine receptors as HIV-1 coreceptors: roles in viral entry, tropism, and disease. *Annual review of immunology*, 17(1):657–700, 1999.

[97] Oyebode J Oyeyemi, Oluwafemi Davies, David L Robertson, and Jean-Marc Schwartz. A logical model of HIV-1 interactions with the T-cell activation signalling pathway. *Bioinformatics*, 31(7):1075–1083, 2014.

[98] Eric J Arts and Daria J Hazuda. HIV-1 antiretroviral drug therapy. *Cold Spring Harbor perspectives in medicine*, 2(4):a007161, 2012.

[99] Keduo Qian, Susan L Morris-Natschke, and Kuo-Hsiung Lee. HIV entry inhibitors and their potential in hiv therapy. *Medicinal research reviews*, 29(2):369–393, 2009.

[100] Dean W Felsher. Reversibility of oncogene-induced cancer. *Current opinion in genetics & development*, 14(1):37–42, 2004.

[101] Kwang-Hyun Cho, Jae Il Joo, Dongkwan Shin, Dongsan Kim, and Sang-Min Park. The reverse control of irreversible biological processes. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 8(5):366–377, 2016.

[102] Xin-Dong Gao, Wen-Xu Wang, and Ying-Cheng Lai. Control efficacy of complex networks. *Scientific Reports*, 6, 2016.

[103] Wesley K Kroeze, Douglas J Sheffler, and Bryan L Roth. G-protein-coupled receptors at a glance. *Journal of cell science*, 116(24):4867–4869, 2003.

[104] Michael Williams and Rita Raddatz. *Receptors as Drug Targets*. John Wiley and Sons, Inc., 2001.

[105] Jose C Nacher and Tatsuya Akutsu. Minimum dominating set-based methods for analyzing biological networks. *Methods*, 102:57–63, 2016.

[106] Jose C Nacher and Tatsuya Akutsu. Analysis of critical and redundant nodes in controlling directed and undirected complex networks using dominating sets. *Journal of Complex Networks*, 2(4):394–412, 2014.

[107] Rui Li, Meng Yang, and Tianguang Chu. Controllability and observability of boolean networks arising from biology. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 25(2):023104, 2015.

Appendix 1: Supplementary information for human cancer signaling network

Data curation: Removing ambiguities in gene names For six genes the Entrez ID were updated and 3 genes had animal orthologs and were replaced with it human orthologous genes. We also deleted duplicated gene that had same approved name but were reported by different symbols. For such genes if they had the same interaction we removed the duplicated one and added one if it was a new interaction and retained any one gene among them. Out of 1186 genes from the original network we wanted to query for HUGO symbols. We queried the HGCN database. Initially 1131 approved symbols were obtained. For remaining 55 manual curation was done by looking up the Entrez ID at NCBI and second by gene symbol using multi-symbol checker http://www.genenames.org/cgi-bin/symbol_checker. For certain genes the Entrez ID and gene symbol did not match, in which case we retained the gene symbol and updated the corresponding Entrez

Gene name	Original ID	Remark
AA	8	Updated ID 7003
MEF2B	4207	Updated ID 100
PKC	50818	Updated ID 11
p65	55566	This Id is withdrawn but the gene symbol matches 5970, RELA.
PDE	64414	Updated ID 50
Actin	280979	This is of Bos Taurus, its human ortholog is ACTB with ID 60
CP1	50820	Replaced with 23314 by NCBI
MUNC-18	282378	This is of Bos Taurus, its human ortholog is STXBP1 with ID 6812
P35611	27360	This is of mouse, its human ortholog is ADD3 with ID 12

Table A1: List of genes that had different/ Updated Entrez ID.

ID. For three genes the ID were not of human gene and hence its corresponding human ortholog were considered.

Statistical tests

Following are the statistical analysis implemented in this work.

