Systems Biology from Cell to Society: Transmission dynamics in complex networks with mesoscopic organization

By

Jesan T

PHYS01200704028

Bhabha Atomic Research Centre, Mumbai

A thesis submitted to the

Board of Studies in Physical Sciences

In partial fulfillment of requirements

For the Degree of

DOCTOR OF PHILOSOPHY

of

HOMI BHABHA NATIONAL INSTITUTE

September, 2013

Homi Bhabha National Institute

Recommendations of the Viva Voce Board

As members of the Viva Voce Board, we certify that we have read the dissertation prepared by **Jesan T** entitled " Systems Biology from Cell to Society: Transmission dynamics in complex networks with mesoscopic organization " and recommend that it maybe accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.

_____ Date:

Chair - B. V. R. Tata

_____ Date:

Supervisor/Convener - Sitabhra Sinha

_____ Date:

Member 1 - Gautam Menon

_____ Date:

Member 2 - Sudeshna Sinha

_____ Date:

Member 3 - Y. S. Mayya

_____ Date:

External Examiner - Narendra M. Dixit

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to HBNI.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it may be accepted as fulfilling the dissertation requirement.

Date:

Place: Supervisor

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at Homi Bhabha National Institute (HBNI) and is deposited in the Library to be made available to borrowers under rules of the HBNI.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgement of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the Competent Authority of HBNI when in his or her judgement the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

Jesan T

DECLARATION

I, hereby declare that the investigation presented in the thesis has been carried out by me. The work is original and has not been submitted earlier as a whole or in part for a degree / diploma at this or any other Institution / University.

Jesan T

*"The righteous man walks in his integrity; His children are blessed after him." (***Proverbs** *20:7)*

– Holy Bible.

Dedicated to the memory of my Father, B. Tharmaraj.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my indebtedness to my research guide and mentor Prof. Sitabhra Sinha for his continuous support and guidance throughout my research work. I am thankful for his patience, inspiration, enthusiasm, immense knowledge and providing the learning environment. I am also grateful to my collaborators in the work described in the different chapters of my thesis, viz., Dr. Uddipan Sarma, Dr Bhaskar Saha, Dr. Subhadra Halder and Prof Gautam I. Menon.

I would also like to take this opportunity to express my heartfelt gratitude to Shri M. L. Joshi, former head of the Health Physics Division, Bhabha Atomic Research Center, Mumbai who approved, supported and allowed my HBNI PhD work to be carried out at the Institute of Mathematical Sciences (IMSc), Chennai.

I benefited from discussing with several people different aspects of the problems discussed in the thesis and in particular, I would like to thank Prof. Indrani Bose, Dr. Shakti N. Menon, Ms. Busola Olugbon, Dr. Thangarajan Rajkumar and Prof. Somdatta Sinha.

I would like to express my sincere thanks to members of my doctoral committee: Prof. Y. S. Mayya, BARC, Mumbai, Prof. B. V. R. Tata, IGCAR, Kalpakkam, Prof. Sudeshna Sinha, IISER, Mohali and Prof. Gautam I. Menon, IMSc, Chennai, for their encouragement and insightful comments. I learnt a lot from my teachers at IMSc, Prof. Purusattam Ray, Prof. Sudeshna Sinha and Prof. Ronojoy Adhikari and I would like to express my gratitude towards them.

My sincere thanks goes to Prof. R. Balasubramanian, Director of IMSc, Chennai, for extending the facilities of the Institute to carry out my research work. I am grateful to the administrative and computer systems staff at IMSc for all their assistance during my PhD. Some of the research work described here were performed as part of projects funded by different agencies, including the Department of Biotechnology, Government of India (BT/PR7521/BRB/10/482/2006) and the Department of Atomic Energy, Government of India (IMSc Complex Systems Project and PRISM Project, XI and XII Plan).

Finally, I would like to acknowledge my mother D. Angeline Dhayamani, wife A. Saral Jeyaselvi and my sons, J.Graham Staines Dharmaraj and J. Carey Jobson Dharmaraj for their continuous support, sacrifice and love that allowed me to complete my Ph.D. work. Words fail to express my gratitude to my father, B. Tharmaraj, to whose memory the thesis is dedicated.

Jesan T

PUBLICATIONS

1. **Jesan, T.**, Menon, Gautam I. and Sinha, S. (2011). *Epidemiological dynamics of the 2009 Influenza A(H1N1)v outbreak in India*.
   Current Science, **100**, 1051-1054. Arxiv Preprint, 1006.0685

2. Sinha, S., **Jesan, T.** and Chatterjee, N. (2009). *Systems biology: From the cell to the brain*.
   Current Trends in Science: Platinum Jubilee Special (Ed. N. Mukunda), Indian Academy of Sciences, Bangalore, 199-205.
   Arxiv Preprint, 1001.4845.

3. **Jesan, T.**, Sarma, U., Halder, S., Saha, B. and Sinha, S. (2013). *Branched motifs enable long-range interactions in signaling networks through retrograde propagation*. PLoS ONE, **8**(5), e64409.

Contents

List of Figures

List of Tables

Synopsis

The complete mapping of human and other genomes has revealed that the remarkable complexity of living organisms is expressed by less than 30,000 protein-coding genes. Thus, the observed complexity arises not so much from the relatively few components (in this case, genes), as from the large set of mutual interactions that they are capable of generating. The focus of research in biology is therefore gradually shifting towards understanding how interactions between components, be they genes, proteins, cells or organisms, add a qualitatively new layer of complexity to the biological world. This is the domain of systems biology which aims at understanding organisms as an integrated whole of interacting genetic, protein and biochemical reaction networks, rather than focusing on the individual components in isolation.

The recent surge of interest in systems thinking in biology has been fuelled by the fortuitous coincidence in the advent of high throughput experimental techniques (such as DNA and protein microarrays) allowing multiplex assays, along with the almost simultaneous development of affordable high-performance computing which has made possible automated analysis of huge volumes of experimental data and the simulation of very large complex systems. Another possible stimulant has been the parallel growth of the theory of complex networks (comprising many nodes that are connected by links arranged according to some nontrivial topology) from 1998 onwards, which has provided a rigorous theoretical framework for analysis of large-scale networks, ranging from the gene interaction network to the foodwebs. Indeed, reconstructing and analyzing biological networks, be they of genes, proteins or cells, is at the heart of systems biology. The role of such "network biology" is to elucidate the processes by which complex behaviour can arise in a system comprising mutually interacting components. While such emergent behaviour at the systems level is not unique to biology, to explain properties of living systems, such as their robustness to environmental perturbations and evolutionary adaptability, as the outcome of the topological structure of the networks and the resulting dynamics, is a challenge of a different order. As networks appear at all scales in biology, from the intracellular to the ecological, one of the central questions is whether the same general principles of network function can apply to very different spatial and temporal scales in biology. In this thesis, we have analysed in detail transmission processes on several networks occurring at different scales from cell to society to show how using a network approach to study the dynamics complex biological systems can reveal unexpected features and allow new insights. In the following paragraphs I briefly describe the work described in the thesis.

In **Chapter 1** we begin with a short overview of biological networks, focusing on intra-

cellular, as well as, contagion transmission networks. Here we introduce important concepts and definitions used throughout the thesis. This is followed by a discussion of mesoscopic motifs of complex networks, in particular modules, and a brief review of earlier studies of diffusion and transmission dynamics on modular networks. We point out here a distinction between structural modules (investigated in the physics literature) and functional modules (as often used in the biological literature). In the structural sense, modules are subnetworks into which the system can be compartmentalized such that members of the same module are more densely or more strongly connected to each other in comparison to their connections to members of other modules. On the other hand, functional modules comprise members who may be involved in the same task or function. We show in our thesis that functional motifs and structural modules, while distinct, may nevertheless have equally significant roles to play in the dynamics of biological systems.

In **Chapter 2** we investigate a very important intra-cellular network motif, the three-component Mitogen-Activated Protein Kinase (MAPK) signalling module. This pathway is found in all eukaryotic cells and is involved in many critical cellular functions including cell cycle control, stress response, differentiation and growth. Its crucial importance is underscore by the fact that it is seen to be affected in many diseases including cancer, as well as, immunological and degenerative syndromes and is, therefore, an important drug target. The basic linear cascade structure involves regulation of the activity of a MAPK kinase kinase (MAP3K) enzyme by an upstream signal. MAP3K on being activated can act as the enzyme for activation of a MAPK kinase (MAP2K) enzyme which in turn controls the activity of a MAPK enzyme. MAPK, on activation, can be involved in many functions, such as initiation of transcription or stimulation of other kinases. However, such linear or chain-like reaction schemes imply a rigid relation between stimulus and response, precluding the possibility of the system switching to a different response for the same signal under altered circumstances. As many linear cascades are actually part of branched pathways (e.g., the MAP3K enzyme MEKK-1 is known to activate multiple types of MAP2K enzymes in the T-cell and B-cell receptor signalling networks involved in immune response), we have investigated here the dynamics of branched MAPK modules. We demonstrate that enzyme-substrate dynamics on such motifs allow surprisingly long-range communication in the absence of direct long-range interaction between molecules through retrograde propagation between the different (non-interacting) branches of MAPK pathways. Our numerical simulations show that perturbing the activation of MAPK enzyme in one branch can result in a series of changes in the activity levels of molecules upstream to that enzyme, eventually reaching the branch-point and thence affecting the other branches. Our results have recently been verified by biological experiments (done by our collaborators at NCCS, Pune). An important aspect of retrograde propagation in branched pathways that is distinct from previous work on retroactivity focusing exclusively on single chains is that varying the type of perturbation, e.g., between pharmaceutical agent mediated inhibition of phosphorylation or suppression of protein expression, can result in opposing responses in the other branches. This can have potential significance in designing drugs targeting key molecules which regulate multiple pathways implicated in systems-level diseases such as cancer and diabetes.