1. Chi-square test :The chi-square is used to determine if there is significant difference between expected and observed data in one or more categories. Observed data is denoted as O_i,where $i = 1, 2, ...N$ where N is number of categories. Expected data is denoted as E_i and is calculated by, $E_i = p_i * \sum_{i=1}^{N} O_i$, were p_i is expected percentage. The chi-square formula is defined as follows $\chi^2 = \sum_{i=1}^{N} (O_i - E_i)^2 / E_i$. The degree of freedom is calculated as $N - 1$, if the number of categories is N and the corresponding value for that particular degree of freedom with a predetermined level of significance is looked from the chi-square table. If the value obtained from the table is less than the calculated chi-square value then the difference between the percentages among different sets is not due to chance.

2. Z-score statistics: Z-score was calculated to find significant over representation of genes among each category than that by chance alone. Random sampling of genes was done based on the size of observed number of genes in each category for 1000 instances. The mean and standard deviation was calculated for each instance and then the Z-score was calculated based on the following formula. $Z - score = O_f - E_f / (S.D)$ where O_f is observed frequency, E_f is expected frequency and $S.D$ is standard deviation.

Statistical test results

Cancer associated							
Category	O	p	E	$(O-E)^2/E$	χ^2	$0.05sig$	$0.01sig$
PDN	119	29.03	127.74	0.6	11.72	5.99	9.21
ODN	115	31.83	140.07	4.49			
BDN	206	39.13	172.19	6.64			
Non- cancer associated							
Category	O	p	E	$(O-E)^2/E$	χ^2	$0.05sig$	$0.01sig$
PDN	223	29.03	214.26	0.36	6.99	5.99	9.21
ODN	260	31.83	234.9321	2.67			
BDN	255	39.13	288.8098	3.96			

Table A2: The statistics of association of cancer among driver nodes.

Chi-square for indispensable BDN							
Category	O	p	E	$(O-E)^2/E$	χ^2	$0.05sig$	$0.01sig$
CA	110	44.69	88.03	5.48	9.91	3.84	6.63
Non-CA	87	55.31	108.97	4.43			
Chi-square for neutral BDN							
Category	O	p	E	$(O-E)^2/E$	χ^2	$0.05sig$	$0.01sig$
CA	96	44.69	117.97	4.09	7.4	3.84	6.63
Non-CA	168	55.31	146.03	3.31			

Table A3: The statistics of association of cancer among indispensable and neutral BDN.

<table>
<tr><td colspan="8">Chi-square test for All ANPDs</td></tr>
<tr><td><b>Category</b></td><td>O</td><td>p</td><td>E</td><td>$(O-E)^2/E$</td><td>χ^2</td><td>$0.05sig$</td><td>$0.01sig$</td></tr>
<tr><td>Indispensable</td><td>190</td><td>42.73</td><td>126.92</td><td>31.35</td><td>54.75</td><td>3.84</td><td>6.63</td></tr>
<tr><td>Neutral</td><td>107</td><td>57.27</td><td>170.08</td><td>23.4</td><td></td><td></td><td></td></tr>
<tr><td colspan="8">Chi-square test for unique ANPDs among indispensable and neutral</td></tr>
<tr><td><b>Category</b></td><td>O</td><td>p</td><td>E</td><td>$(O-E)^2/E$</td><td>χ^2</td><td>$0.05sig$</td><td>$0.01sig$</td></tr>
<tr><td>Indispensable</td><td>129</td><td>42.73</td><td>74.78</td><td>39.31</td><td>68.64</td><td>3.84</td><td>6.63</td></tr>
<tr><td>Neutral</td><td>46</td><td>57.27</td><td>100.22</td><td>29.33</td><td></td><td></td><td></td></tr>
</table>

Table A4: Statistics for total and unique number of ANPDs targeting indispensable and neutral BDNs.

Appendix 2: Supplementary information for HIV-human molecular interactome

Complex network parameter which determine HIV-human molecular interactome

Holistic view of the complex host-pathogen interactome is necessary for the understanding of complex diseases. With the plethora of omics data available it is now possible to study these complex systems at network level. In this study, we explore different complex network parameters for the HIV-1 human protein interaction network with an aim to understand how these biological networks interact with each other. Some questions that we hope to answer through this study are: What class or module of proteins does HIV interact with? How can we identify these proteins through network parameters?

Methods

Network construction We used HPRD (Human Protein Reference Database), one of the most comprehensive resources of human protein-protein interactions (PPIs), to construct the human interactome [6]. The database is manually curated, contains the maximum number of binary non-redundant human PPIs, gene annotations and largest citations of PPIs curated. This human interactome comprises 9587 proteins and 39240 interactions. The data for HIV-1 human interaction network was complied from BioGRID. BioGRID is a general repository for interaction datasets. The database is complied through comprehensive curation efforts and captures data from published experimental results. The HIV-1 human interactome has 1020 proteins with 1811 interactions (Figure A1, A2).

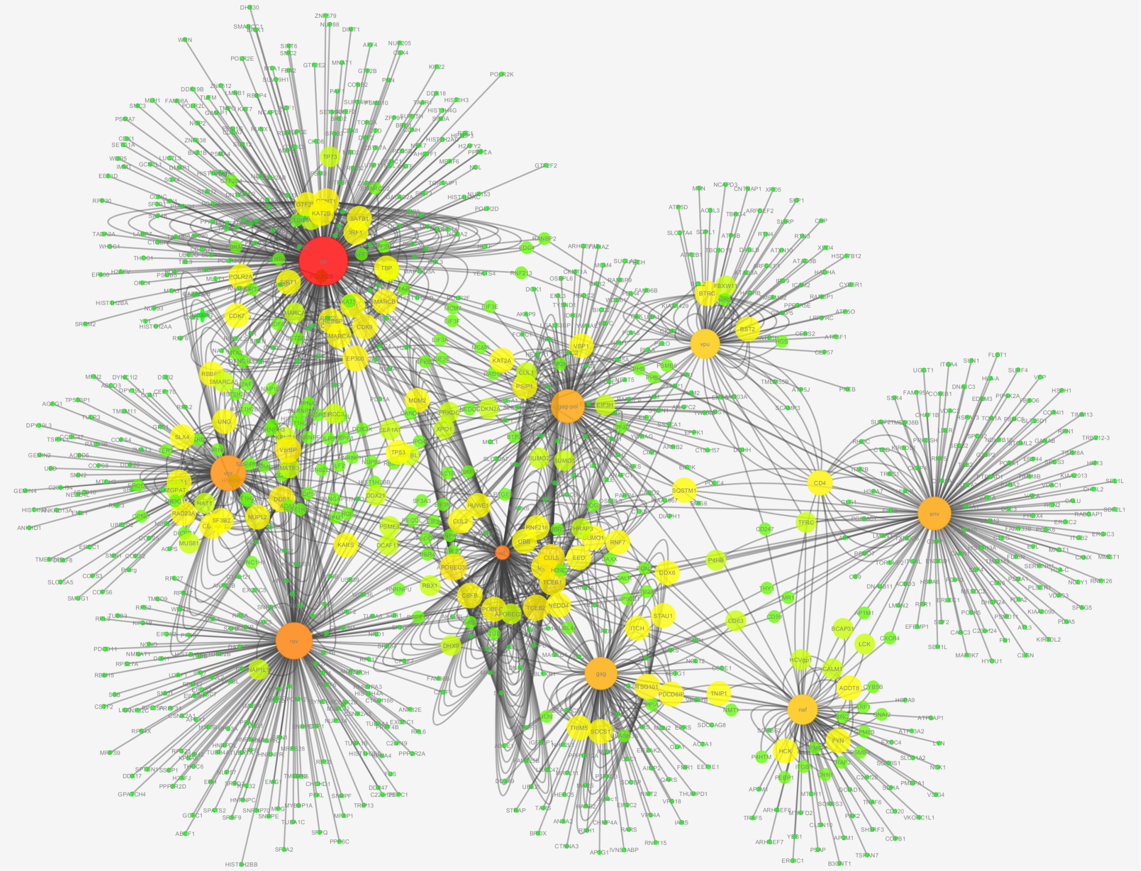


Figure A1: HIV-1 Human Protein interactome.Size of node corresponds to the degree.