In **Chapter 3** we have investigated the mesoscopic structural organization of the human cancer disease-gene network. With the growing recognition that cancer is a "systems-

disease", the focus of research in this area has been gradually shifting away from the study of individual molecules and the effect of single gene mutations to an emerging consensus that this complex disease involves significant disruption of the intra-cellular signalling network. One of the drawbacks of a network-based approach to analyzing cancer is the extremely large number of cellular agents whose interactions need to be investigated. We have tried to circumvent this by taking a mesoscopic view of the cancer network, decomposing the network into modules each of which comprises a relatively small number of agents, which are amenable to detailed investigation. We begin with the bipartite network of 146 tumour types and 927 cancer genes (and corresponding proteins) obtained by scanning the relevant databases. Projecting this data onto a single network, we construct a network whose largest connected component consists of 910 genes. Partitioning this network using efficient community-finding algorithms yields 25 modules, the genes within each community having relatively stronger interaction with each other than with members belonging to other communities. We use this result to perform a modular decomposition of the human protein-protein interaction network comprising 9270 proteins grouped into 542 communities. Considering the distance of the cancer gene modules in the abstract protein-module space allows us to build a relational dendrogram between different tumour types, as well as, between different classes of cancer, which gives one a new appreciation of the complex relationships between different categories of tumours and cancer disease types. For example, our analysis shows that the hormonally related disease types of breast cancer and ovarian cancer occur very close to each other in the dendrogram hierarchy. We also investigate the functional role of different cancer genes as revealed by their importance in the modular organization of the network by investigating the joint distribution of their participation coefficients and their within module degree z-scores. We have identified about 36 genes as "connector hubs" occupying critical positions in the cancer network which can be potential targets for therapeutic efforts.

While the chapters discussed above focused on the networks implicated in health and disease at the level of a single cell, in the next three chapters we shift our attention to disease transmission through contact networks, i.e., at the level of a society of human individuals. For such networks, one of the most important dynamical processes to understand is how epidemics are initiated and propagated. In **Chapter 4** we present our work on understanding the outbreak dynamics of the 2009 Influenza A(H1N1)v in India by estimating the initial transmissibility of the disease. This is done by analysing the time-series data for the onset of the influenza pandemic in India during the period June 1 - September 30, 2009. The novel influenza strain (later termed influenza A(H1N1)v) was first identified in Mexico in March 2009, after which it rapidly spread to different countries. The first confirmed case in India, a passenger arriving from USA, was detected on 16 May 2009 in Hyderabad. In fact, most of the initial cases in India were passengers arriving by international flights. However, towards the end of July, the infections appeared to spread to the resident population with an increasing number of cases being reported for people who had not been abroad. To devise effective strategies for combating the spread of pandemics, it is essential to estimate their transmissibility in a reliable manner. This is generally characterized by the reproductive rate R, defined as the average number of secondary infections resulting from a single (primary) infection. A special case is the basic reproduction number R_0, which is the value of R measured when the overall population is susceptible to the infection, as is the case at the initial stage of an epidemic. Using a variety of sta-

tistical fitting procedures, we have obtained a robust estimate of the exponential growth rate $\lambda \simeq 0.15$. This corresponds to a basic reproduction number $R_0 \simeq 1.45$ for influenza A(H1N1)v in India, a value which lies towards the lower end of the range of values reported for different countries affected by the pandemic. We have also separately obtained estimates for different regions of the country which varied over the range 1.34-1.74. This suggests that seasonal and regional variations need to be taken into account to formulate strategies for countering the spread of the disease.

Models of epidemic propagation very often assume that populations are well-mixed (i.e., an infected individual can infect any other individual in the population with equal probability) for mathematical simplicity. However, in reality, individuals very often confine most of their interactions to members of their own social group. Thus, the contact network of individuals in a society can be considered to be modular (of which there is sufficient empirical evidence), with the members of the same module having much higher probability of being infected by each other (as a result of the increased frequency of interactions) as compared to members of different modules. In **Chapter 5** we show that contagion transmission dynamics on modular networks can have startling consequences, in particular, resulting in the persistence of highly infectious diseases. In our study, we have considered a situation where individuals after having recovered from an infection can again become susceptible with a certain probability either as a result of loss of immunity or through removal and subsequent replacement by new individuals. Our study of this SIRS (Susceptible-Infectious- Recovered-Susceptible) epidemic model dynamics over a modular contact network suggests that under certain circumstances an epidemic can become persistently recurrent. Through numerical simulation, we show the dependence of the probability of persistence on the parameters of the network mesoscopic organization as well as on epidemic parameters such as the infection rate α, recovery rate β and the rate at which recovered individuals become susceptible γ. In particular, we show that highly contagious diseases (large α), which quickly die out in a population with homogeneous contact structure, can survive indefinitely (becoming endemic) when there is strong community organization in the population.

The epidemic model studied earlier assumes that infections spread by direct contact between infected individuals. While this is a good model for several types of diseases (such as influenza or chicken-pox), there are several other important diseases which are spread indirectly via an intermediate host (such as malaria, where the vector is the *Anopheles* mosquito). In such situations, apart from the contact network structure we also need to consider the dependence of the vector population density on spatial geography upon which the network is embedded. In **Chapter 6** we consider the spatio-temporal dynamics of malaria transmission by first presenting an empirical analysis of epidemic data from a rural block in north Bengal. The time series data of malaria incidence for two different malaria strains (*Plasmodium falciparum* and *Plasmodium vivax*) recorded over the period Jan 2005- Feb 2009 and obtained from 51 health centers located at different regions in the block are subjected to wavelet phase analysis in order to identify travelling waves of increasing malaria incidence. This has allowed us to locate the epicentres where the outbreaks arise initially and then spread to neighbouring regions. There is significant correlation between phase angle difference between epicentre with other regions and the distance of those regions from the epicentre which substantiates the travelling wave

nature of the epidemic spatio-temporal transmission dynamics. The epicentres are characterized by (a) favourable conditions for high rates of mosquito reproduction such as forest coverage, (b) relatively high human population and (c) high degree of connectedness with neighbouring regions. By correlating epidemic incidence with rainfall data, we find that the latter plays a significant role in the incidence dynamics of malaria with a delay of about 1-2 months. However, rainfall variations have a periodicity of 12 months whereas we observed a dominant periodicity of 8 months in the malaria incidence for certain regions. This suggests that the periodicity of rainfall seasonal variations is not the sole driving factor of the epidemic dynamics. Using a spatially detailed model of malaria transmission we have presented (in the later portion of this chapter) an investigation of the interaction of an externally imposed environmental signal (rainfall with a periodicity of 12 months) with the intrinsic spatio-temporal dynamics of malaria transmission having a different period.

In **Chapter 7** we conclude with a general discussion on how the transmission dynamics on a biological network is governed by modular structure at different scales. The possible implications and applications of our study, e.g., in drug design and assessing the effectiveness of public health intervention procedures, are explored. We also discuss possible future extensions of our research program to investigate related questions such as, the role of convergent pathways on possible congestion of information transmission in signaling networks, the mesoscopic structure of the bipartite network comprising genes related to diseases and the pharmaceutical drugs used for treating such diseases, and the effect of increasing communication speed between population centers on the basic reproduction number for an epidemic.

1

Introduction

We are caught in an inescapable network of mutuality, tied in a single garment of destiny. Whatever affects one directly, affects all indirectly.

– Martin Luther King Jr.