Interactome analysis We performed network analysis of the HPRD and HIV-1 Human interactome to compute various graph theoretic parameters. We computed network centrality measures like degree, betweenness, stress, average shortest path length, clustering coefficient and topological coefficient. Identification of Hub nodes: We ranked the proteins based on the degree in the human interactome. Then we compiled them into ten hub sets (top 100-top 1000) containing the proteins with ranks above certain cutoffs.

Results and Conclusion

We identified that 812 proteins out of 1028 were targeted by HIV-1 from the HPRD(Figure A3). The hub sets classified top 100, top 200, top 300 had 39, 61, 88 hub proteins respectively targeted by HIV. Similarly we classified sets based on the betweenness centrality and had 32, 62, 85 bottleneck proteins in top 100, 200, 300 respectively targeted by HIV-1. We randomly sampled a corresponding

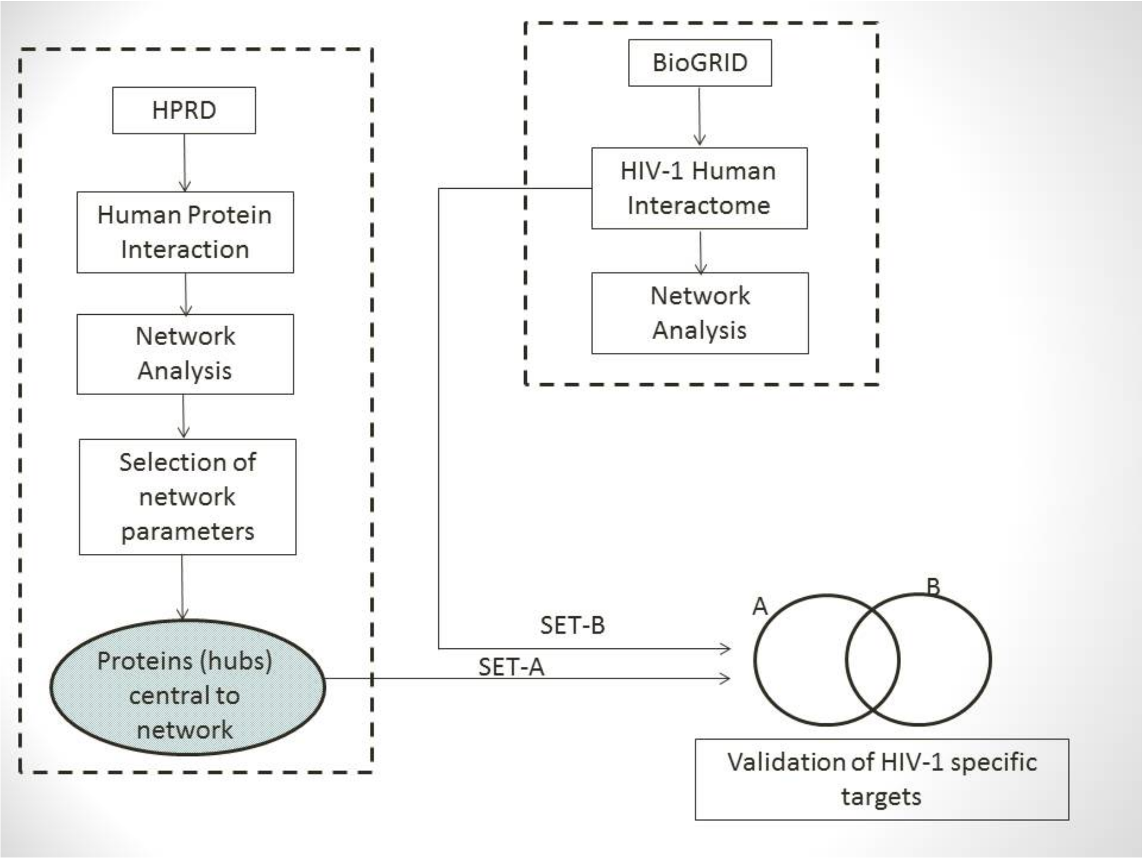

HPRD
Human Protein Interaction
Network Analysis
Selection of network parameters
Proteins (hubs) central to network
BioGRID
HIV-1 Human Interactome
Network Analysis
SET-B
SET-A
A
B
Validation of HIV-1 specific targets

Figure A2: Strategy implemented for identification of HIV-1 specific targets.

number of proteins from the human interactome for 100 instances and found that around 8% as expected are found to be targeted by the virus (Figure A4).