With the completion of human genome mapping [1], the focus of scientists seeking to explain the biological complexity of living systems is shifting from analyzing the individual components such as a particular gene or biochemical reaction to understanding how the interactions amongst the large number of components results in the different functions of an organism. To this end, the area of *systems biology* attempts to achieve a integrated or 'systems-level' description of biology by investigating the network of interactions connecting the various elements together instead of studying in isolation the properties of a few of the components. While the term 'systems biology' itself is of recent coinage [2,3], the field has had several antecedents, most notably, *cybernetics*, as pioneered by Norbert Wiener [4] and W. Ross Ashby [5] (who indeed can be considered to be one of the founding figures of the related discipline of systems neuroscience along with Warren McCulloch [6]) and the *general system theory* of Ludwig von Bertalanffy [7] which have profoundly influenced many disciplines including biology. However, what distinguishes

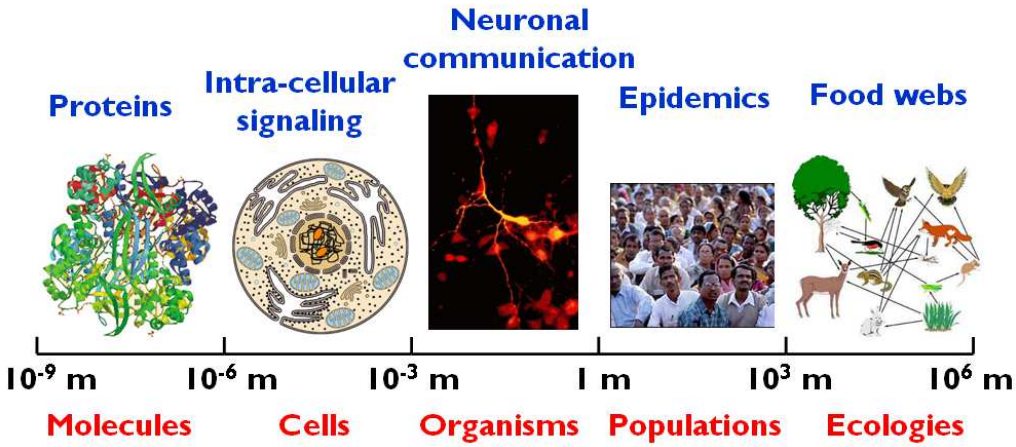

Figure 1.1: **Biological networks appear at different length scales.** Dynamical processes mediated through complex networks occurring at different length scales in the biological world, from the protein contact networks at the molecular level to the network of trophic relations spanning several tens or hundreds of kilometers. In between these two extremes, we see intra-cellular networks (e.g., those involved in intra-cellular signaling), inter-cellular networks (e.g., neuronal networks in the brain) and networks among individuals comprising a population (e.g., the social contact network through which various epidemic diseases spread) [8].

systems biology from earlier attempts at analyzing the complexity of biological systems is the focus on network representation of such systems. This has facilitated the integration of vast quantities of experimental data using high-performance computing allied with various efficient algorithms and has provided a platform for comprehending the key features of the processes being investigated. This approach has been very productive as complex networks are seen in the biological world at all length scales, from that of molecules (protein contact networks) to ecologies (food webs) (Fig. 1.1).

Networks, i.e., systems comprising many interacting elements, are ubiquitous in the world around us [9, 10]. They are often represented as *graphs*, consisting of a set of *vertices* or *nodes* that are connected to each other by *links* or *edges*. While networks have been studied extensively in the social sciences for a long time [11], the past decade has seen a techniques of network analysis being applied to understand an enormous variety of complex systems in the biological (e.g., metabolic and protein interaction networks) and technological (e.g., the internet and electrical power grid) domains [12, 13]. The availabil-

ity of affordable high-performance computing allowing the representation and analysis of systems consisting of thousands and in some cases, millions of nodes, that may also be evolving over time, has revealed unexpected universal features in the large-scale statistical properties of such systems. These discoveries, beginning with the identification of "small-world" property, i.e., the coexistence of low average path length between nodes (characteristic of random networks) and highly clustered neighborhoods (characteristic of regular grids or lattices) [14], and the observation of a scale-free distribution of degree (i.e., links per node) in many networks that occur in real life [15], have made networks one of the most exciting areas of research in statistical physics. Along with the study of structural properties of networks, there has been increasing interest in the dynamics on and of complex networks, as most biological, social and technological complex systems are inherently dynamic. The idea of dynamics being strongly influenced by the structure of the underlying network was suggested by Watts and Strogatz in their work on small-world networks using the example of epidemic spreading, it is also of interest to see how the inherently nonlinear dynamical processes in complex systems can constrain the topological structure of the networks.

In this thesis, we present results of theoretical and simulation studies exploring in detail at the interplay between structural and dynamical aspects of biological networks occurring at very different scales - from intra-cellular signaling networks to the network of epidemic or contagion propagation in human societies. We give below a brief overview of biological networks, looking in turn at the different spatial scales ranging from intra-cellular to inter-organism and inter-species networks. This is followed by a short description of the key network concepts and definitions used throughout the thesis. Next we discuss mesoscopic structural features of complex networks, in particular, modular organization. Finally we briefly review earlier work on diffusion and transmission dynamics on such networks before concluding with an overview of the thesis.

1.1 Complex Networks in Biology

Several biological processes involve numerous interactions between the constituent elements, often across different length scales, varying from molecular processes to species interactions. At the intra-cellular length scale, one finds networks of chemical interactions between cellular molecules; for instance protein networks, gene regulatory networks, metabolic networks and signalling networks. At the inter-cellular length scale, one finds the neuronal network, which represents the connections between neurons in the brain. We observe network of interactions between individuals of a group at the inter-organisms length scale. Biology typically overlaps with social science at this scale, most prominently in the widely studied issue of epidemic spread through social networks. At a much larger scale, we have inter-species networks, e.g., the food web of an entire ecosystem. In these various contexts, the role of network biology is to elucidate the processes by which complex behavior can arise in a biological system comprising mutually interacting components and to provide a theoretical framework that allows a convenient representation of inter-relations in such complex biological systems.

1.1.1 Intra-cellular networks

A variety of networks control the multitude of dynamical processes responsible for the proper function of a biological cell [16]. These include genetic regulatory networks [17, 18] in which genes regulate the expression of other genes through activation or inhibition (i.e., by expressing proteins that act as promoters or suppressors of other genes) as seen, for example, the formation of patterns that occur during the development of a fertilized cell into an embryo [19]. Metabolic networks includes bio-chemical reactions [21–23], that are responsible for breaking down organic compounds to extract energy as well as those reactions which use energy to construct vital components of the cell such as amino acids and nucleic acids. While the glycolytic pathway that converts glucose into pyru-

vate (the first significant portion of the metabolic network to be reconstructed) took many years to be elucidated, there are now experimental techniques such as the yeast two-hybrid screening method that test for physical interactions between many pairs of proteins at a time, allowing for rapid reconstruction of such networks. One of the most intriguing cellular networks is the protein-protein interaction network that is responsible (among other things) for intracellular signaling, the mechanism by which a cell responds to various stimuli through an ordered sequence of biochemical and physical association reactions [20]. These reactions regulate processes vital to the development and survival of the organism (e.g., differentiation, cell division, apoptosis, etc.) by transmitting information from receptors located at the cell surface (that receive external signals) to specific intracellular targets in a series of enzyme-substrate reaction steps [24].

Protein Interaction Network (PIN)

Protein interactions play central role in almost all cellular functions such as regulation and signaling. Protein molecules are built from a long chain of twenty amino acids, each linearly linked to its neighbor through a covalent bond. Each unique side chain of amino acids within this linear primary structure of protein bond with one another and then fold into a unique three dimensional structure. The precise shape of this structure decides its role in the cellular processes. Protein structure can be represented as a network of non-covalent connections (links) between the constituent amino acids (nodes). This is referred to as a protein contact network, and it is one of the smallest networks (length scale: $\sim 10^{-8}$ m) in the natural world [25]. The small-world property of PCNs for different protein molecules has been noted numerous times in the literature [26, 27]. The existence of modular structure and distinct time scales that correspond to the inter- and intra-modular modes of motion has been shown through the study of protein contact network dynamics [25].

The analysis of protein-protein interaction (PPI) networks provide crucial information re-

lated to the coordination of protein function required to explain the cellular structure and dynamics. In a protein-protein interaction network, the nodes are proteins and links between them are assigned if it has been experimentally confirmed that they are connected. Generally, protein interaction data has binary nature and their strength is not quantified. The representation as a network of a protein's interactions with other proteins, DNA, RNA, and small molecules provides interesting insights into the biological process within a cell and highlights the importance of the various proteins in terms of both structure and dynamics [28]. Protein interaction data are derived from different experimental methods and a variety of large databases are available, most of which are organism specific, such as the Human Protein Reference Database (HPRD). By analyzing different databases one can identify several robust features of these network, a prominent one being the scale free property of the degree (i.e., number of links per node) distribution [29]. This property underlines the danger of mutations in highly interacting proteins, which are expected to be lethal for the cell, a prediction that is supported by explicit measurements [30, 31]. The rapid increase in the availability of human protein interaction data and studies of networks of human disease/ human gene associations [32] has led to identification of the networks underlying human disease [33]. Studies have also found that disease genes exhibit an increased tendency for their protein products to interact with one another, that they tend to be coexpressed in specific tissues, and display coherent functions with respect to all three branches of the Gene Ontology hierarchy [34]. Similarly, cancer-related proteins tend to have, on average, twice as many interaction partners as non-cancer-related proteins [35,36]. Also, cancer-related proteins tend to reside in larger clusters and participate in more clusters than non-cancer-related proteins [37]. Capturing the spatio-temporal changes in protein connectivity that are associated with the progression towards disease permits scientists to select protein targets for therapeutic intervention through an understanding of the underlying mechanisms of drug-protein interactions and potential toxic side-effects [38, 39].