In this study we provide an overview of the human proteins interaction with HIV-1 and demonstrate that the viruses target hubs (nodes/proteins that interact with many other proteins). We also demonstrated the results for bottleneck (proteins that lie on many shortest paths) and obtained the same result that HIV-1 targets hub and bottleneck proteins. Through this study we identify certain network parameters through graph theoretic metrics that could be useful in identifying certain class of proteins and help identifying drug targets.

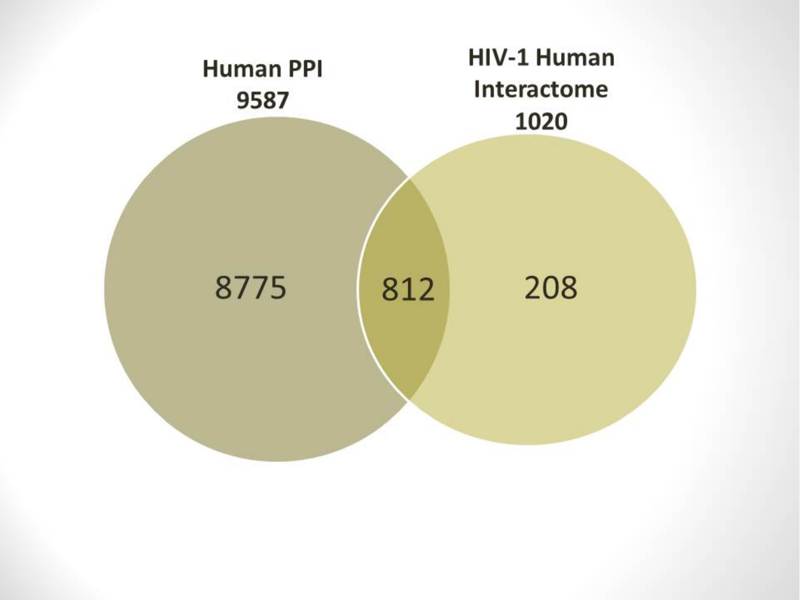
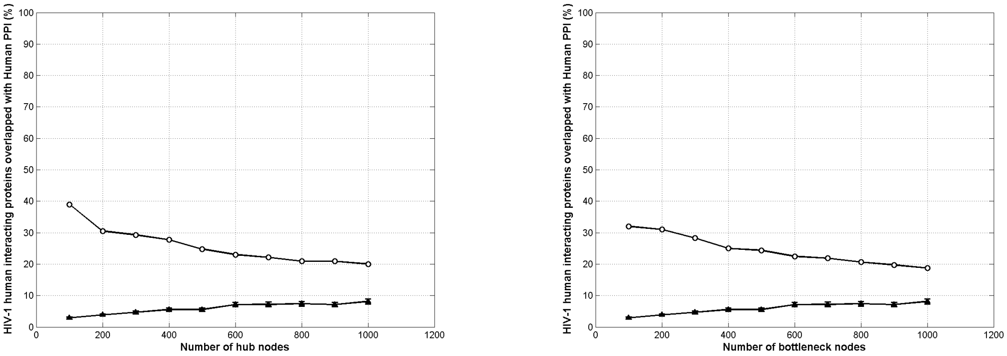


Figure A3: Venn diagram depicting the number of proteins targeted by HIV-1 from the human interactome.



(a) Hubs of human interactome targeted by HIV-1.

(b) bottlenecks of human interactome targeted by HIV-1.

Figure A4: Network parameters central to HIV-1 human molecular interactome.