Gene Regulatory Network (GRN)

Cells must continually adapt to changing conditions by altering their gene expression patterns. Gene expression is a complex sequential processes regulated at several stages that ultimately results in a specific quantity of target proteins known as transcription factors. Proteins are created in the cell by a two stage mechanism. In the first stage, known as transcription, an enzyme called RNA polymerase makes a copy of the coding sequence of a single gene. This copy consists of RNA, as well as another information-bearing biopolymer that is chemically similar but not identical to DNA. RNA copies of this type are known as messenger RNAs. In the second stage, known as translation, the protein is assembled, step by step, from the RNA sequence by an ingenious piece of molecular machinery known as a ribosome, which is a complex of interacting proteins and RNA. The translation process involves the use of transfer RNAs, which are short molecules of RNA that have a region at one end that recognizes and binds to a codon in the messenger RNA and a region at the other end that pulls the required amino acid into the correct place in the growing protein. The end result is a protein, assembled following the exact prescription spelled out in the corresponding gene. In the jargon of molecular biology, one says that the gene has been expressed. Proteins synthesized from genes may function as transcription factors promoting or inhibiting production of one or more proteins, as enzymes catalyzing metabolic reactions, or as components of signal transduction pathways. The complete set of such interactions forms a Genetic Regulatory Network(GRN) [10]. In such networks of genes, proteins and other biological molecules are represented as nodes and the corresponding series of regulatory interactions are represented as positive (promoting) and negative (inhibiting) links. Genetic regulatory networks were one of the first networked dynamical systems to be simulated using large scale modelling [13]. The very early pioneering work on GRN looks at the genetic control of cellular differentiation by studying random networks of elements controlled by Boolean logic [17]. GRNs often show specific motifs and they mostly follow the power-law degree distributions (scale-

free network), even though some of them, like the transcriptional regulatory networks of *E. coli* and *S. cerevisiae* have been shown to possess mixed scale-free and exponential properties [19, 40]. The importance of structure and dynamics of gene regulatory networks in every cellular process is recognized, including cell differentiation, metabolism, the cell cycle and signal transduction. Understanding the dynamics of these networks may provide insights in to mechanisms of hereditary diseases that occur when these cellular processes are dysregulated and help in identifying appropriate drugs targets [41].

Metabolic Network

Metabolism in the cell is the complicated sequence of chemical reactions that convert substrates (food or nutrients) from the environment to energy (e.g., in the form of ATP) as well as synthesize biological molecules that are essential for the growth and survival of the cell. Metabolic networks are directed networks whose nodes are chemicals generally known as metabolites that are connected to one another through metabolic reactions (links). Metabolic network analysis helps in understanding the chemical dynamics in the cell and to identify the fundamental properties of living systems. It also has many industrial applications such as production of chemicals, antibiotics, enzymes, antibodies, drugs that cure metabolic diseases of human beings, predicting selective drug targets and the control of pathogens [42, 43]. Metabolic networks are extremely heterogeneous and vary from organism to organism [29]. The scale-free structure remains robust even after removal of some central nodes (metabolites) [44]. The architecture of the metabolic network rests on highly connected metabolites [21]. The structural analysis of metabolic networks in many species indicates the existence of a small-world structure, with the nodes in the network connected through short paths [21]. The metabolic network of *E. coli* is an example of such a small-world network. The occurrence of such structure may contain traces of the evolutionary history of metabolism [22]. It has been found that metabolic networks consists of functional modules - the existence of a set of connected molecules

(nodes) that work together to achieve a function - whose nodes can be classified into universal roles according to their patterns of intra- and inter-modular connections [23]. It has been shown that metabolic networks have a hierarchical modular nature that can be decomposed into several small, but highly connected modules that combine in a hierarchical manner to form larger, less cohesive units with their number and degree of clustering following a power law [45].

Signal Transduction Network

Signal transduction is the process by which a cell receives information from the extracellular environment, modulates and process the information and then regulates the related intracellular processes. The inputs to signal transduction networks are the binding of extracellular ligands (e.g., hormones, cytokines, growth factors) to the corresponding receptor proteins embedded in the cell membrane. This binding triggers a cascade of downstream signal transduction reactions, i.e., each key component becomes activated by the previous step to activate the key molecule of the subsequent reaction. Signal transduction processes rely on a cascade of reversible chemical modification of proteins, as well as, on the formation of complexes. The final targets of signal transduction processes are transcription factors and metabolic enzymes. The signaling networks of a cell are capable of recognizing sets of inputs and responding appropriately, with their connection strengths having been selected during evolution [46]. In contrast to metabolic networks, signal transduction networks consist of a rather limited mass flow and mainly facilitate the transmission of information along a sequence of reactions [47]. Signal transduction networks display small-world properties and scale-free topology [47]. Although protein-protein interactions are typically isotropic and undirected, in signal transduction networks they are anisotropic (a set of inputs is transformed into a set of outputs, but the reverse is not possible) and directed [46, 47].

Intra cellular signaling subnetworks (modules) that act as distinctly identifiable functional

units performing specific tasks interact with one another to form complex networks [48]. Models of the dynamics of reactions occurring in individual modules are an essential tool that allow us to understand the set of interactions over the complex bio-molecular network and facilitates the eventual development of a coherent picture for the functioning of the entire cellular network. Complex organization of signaling modules may significantly contribute to the robustness of cellular functions. Robustness refers to the ability of a system to maintain its functionalities against external and internal perturbations [49]. Another important tool to characterize the signaling network are motifs such as self-sustaining feedback loops which describe the regulatory features and reveal information about the topology of the network that are important for biological functionality [50]. Negative feedback loops can give rise to adaptation and desensitization, while positive feedback loops can lead to emergent network properties such as ultrasensitivity and bistability [51–53].

1.1.2 Inter-cellular networks

As we move up in scale from cellular to multi-cellular systems, we encounter issues related to inter-cellular communication and the response of such systems to events in the external environment. Possibly the most intriguing questions in this domain of system biology are concerned with the activity of the brain and the nervous system, and comprise the discipline of systems neuroscience. This field explores the neural basis of cognition, as well as of motivational, sensory, and motor processes, and also fills the gap between molecular and cellular approaches to the study of brain and the behavioral analysis of high-level mental functions. Given the complexity, inaccessibility and heterogeneity of the brain, the tools associated with systems neuroscience are interdisciplinary in nature. The aim is to use all these tools to have a better understanding of the integrated functioning of large-scale distributed brain networks and how disruptions in brain function and connectivity impact behavior, as well as to addresses questions on both the normal

and abnormal functioning of the nervous system. One of the main functions of the brain is to process information. The primary information processing element is the neuron, a specialized brain cell that integrates several inputs to generate a single output. The actual signals that travel within neurons are electrochemical in nature. In recent years the network approaches have offered significant new insights into how the structure of the brain shapes its dynamics and how the elements of a neural network can make different contributions to brain function on the basis of how they are interconnected. In network terms, a neural network is represented as a set of neurons (nodes) and that are connected by two types of directed links, excitatory and inhibitory. The relation between brain structure and dynamics presents unique challenges to neuroscientists.

The worm *Caenorhabditis elegans* is an excellent experimental system for understanding the relationships between structure and function in an entire nervous system. The nervous system of this nematode has been completely mapped and it consists of only 302 neurons connected by approximately 2000 links. The unique identification of neurons and availability of a detailed physical connectivity map derived from ultrastructural analysis with electron microscope and electrophysiological data [54] makes this the system of choice for those wishing to use technique of network analysis to understand nervous system functioning. It has been shown that principles of wiring economy and developmental constraints do not completely decide the connection structure of the network [55]. The large number of recurrent connections observed among interneurons suggests that a hierarchical structure (having a densely connected core and an overall sparse structure) prevents indiscriminate global activation of the nervous system, while at the same time permitting large density of connections that allows high communication efficiency, so that information can propagate rapidly from sensory to motor neurons [55].

1.1.3 Inter-organism networks

At the scale of individual organisms, such as human beings, one of the most widely studied networks is the social contact network along which epidemics propagate [25]. Many infectious diseases spread in human populations through contact between infected and susceptible individuals. Traditionally, mathematical studies of epidemic propagation utilized the homogeneous mixing hypothesis meaning that an infected individual of the population has the same probability to spread the disease to any other individual chosen randomly [56, 57]. However, in the real world, each individual only has contact with a small fraction of the total population, although the number of contacts that people have can vary from one person to another. Network models of epidemic refer to individuals (nodes) and their network of potential contacts (links) instead of assuming that contact is possible with the entire population. Small-world networks play an important role in the study of the influence of the contact network structure on dynamics of social processes of epidemic spreading.

Modelling of epidemic spreading on networks has to take into account the ubiquity of small-world networks in nature. In particular, the worldwide spread of SARS demonstrated that even a few long-range links can significantly enhance the spread of diseases [61]. There have been many studies of epidemic spreading on small world or related network models [60, 62]. Note that modular networks share all the structural characteristics of such small world networks, although having very different dynamical properties [25]. It has been shown that small world networks have a continuous range of time-scales and modular networks exhibit two sets of distinct time-scales that are related to intra- and inter-modular events [63]. Epidemiology modelling results are much needed for planning, executing and evaluating different prevention, therapy and control methods. Thus, devising an effective strategy to counter the spread of epidemics will have to take into account a detailed knowledge of such structures in the social network of susceptible and infected individuals.

1.1.4 Inter-species networks

In terms of length scale, the largest possible biological networks on earth are the interactions between different species in an ecosystem [25]. Ecological networks consist of all possible links in the ecosystem, such as cooperation, competition and predator-prey relations between species. In food webs, the nodes represent species in an ecosystem and the directed links represent predator-prey relationships. The direction of the links indicate the flow of biomass and the links are usually weighted to represent the amount of energy that is transferred. Most observed food web networks are highly structured. This is one of the principal arguments advanced against the May-Wigner stability theorem, which suggests that increased complexity of a network inevitably leads to its destabilization [64]. Food webs have been shown to have modular structure, with species in each module interacting between themselves strongly and only weakly with species in other modules [66]. It is thought that modularity plays a vital role in stabilizing the dynamics of ecosystems [25].

1.2 Basic concepts of complex networks

A complex network can be represented as a graph comprising a set of N nodes (or vertices) and L links (or edges) connecting pairs of nodes.

Adjacency Matrix

Any pair of nodes connected by an link are said to be adjacent or neighboring. The adjacency matrix provides a mathematical representation of a network. The adjacency matrix $A = \{a_{ij}\}_{N\times N}$ for a graph G with N nodes is an $N \times N$ matrix whose elements a_{ij} =1 if the i-th and j-th nodes are connected, and a_{ij} =0 if they are not. The diagonal matrix elements a_{ii}=0 as there are no self-connections of the nodes. An undirected network has a symmetric adjacency matrix as a link between i and j implies that there is an link between j and i: $a_{ij} = a_{ji}$. For directed networks each link has an associated direction and thus, the adjacent matrix is asymmetric. In addition to the number of connections, nodes in

many real networks can have links of different intensities with other nodes. The weights of the links in weighted networks are represented by a weight matrix W whose elements w_{ij} represent the intensity of the connection between i-th and j-th nodes.

Degree

The degree k_i of a node i in a network is the total number of links that it has with other nodes in the network. Degree can be calculated from the adjacency matrix as

$$k_i = \sum_{j=1}^{N} a_{ij}.$$

In directed networks, one can define the in-degree k_i^{in} and the out-degree k_i^{out}, corresponding to the total number of incoming links to node i and outgoing from node i, respectively.

Strength

In a weighted, undirected network the strength s_i is the sum of the weights for the links of a node i:

$$s_i = \sum_{j=1}^{N} w_{ij}.$$

In directed networks, one can define the in-strength s_i^{in} and the out-strength s_i^{out}, corresponding to the total intensity of incoming links to node i and outgoing from node i, respectively.

Degree Distribution

The variations in the number of links of different nodes in a network can be characterized by the degree distribution $p(k)$ defined as the probability that any node in the network will have degree $k = 1, 2, 3, ...N - 1$. Several networks occurring in nature display degree distribution having power law form with exponent γ having values ranging between 2 to 3. As power-laws have the property that the same form appears at all scales, the networks showing power-law distributions are often referred to as scale-free networks. Other networks can exhibit truncated power law or exponential degree distributions [67].

Path length

In many networks it is possible to go from any node to any other node in a very small number of steps following the connections of the network. A network is called *connected* if there is a path connecting any node to any other node in the network. As there are many possible paths connecting a given pair of nodes i and j, one defines the distance between the two nodes as the shortest path length d_{ij} between them, i.e., the number of links that must be traversed to go from one node to another using the shortest route. The average path length l is defined as the average value of d_{ij} over all the possible pairs of nodes in the network, and is also known as characteristic path length:

$$l = \frac{1}{\frac{1}{2}N(N-1)} \sum_{i \geq j} d_{ij},$$

where d_{ij} is the shortest path length from node i to j and N is the number of nodes in the network. For networks having disconnected components, the average path length l can be calculated as

$$E \equiv l^{-1} \equiv \frac{1}{\frac{1}{2}N(N-1)} \sum_{i > j} \frac{1}{d_{ij}},$$

where E is termed as the communication efficiency. It quantifies the information propagation speed over the network [68]. Another measure of a network size is its diameter D, the longest of all shortest path lengths between two nodes in the network. The diameter is defined as $D = max\{d_{ij}\}$ $\forall$ pairs (i, j) of shortest path lengths.

Clustering

Clustering implies that neighbors of a node are also mutual neighbors. The clustering of an undirected network is quantified by the clustering coefficient C_i which measures the degree to which the neighbors of a node are connected to each other [14]. Given a node i, the clustering coefficient C_i of a node i is defined as the ratio of the number of links between the neighbors of i to the maximum possible number of such links. If the degree of node i is k_i and these nodes have T_i links among them, then $C_i = \frac{2T_i}{k_i(k_i-1)}$. If $C_i = 1$, the neighborhood of node i is completely interconnected while if $C_i = 0$ none of

its neighboring nodes are connected to each other. The average clustering coefficient C for an entire network is

$$C = \frac{1}{N} \sum_{i=1}^{N} C_i.$$

A network with short average path length and a high clustering coefficient C is termed a small-world network (SWN). Many of the networks observed in reality have been shown to be SWN [14].

Centrality

The importance of a node in a network depends on many factors. One of the key measures quantifying the topological importance of a node is its centrality. The *degree centrality* depends on the degree of a node, with a higher degree centrality implying that the node is connected to many other nodes. The *closeness centrality* g_i is measured using the average path length of a node to all other possible pairs of nodes as

$$g_i = \frac{1}{\sum_{j \neq i} l_{ij}}.$$

This measure gives a large centrality to nodes which are connected via short paths to many other nodes. The concept of *betweenness centrality* b_i underlines the crucial role certain nodes play in connecting different regions of the network by acting as bridges [69, 70]. This is measured as follows. For each node i, in the network, the number of routing paths to all other nodes going through i is counted. Then the the betweenness centrality b_i of node i is defined as

$$b_i = \sum_{j,k \in N, j \neq k} \frac{n_{jk}(i)}{n_{jk}},$$

where n_{jk} is the number of shortest paths connecting j and k while $n_{jk}(i)$ is the number of shortest paths connecting j and k and passing through node i. The betweenness centrality b_i is useful for determining the community structure of biological and social networks [30, 71].

1.3 Mesoscopic organization of complex networks

Most of the real-world complex networks display more complex architectures than classical random or regular networks. These complex structures which are the result of continually evolving dynamics can in turn affect the function of the network [72]. This has directed the focus of research work in the area of complex networks from purely structural aspects of the connection topology to dynamical processes. For instance, we may be interested in the collective behavior of a large ensemble of dynamical systems that interact through a complex connection topology [72]. New emergent features can rise from the interplay between structure and function in complex networks [73]. In the last few years, researchers have made important steps toward understanding qualitatively different properties of critical phenomena in complex networks, e.g., in the context of synchronization, diffusion, etc. [74]. The focus of research activity related to interplay between structure and functions of complex networks has shifted from the properties of individual nodes (microscopic level) such as the degree k distribution $P(k)$ [15] as well as from the properties characterizing the entire network in terms of a global value (macroscopic level), such as average path length or clustering coefficient [14]. Indeed, recent work has revealed significant large-scale heterogeneity in many complex networks: the statistical properties of nodes may strongly differ in different parts of a network. Thus networks which are indistinguishable at both the micro-scale and the macro-scale may nevertheless have radically different behavior [63, 74, 75]. The origin of this difference lies in their mesoscopic organization which can be structurally manifested as patterns in the arrangement of links between subparts of the network [75, 76]. The term *mesoscopic* refers to intermediate scale between the micro and macro scales of a network . Analyzing networks at the mesoscopic level and considering the local patterns in the inhomogeneous distribution of connections may reveal vital clues that may lie hidden in a macroscopic statistical description [77]. One of the prominent examples of mesoscopic organizing principles operating in networks is the existence of communities (modules) in many real-world networks [78].

1.3.1 Network Modules: Structural and Functional

Modules in a network is defined as subnetworks whose components (nodes) are much more densely and/or strongly connected with each other compared to connections with nodes that belong to different subnetworks. Modules have a intrinsic importance in revealing the organizational principle of networks and ultimately link the structure of the network with its functionality [23,79]. In metabolic networks, a module might correspond to a circuit, pathway or motif that carries out a certain function, such as synthesizing or regulating a vital chemical product [21,23]. A module might also be equivalent to an actual community in society, i.e., a group of people with a common interest, people who live in the same locality or belong to a common workplace, or a group of relatives and friends in social networks [80]. Analyzing the modules can reveal vital clues about the network organization, enhance understanding of the dynamic processes taking place on the network and uncover relationships between the nodes that are not evident by examining the whole network. This may help in uncovering the possible mechanism for module formation, understanding the effect of modular structure on dynamic processes (e.g., spreading processes of epidemics) and revealing the functions of a system [81].

From a topological perspective, modules are groups of highly connected set of nodes having relatively less number of links with nodes of other groups in the network. The statistical properties of modules are found to be quite different in many networks and affect the dynamical functions of networked systems [71]. In various social and biological networks, the existence of modular structure can be associated to the specific functions of such modules [23]. Functional modules are groups of nodes (components) that perform particular tasks. This may be intimately related to the nature of interactions among the nodes, such that the function cannot be explained by considering only the properties of isolated nodes. Biological networks are known to often exhibit functional modules [28]. Nodes may belong to different modules at different times and can have different functions [28]. For example the Mitogen-Activated Protein Kinase (MAPK) pathway in combina-

tion with other proteins forms a functional module and perform many different functions including signal-amplification in the intra-cellular signaling network [83–86]. In general, the functional units within a biological network correspond to modules. Structural modules need not always correspond to functional units in the network. However, structural organization exhibiting modules may often provide vital clues to the modular functionality of the network. The existence and robustness of modular structure and functional modules in biological networks implies that modular organization is a result of evolutionary processes [82, 87]. Thus, identifying the modular structure in real-world networks is a vital step towards understanding complex biological systems.