Classification of edges based on controllability

source	target	Reaction type
BAD	BCL2L1	inhibits
GRAP2(GADS)	LCP2(SLP76)	activates
Cam	CaN	activates
GAB2	PTPN11(SHP2)	activates
GRB2	GAB2	activates
IP3	Ca2+	activates
MAPK8(JNK)	JUN	activates
MAPK14(p38)	PXN	activates
PAG	CSK	activates
RAF	MAP2K1/2(MEK1/2)	activates
RAS	RAF	activates
RPS6KA1(RSK)	CREB	activates
RASGRP1	SOS	activates
MAP3K7(TAK1)	MAP2K7(MKK7)	activates
PDCD1(PD-1)	PTPN6(SHP1)	activates
PPP2R4(PP2A)	AKT	inhibits
MAP3K8(COT)	MAP3K14(NIK)	activates
SHC	GRB2	activates
VAV1	RHOA	activates
Chemokine	CCR5/CXCR4	activates
GRKL	GRK	activates
CCR5/CXCR4	JAK3	activates
Gai	AC	inhibits
AC	cAMP	activates
cAMP	PKA	activates
Gby	PLCB	activates
PKA	P-REX1	inhibits
PTK2B(Pyk2)	Src	activates
NCF1	NOX4	activates
JAK3	STAT	activates
RHOA	ROCK	activates
RAP	CDC42	activates
RASGRP2	RAP	activates
CD4	PTK2	activates
CBLB	PTEN	activates

Table A5: Critical links in T-cell activation network.

source	target	Reaction type
BAD	BCL2L1	inhibits
GRAP2(GADS)	LCP2(SLP76)	activates
GSK3	CYC1	inhibits
GADD45	MAPK14(p38)	inhibits
Cam	CaN	activates
GAB2	PTPN11(SHP2)	activates
GRB2	GAB2	activates
IP3	Ca2+	activates
MAPK8(JNK)	JUN	activates
MAPK14(p38)	PXN	activates
PAG	CSK	activates
PIP2	PIP3	activates
RAF	MAP2K1/2(MEK1/2)	activates
RAS	RAF	activates
RPS6KA1(RSK)	CREB	activates
RASGRP1	SOS	activates
MAP3K7(TAK1)	MAP2K7(MKK7)	activates
PDCD1(PD-1)	PTPN6(SHP1)	activates
PPP2R4(PP2A)	AKT	inhibits
MAP3K8(COT)	MAP3K14(NIK)	activates
SHC	GRB2	activates
VAV1	RHOA	activates
Chemokine	CCR5/CXCR4	activates
GRKL	GRK	activates
CCR5/CXCR4	JAK3	activates
Gai	AC	inhibits
cAMP	PKA	activates
PTK2B(Pyk2)	Src	activates
NCF1	NOX4	activates
JAK3	STAT	activates
RHOA	ROCK	activates
RAP	CDC42	activates
RASGRP2	RAP	activates
CD4	PTK2	activates
PI3K	BAD	activates
CBLB	PTEN	activates
protease	B-arrestin	activates
STAT	capsid	activates
PAK	matrix	activates
vif	CD4	activates
CDKN1A(p21)	in	inhibits
vpu	CTNNB1(bcat)	inhibits
CD45	gp120	activates

Table A6: Critical links in HIV T-cell activation network.

	MDS model			MM model		
	Uninfected	**Infected**		**Uninfected**	**Infected**	
		HIV	**Human**		**HIV**	**Human**
Driver nodes	1398	1232		2283	2264	
Critical	874	12	688	377	1	266
Ordinary	1250	5	1231	3330	2	3443
Redundant	4215	5	4420	2632	19	2630

Table A7: Classification of minimum dominating set (MDS) and maximum matching driver nodes.

Identification of driver nodes using minimum dominating set method

To further gain insights in the hijack mechanism we identified driver nodes in the uninfected and HIV-1 infected network using the minimum dominating set model (MDS) [31]. For a directed graph, a set of nodes S is said to be a dominating set if for any node $v \in V, v \in S$ holds or there is a node $u \in S$ such that there exists a directed edge (u, v). A minimum dominating set (MDS) is a dominating set with the minimum number of nodes. For a given graph the MDS need not be unique. Each MDS forms the driver node set. Because there are multiple MDS configurations, nodes are classified as critical driver node, if a node is always present in all MDS, occasionally present in MDS then it is an ordinary driver node and if a node is never part of any MDS then it is a redundant/non-driver node [106]. After classifying the driver nodes based on MDS model, we compared it with the maximum matching (MM) model. Since MDS computation is NP-hard we should consider the network individually to find an MDS in it or we should use approximation algorithms to compute an MDS in an arbitrary network.