1.3.2 Identifying structural modules in networks

Identifying structural modules is one of the most important problems related to understanding the link between structure and function in complex networks and has applications in many disciplines. The mathematical formalization of this problem is closely related to the ideas of graph partitioning and the first algorithms for graph partitioning were proposed in the early 1970's [88]. In 2002, Girvan and Newman proposed a new algorithm, that identified links between modules and recursively removed them. After several iterations this process results in the isolation of the modules [71,88]. This triggered substantial work on the problem and in subsequent years many new methods have been proposed . The most successful solutions to the community detection problem, in terms of accuracy, are those based on the optimization of a quality function called modularity proposed by Newman and Girvan, which allows the comparison between different partitionings of the network [77, 80]. This is based on the idea that a random network is not expected to have modular structure. So the existence of modules is revealed by the comparison between the actual density of links in a subnetwork and the density one would expect in the subnetwork if community structure was absent [88]. Given a network partitioned into modules, m_i being the community to which node i is assigned, the mathematical definition

of modularity Q is expressed as

$$Q = \frac{1}{2s} \sum_{ij} \left(w_{ij} - \frac{s_i s_j}{2s} \right) \delta(m_i m_j),$$

where w_{ij} is the weight of the connection between nodes i and j (the total strength is $2s$). The Kronecker delta function $\delta(m_i m_j) = 1$ if node i and j are into the same community, $= 0$ otherwise. As the only non-zero contributions to the sum are from pairs of nodes belonging to the same module, for a given partition of the nodes of a network into M modules,

$$Q = \sum_{m=1}^{M} \left[\frac{l_m}{s} - \left(\frac{d_m}{2m} \right)^2 \right],$$

where l_m and d_m are the links between nodes and the total strength of all nodes belonging to module m, respectively. By maximizing the modularity Q, an optimal partitioning of the network into its constituent modules can be obtained.

There are currently numerous algorithms that have been proposed to find the partition resulting in maximum modularity in a reasonable time. The algorithm proposed by Girvan and Newman (GN) [71, 89] is one of the early algorithms used to identify modules in networks by this method. It is a hierarchical divisive algorithm in which links are iteratively removed based on the value of their betweenness which is related the number of shortest paths between pairs of nodes that pass through the link. The process of link removal ends when the modularity of the resulting partition reaches a maximum. A fast greedy optimization algorithm developed by Clauset *et al.* [90] maximizes the modularity by starting initially with a set of isolated nodes and then iteratively adding the links of the original network so as to produce the largest possible increase in the modularity at each step. Higher precision in maximizing modularity Q can be achieved via simulated annealing [23] at the expense of computational speed [91]. The fast modularity optimization by Blondel *et al.* [92] is a multi-step technique based on a local optimization of the modularity in the neighborhood of each node. After a partition is identified, the communities are replaced by super-nodes, yielding a smaller weighted network. The procedure is then

iterated until modularity (which is always computed with respect to the original graph) does not increase any further [91]. *CFinder* is an algorithm proposed by Palla *et al.* [93] that looks for modules which may overlap, i.e., share nodes. It is based on the concept that the links within a module are very likely to form cliques due to their high density while it is unlikely that inter-modular links will form cliques [88, 91].

Rosvall and Bergstrom [94] have used information theoretic concepts to identify modules in networks. They consider a compressed description of the network that approximates the information contained in the adjacency matrix, the premise being that the maximum compression can be achieved when any existing modular organization is incorporated into the description. This is achieved by computing the minimum of a function which expresses the best tradeoff between the minimal conditional information between the original and the compressed information and the maximal compression. The optimization of the function is carried out via simulated annealing (referred to as *Infomod*). A dynamic algorithm developed by Rosvall and Bergstrom [95] is based on the same principle as *Infomod*, except that a dynamic process (random walk) takes place on the network. This method is also referred to as *Infomap*. Performance results of several benchmarks tests have shown the *Infomap* method to be very efficient in detecting modules in a given complex network [88, 91].

1.3.3 Role of modules in dynamics: multiple time-scales

Dynamical processes taking place on networks with modular structure can result in complex non-linear behavior. The spectral analysis of the Laplacian matrix of a network is an extremely useful method for describing the dynamics of such a system [96]. The Laplacian matrix L for a network is defined as $L = D - A$, where A is adjacency matrix of the network and $D = d_{ij}$ is the degree matrix, which is a diagonal matrix with $d_{ii} = k_i$, the degree of the i-th node, $i = 1, 2, ..., N$. For undirected networks, the Laplacian matrix is symmetric and positive semi-definite, with nonnegative real eigenvalues

$\lambda_1 = 0 < \lambda_2 \leq \lambda_3 .... \lambda_N$. The eigenvalue λ_2 is nonzero if and only if the network is connected. Indeed, the number of zero eigenvalues of L equals the number of isolated modules. The smallest non-zero eigenvalue can be related to the time required for synchronization or diffusion to occur over the entire system. The increase in the number of links between modules can result in increase in value of λ_2, which also implies improved synchronizability. For a network with M modules, the difference in the eigenvalues λ_M and λ_{M+1} (spectral gap) is larger than the difference between any other consecutive eigenvalues of L [63]. This property is often considered to be a signature of the presence of modular structure in the network. For modular networks, a typical synchronization process generally starts from partial synchronization through local modular synchronization to global complete synchronization. Dynamics of synchronization in modular networks tend to produce separation between two distinct time-scales: fast intra-modular processes and slow inter-modular processes. Many complex systems across social, technological, and biological systems have the property of hierarchical modular network organization (modules-within-modules), meaning that modular structure is expressed repeatedly at different hierarchical levels within each module [97]. Dynamical processes on hierarchical modular network exhibit many distinct time-scales corresponding to the many gaps in the Laplacian spectrum (the number of gaps exactly equal the number of hierarchical levels) [75, 76].

1.3.4 Dynamics of propagation on networks

Propagation (of signals, traffic, etc.) is an important concept in biological complex networks due to its wide range of applications spanning the epidemics to signaling in the cell. The interplay between structural and dynamical features characterize the propagation property of a network.

Diffusion on networks

Diffusion on a network involves a stochastic flow along all possible links between nodes.

Let us suppose that a node i contains ψ_i fraction of random walkers at a given time. Then, in a small interval of time dt, random walkers move from node i to an adjacent node j at a rate $D_c(\psi_i - \psi_j)$ where D_c is the diffusion constant [10]. The rate of change ψ_i is given by

$$\frac{d\psi_i}{dt} = D_c \sum_j A_{ij} \left(\psi_j - \psi_i\right).$$

The generalized equation in matrix form can be represented as

$$\frac{d\psi}{dt} = D_c \left(A - D\right) \psi,$$

where A is the adjacency matrix and D is the diagonal degree matrix mentioned above. Thus, the diffusion dynamics in networks can be explained using the Laplacian matrix L [10, 63] as

$$\frac{d\psi}{dt} + D_c L \psi = 0,$$

where $L = D - A$.

Transmission on networks

Transmission on a network involves transfer or directed flow along directed links between nodes. An example is the transmission of electrical power along high-voltage lines over long distances. Similar directed flow can be seen in metabolic networks where biomolecular mass flows along a specific sequence of chemical reactions. Information flow along the synaptic connections of neuronal networks provides another example.

1.4 Overview of the thesis

In the following paragraphs we briefly describe the work reported in the thesis.

Branched structures arise in the intra-cellular signaling network when a molecule is involved in multiple enzyme-substrate reaction cascades. Such branched motifs are in-

volved in key biological processes, e.g., immune response activated by T-cell and B-cell receptors. In **Chapter 2** we have demonstrated that long-range communication can occur on the network through retrograde propagation between branches of signaling pathways whose molecules do not directly interact. Our investigation of a model system comprising branches, with JNK and p38MAPK as terminal molecules respectively, that share a common MAP3K enzyme MEKK3/4 show that perturbing an enzyme in one branch can result in a series of changes in the activity levels of molecules "upstream" to the enzyme that eventually reaches the branch-point and affects other branches. In the absence of any evidence for explicit feedback regulation between the functionally distinct JNK and p38MAPK pathways, experimentally observed modulation of phosphorylation amplitudes in the two pathways when a terminal kinase is inhibited implies the existence of long-range coordination through retrograde information propagation as predicted by our simulations in models comprising two or more branches. An important aspect of retrograde propagation in branched pathways that is distinct from previous work on retroactivity focusing exclusively on single chains is that varying the type of perturbation, e.g., between pharmaceutical agent mediated inhibition of phosphorylation or suppression of protein expression, can result in opposing responses in the other branches. This can have potential significance in designing drugs targeting key molecules which regulate multiple pathways implicated in systems-level diseases such as cancer and diabetes.

In **Chapter 3** we have investigated the mesoscopic structural organization of the human cancer disease-gene network. With the growing recognition that cancer is a "systems-disease", the focus of research in this area has been gradually shifting away from the study of individual molecules and the effect of single gene mutations to an emerging consensus that this complex disease involves significant disruption of the intra-cellular signalling network. One of the drawbacks of a network-based approach to analyzing cancer is the extremely large number of cellular agents whose interactions need to be investigated. We have tried to circumvent this by taking a mesoscopic view of the cancer network, decomposing the network into modules each of which comprises a relatively small number of

agents, which are amenable to detailed investigation. We begin with the bipartite network of 146 tumor types and 927 cancer genes (and corresponding proteins) obtained by scanning the relevant databases. Projecting this data onto a single network, we construct a network whose largest connected component consists of 910 genes. Partitioning this network using efficient community-finding algorithms yields 25 modules, the genes within each community having relatively stronger interaction with each other than with members belonging to other communities. We use this result to perform a modular decomposition of the human protein-protein interaction network comprising 9270 proteins grouped into 542 communities. Considering the distance of the cancer gene modules in the abstract protein-module space allows us to build a relational dendrogram between different tumor types, as well as, between different classes of cancer, which gives one a new appreciation of the complex relationships between different categories of tumors and cancer disease types. For example, our analysis shows that the hormonally related disease types of breast cancer and ovarian cancer occur very close to each other in the dendrogram hierarchy. We also investigate the functional role of different cancer genes as revealed by their importance in the modular organization of the network by investigating the joint distribution of their participation coefficients and their within module degree z-scores. We have identified about 36 genes as "connector hubs" occupying critical positions in the cancer network which can be potential targets for therapeutic efforts.

In the preceding chapters we had considered networks implicated in health and disease at the level of a single cell, in the next three chapters we shift our attention to disease transmission through contact networks, i.e., at the level of a society of human individuals. For such networks, one of the most important dynamical processes to understand is how epidemics are initiated and propagated. In **Chapter 4** we present our work on understanding the outbreak dynamics of the 2009 Influenza A(H1N1)v in India by estimating the initial transmissibility of the disease. This is done by analyzing the time-series data for the onset of the influenza pandemic in India during the period June 1-September 30, 2009. The novel influenza strain (later termed influenza A(H1N1)v) was first identified in Mexico in

March 2009, after which it rapidly spread to different countries. The first confirmed case in India, a passenger arriving from USA, was detected on 16 May 2009 in Hyderabad. In fact, most of the initial cases in India were passengers arriving by international flights. However, towards the end of July, the infections appeared to spread to the resident population with an increasing number of cases being reported for people who had not been abroad. To devise effective strategies for combating the spread of pandemics, it is essential to estimate their transmissibility in a reliable manner. This is generally characterized by the reproductive rate R, defined as the average number of secondary infections resulting from a single (primary) infection. A special case is the basic reproduction number R0, which is the value of R measured when the overall population is susceptible to the infection, as is the case at the initial stage of an epidemic. Using a variety of statistical fitting procedures, we have obtained a robust estimate of the exponential growth rate λ = 0.15. This corresponds to a basic reproduction number R0 = 1.45 for influenza A(H1N1)v in India, a value which lies towards the lower end of the range of values reported for different countries affected by the pandemic. We have also separately obtained estimates for different regions of the country which varied over the range 1.34-1.74. This suggests that seasonal and regional variations need to be taken into account to formulate strategies for countering the spread of the disease.

Models of epidemic propagation very often assume that populations are well-mixed (i.e., an infected individual can infect any other individual in the population with equal probability) for mathematical simplicity. However, in reality, individuals very often confine most of their interactions to members of their own social group. Thus, the contact network of individuals in a society can be considered to be modular (of which there is sufficient empirical evidence), with the members of the same module having much higher probability of being infected by each other (as a result of the increased frequency of interactions) as compared to members of different modules. In **Chapter 5** we show that contagion transmission dynamics on modular networks can have startling consequences, in particular, resulting in the persistence of highly infectious diseases. In our study, we have

considered a situation where individuals after having recovered from an infection can again become susceptible with a certain probability either as a result of loss of immunity or through removal and subsequent replacement by new individuals. Our study of this SIRS (Susceptible-Infectious- Recovered-Susceptible) epidemic model dynamics over a modular contact network suggests that under certain circumstances an epidemic can become persistently recurrent. Through numerical simulation, we show the dependence of the probability of persistence on the parameters of the network mesoscopic organization as well as on epidemic parameters such as the infection rate α. In particular, we show that highly contagious diseases (large α), which quickly die out in a population with homogeneous contact structure, can survive indefinitely (becoming endemic) when there is strong community organization in the population.

So far we have considered epidemics in which infections spread by direct contact between infected individuals. While this is a good model for several types of diseases (such as influenza or chicken-pox), there are several other important diseases which are spread indirectly via a different species acting as a vector (e.g.,Anopheles mosquitos in the case of malaria). In such situations, apart from the contact network structure we also need to consider the dependence of the vector population density on spatial geography upon which the network is embedded. In **Chapter 6** we consider the spatio-temporal dynamics of malaria transmission by first presenting an empirical analysis of epidemic data from a rural block in north Bengal. The time series data of malaria incidence for two different malaria strains (*Plasmodium falciparum* and *Plasmodium vivax*) recorded over the period Jan 2005- Feb 2009 and obtained from 51 health centers located at different regions in the block are subjected to wavelet phase analysis in order to identify travelling waves of increasing malaria incidence. This has allowed us to locate the epicenters where the outbreaks arise initially and then spread to neighboring regions. There is significant correlation between phase angle difference between epicenter with other regions and the distance of those regions from the epicenter which substantiates the travelling wave nature of the epidemic spatio-temporal transmission dynamics. The epicenters are characterized by (a)

favorable conditions for high rates of mosquito reproduction such as forest coverage, (b) relatively high human population and (c) high degree of connectedness with neighboring regions. By correlating epidemic incidence with rainfall data, we find that the latter plays a significant role in the incidence dynamics of malaria with a delay of 1-3 months. In the later portion of the chapter we have presented an investigation of the interaction of an externally imposed environmental signal (rainfall with a periodicity of 12 months) with the spatio-temporal dynamics of malaria transmission.

We conclude with a summary of the principal results reported in this thesis and indicate possible future directions of research.

2

Branched motifs enable long-range interactions in signaling networks through retrograde propagation

2.1 Introduction

The intra-cellular signaling machinery is an extremely large and complex network that is best understood in terms of interactions between *modules*, i.e., well-defined sub-networks of interacting proteins. Such modules, often associated with specific functions, are distinguished by a relative level of insulation from the activity of other molecules [28]. However, as they are connected via the network, functions of individual modules can affect that of others in various complicated ways. The resulting adaptation of response to different circumstances allows the same module to be reused in many distinct contexts. Investigating the dynamical response of a basic module to various perturbations may give us a deeper understanding of its global role in the overall functioning of the network. Such a standard signaling module, found in all eukaryotic cells, is the three component mitogen-activated protein kinase (MAPK) cascade involved in many critical cellular functions in-

cluding cell cycle control, stress response, differentiation, growth, etc. [83, 98, 99]. It is affected in many diseases including cancer, as well as, immunological and degenerative syndromes and is an important drug target [84]. The well-understood linear cascade involves the regulation by an input signal of the activity of a MAPK kinase kinase (MAP3K) that controls the activation of a MAPK kinase (MAP2K) which in turn controls the activity of a MAPK (Fig. 2.1). The end-result of MAPK activation is to initiate transcription or to stimulate the activity of other kinases [100].

However, such linear or chain-like reaction schemes imply a rigid relation between stimulus and response, precluding the possibility of the system switching to a different response for the same signal under altered circumstances. This adaptive ability is essential for survival of a cell in a noisy and dynamic environment. In fact, many linear cascades are actually part of *branched* pathways. For example, the MAP3K MEKK-1/2/3 are known to be capable of activating both JNK and ERK pathways [101]. The MAP3K MEKK-1 has also been seen to activate both the JNK and p38 pathways in the T-cell receptor signaling network [102], as well as, in the network downstream of the B-cell antigen receptor [103]. Such a design provides the cellular signaling apparatus the complexity necessary to allow integration of several signals and to regulate multiple functions at the same time [48, 100]. In particular, divergent signaling, where the activity of one molecule provides the input to multiple linear cascades allow the same external signal to generate different possible responses, the actual output being decided by the internal cellular context [100]. Such differential regulation can be achieved through reciprocal inhibition between the different branching pathways resulting in the eventual dominance of one of the possible responses. Thus, the ubiquity of branched pathway modules in the signaling network may be a consequence of the adaptability they provide to a cell in terms of making context-dependent choice between different responses.