The analysis classified 1398 (22%) nodes as driver nodes based on MDS method, compared to 2282 (36%) of them classified by the maximum matching in the uninfected network. In the HIV-1 infected network, MDS classified 19% of the nodes as driver nodes compared to 36% of them classified by the MM model. This captures the ease with which the network can be controlled and that MDS yielded fewer driver nodes compared to maximum matching. Since the driver node set is not unique for both the models, driver nodes are classified further as critical, ordinary and redundant. In the uninfected network,the MDS classified 14% of the nodes as critical compared to only 6% of the nodes classified as critical by the MM model. Similarly, the MDS classified more nodes as redundant compared to MM. On the other hand MM model classified most of the nodes as ordinary (53%) compared to MDS (20%) (Table A7). This explains why fewer overall driver nodes are required to control the network with MDS. A similar trend in the classification of driver nodes is observed in the infected network. As was observed in Chapter 4 for the MM model, we see that in the MDS model also there is a decrease in the number of critical driver nodes, while the number of ordinary and redundant nodes remain almost the same for the infected and uninfected states. These results indicate that inclusion of the virus has increased the controllability of the network, i.e., the virus set of interactions increases the number of interactions facilitating the hijacking of the cell, and controlling the network more efficiently.

We also looked at the association of the virus proteins with the driver nodes to identify if they are preferentially targeted by the virus. Out of the 6339 proteins in the uninfected human PPI network, 2529 nodes have been reported to be targeted by HIV-1. Of the different driver node classified by the MDS model, 50% of the critical driver nodes were significantly targeted by HIV-1 (Z-score=6.50) (Table A8). The critical driver nodes identified by MDS are mostly high degree nodes. In terms of understanding infection, the virus is mainly driving the network by targeting the critical driver nodes.

Node type	Observed	Percentage	Random mean	Z-score	P-value
Critical	438	50.11	349.38	6.5	8.03E-011
Ordinary	489	39.12	498.4	-0.61	0.542
Redundant	1602	38.01	1681.45	-4.41	1.03E-005

Table A8: HIV-1 targets among driver nodes. Numbers of observed critical, intermittent or redundant nodes.

We also analysed both the networks for robustness to control based on MDS model and, classified each node into (1) indispensable, i.e., if we have to control more driver nodes in its absence; (2) dispensable, if we have to control fewer driver nodes and (3) neutral, if in its absence there is no change in the number of driver nodes. We compared this classification with that obtained using MM model in Chapter 4.

Interestingly, MDS classified a much smaller number of nodes as indispensable (503) or dispensable (770) compared to MM (1330 versus 2347 respectively), with twice the number of neutral nodes identified by MDS compared to MM (Table A9). As a consequence MDS performs much more efficiently than MM. A similar trend was seen in the HIV-1 infected network with a smaller number of nodes classified as indispensable (397 human and 11 HIV) or dispensable (719 human and 3 HIV-1) compared to MM (1331 human and 19 HIV-1 proteins verses 2346 human and 1 HIV-1). In addition, on comparing the difference in node characterisation in both states (uninfected and infected), the indispensable nodes reduced by about 20% to 397 for the MDS model but showed no change for the MM model (Table A9).

	MDS model			**MM model**		
	Uninfected	**Infected**		**Uninfected**	**Infected**	
		HIV	**Human**		**HIV**	**Human**
Indispensable	503	11	397	1330	19	1331
Dispensable	770	3	719	2347	1	2346
Neutral	5066	8	5223	2662	2	2662

Table A9: Control robustness analysis: Classification of nodes based on deletion studies between normal/uninfected and HIV-1 infected networks.

Our preliminary analysis on the HIV-1 human molecular interactome using the minimum dominating set model shows that the results obtained by this model better reflects the viral hijack mechanism than the maximum matching model. Further analysis is required to understand the robustness of this model.

Appendix 3: Controllability analysis using switchboard model

Identifying nodes based on switchboard model

Structural controllability has widely focused on nodal dynamics, i.e., identifying driver nodes that when provided with certain input can change the state of a network from any initial state to desired final state. While in many systems, it is interesting to know how the signals flow along the connections and the nodes do something with these signals. For instance, a transcription network, where information flow across the pathways influences genes and proteins or a neuronal network where signals received triggers the activation of certain neurons. In such cases, it is useful to understand the dynamics on the edges. Nepusz and Vicsek [32] proposed a model that studies dynamical process on edges termed switchboard model. The model identifies minimum set of driver nodes that can driver certain edges to maintain structural controllability. Here the nodes acts as a small switchboard- like device mapping the signals of incoming edges to the outgoing edges.

To identify the driver nodes using the switchboard model, a directed graph is converted to its equivalent line graph. A line graph $L(G)$ of a graph G is a graph, where the each vertex in $L(G)$ represents a edge of G and two vertices in $L(G)$ are adjacent if and only if their corresponding edges are incident in G. The minimum input theorem described in Chapter 2 is then applied on the line graph $L(G)$. This gives set of control paths and driven nodes in line digraph or equivalently a set of driven edges in the original graph (G). The set of driven nodes is those nodes/vertices that have at least one out-going driven edge. The maximum matching on $L(G)$ consists of vertex-disjoint open and closed paths (stems and buds), mapping these paths back to G yields edge-disjoint open and closed walks in G. The walks together form a complete edge cover of G. As the first vertex of each stem has to be driven in $L(G)$, the driver nodes in G are those from which the corresponding open edge-disjoint walks originate. Thus we need to find an edge cover that minimizes the number of nodes from which open walks originate in G.

We identified the driver nodes for networks in our study and compared it with Liu *et.al.* nodal dynamics model [107]. The switchboard model was implemented using netctrl package obtained from https://github.com/ntamas/netctrl.

Comparison of driver nodes

Network	**Nodes**	**Edges**	N_{SBD}	N_{liu}
Cancer Signalling	1232	3060	576 (0.468)	584 (0.474)
Directed Human PPI	6339	34813	2269 (0.358)	2283 (0.360)
HIV-human molecular	6361	40625	2083 (0.327)	2265 (0.356)
T-cell activation	121	255	42 (0.347)	35 (0.289)
HIV-T-cell activation	137	367	44 (0.321)	34 (0.248)
E. coli transcription	423	578	98 (0.232)	308 (0.728)

Table A10: Driver node comparison between switchboard and Liu model.

The cancer signalling and directed human PPI network have no difference in the number of driver nodes. The Liu's model identifies few driver nodes for the T-cell activation and HIV-T-cell activation networks. The switchboard model identifies fewer driver nodes in the *E. coli* transcription network compared to Liu model. One reason could be that in the transcription network a particular gene can up-regulate or down-regulate another gene. The driven node can independently influence its subordinates in the switchboard model thus yielding fewer driver nodes compared to linear nodal dynamics. Our analysis are at preliminary stage and further investigation on what it means to control edges in biological networks is to be done.

Appendix 4: Publication

List of publications

Peer-reviewed journal

- Vandana Ravindran, V Sunitha, Ganesh Bagler. Identification of critical regulatory genes in cancer signaling network using controllability analysis. *Physica A-Statistical Mechanics and its Applications* Vol 474,pp.134-143, 2016.

Peer-reviewed conferences

- Vandana Ravindran, V Sunitha, Ganesh Bagler. Investigation of control profile in biological networks. In the proceedings of the 6th International Conference on Complex Networks and Their Applications, pp.69-71, 2017.

- Vandana Ravindran, V Sunitha, Ganesh Bagler. Controllability of human cancer signaling network- Signaling as a paradigm for disease control through drug-gene interactions. In the proceedings of IEEE International Conference on Signal Processing and Communication, pp.363-367, 2016.