In this chapter we show that, in addition to having more flexibility compared to linear cascades, branched pathways allow complex long-range coordination of activity even in

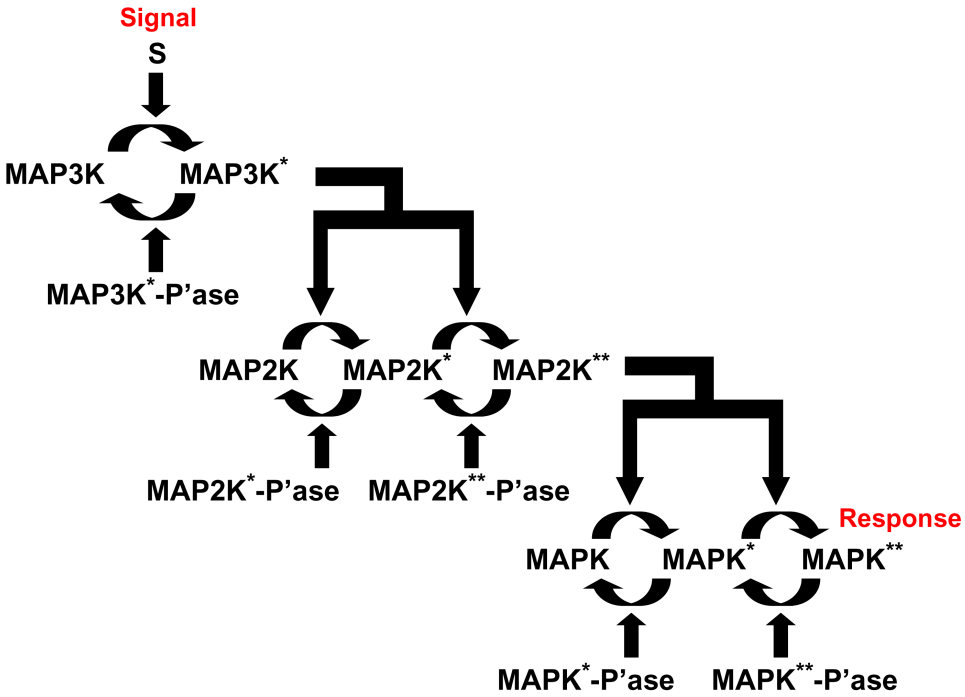

Figure 2.1: **Schematic representation of the mitogen-activated protein kinase (MAPK) cascade.** Activation through phosphorylation of the substrate kinase at each step is mediated by the upstream kinase, the single and doubly phosphorylated states of the substrate being denoted by the superscripts "*" and "**", respectively. Activated kinases are eventually dephosphorylated into their inactive forms by the corresponding phosphatases (indicated by the suffix "-P'ase"). Activity in the three-component pathway is initiated by a signal S regulating the activation of the MAPK kinase kinase (MAP3K). Activated MAP3K controls the activation of MAPK kinase (MAP2K), which in its turn regulates the activation of MAPK. Note that unlike the single phosphorylation of MAP3K, both MAP2K and MAPK require double phosphorylation in order to become active, i.e., capable of acting as the enzyme for the corresponding downstream substrate protein. The eventual response of the cascade is quantified by the concentration of activated MAPK, which can be used in initiating transcription, activating other protein molecules, etc.

the absence of any direct long-range connections (i.e., reactions) between molecules. This allows complex dynamical control in conjunction with economy of wiring, measured by the total number of different possible chemical reactions, in the intra-cellular signaling network. Restricting the total number of links is desirable in a complex system, as high connectivity in a network can reduce the specificity of the response as many different signals can elicit the same activity [104, 105]. We show that reciprocal control between parallel reaction cascades stimulated by a common signal can occur in the absence of any coupling between the reactants in the different pathways. This non-local control takes place through retrograde information propagation previously demonstrated for linear reaction cascades [106, 107], in contrast to the classic forward flow from the input signal to the response of the terminal molecular species (e.g., from MAP3K to MAP2K to MAPK). Inhibition of the terminal molecule of one pathway in such a system can initiate a series of perturbations which travel upstream all the way to the molecule at the branch-point and from there, downstream along the other parallel pathways, altering the activity of several molecules all over the network. These predictions have been experimentally verified by our collaborators at the National Centre For Cell Science (NCCS), Pune, India in macrophage cells, where blocking JNK phosphorylation is observed to bring about amplification of p38MAPK activity, and vice versa [108].

The possibility of reverse communication between components of the intracellular signaling network implies that branched pathways cannot be considered as simple open-loop circuits. Instead, retrograde propagation effectively implement closed-loop or feedback circuits, where perturbing one of the "output" elements can result in changes at the branch-point (at the "input"-end of the module), and thus, eventually to all other outputs of the module. Although the existence of implicit feedbacks in activation-deactivation reaction pathways have been shown earlier in the case of a simple linear MAPK cascade [106, 107], the consequences of combined forward and retrograde propagation of information in complex signaling network motifs have so far been unexplored. Our results reveal that even in absence of explicit long-range connections, local perturbations in

one section of a signalling network can have systems-level consequences.

2.2 Materials and Methods

Mathematical modeling: The time-evolution of the different molecular concentrations in the branched motif has been described using a set of coupled ordinary differential equations (ODEs) where each enzyme (E)-substrate kinase (SK) reaction has two steps: (i) a reversible enzyme-substrate kinase complex (ESK) formation step (with the forward and reverse reaction rates denoted by k_i and k_{-i} for the i-th reaction) and (ii) an irreversible step of product (i.e., the activated substrate SK*) formation from the complex (with a rate k_{i+1}) (Fig. 2.2). In a linear cascade, the product (activated substrate) of an earlier step can be the enzyme for the next step; thus, a MAP3K-MAP2K-MAPK reaction cascade (as in the Huang-Ferrell model [109], considering the kinase as well as phosphatase-mediated reactions) is described by 18 coupled ODEs while our model with two MAP2K-MAPK branches emerging from a common MAP3K is described by 32 coupled ODEs. For simplicity, the branches have been considered to be symmetric and the corresponding parameters in each branch are assigned the same values. The system of coupled ODEs are solved numerically using the `ode23s` routine for solving stiff equations implemented in Matlab 7, without invoking the quasi-steady-state hypothesis of Michelis-Menten kinetics [110]. The model parameters, viz., the different reaction rates and initial concentrations of the substrate molecules, are adapted from the Huang-Ferrell model [106, 109].

Analysis of robustness with respect to parameter variation: Robustness analysis for the results reported here with respect to variation in the parameters have been carried out by performing Monte Carlo simulations over $\sim 10^4$ realizations of the model system. In each realization, the 38 model parameters are randomly sampled from a biologically plausible range [106, 109] having an uniform distribution about the respective Huang-Ferrell (HF) values [109]. The branches have been considered to be symmetric with correspond-

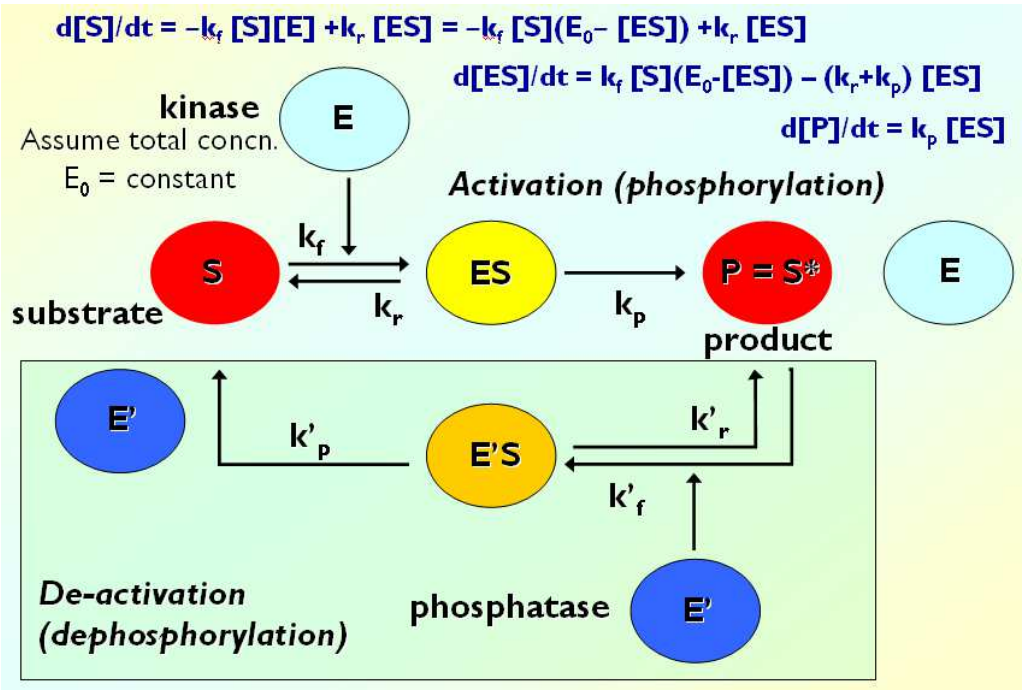

Figure 2.2: **Dynamics of kinase activation (phosphorylation) and de-activation (dephosphorylation).** The substrate kinase (S) is activated by its corresponding enzyme (E), which is the kinase located immediately upstream in the cascade. The enzyme-substrate complex (ES) formation is a reversible reaction with forward and reverse rates of k_f and k_r, respectively. The product (P) of the enzyme-substrate reaction is the phosphorylated kinase S*, which is generated by an irreversible reaction step from ES with the rate k_p. The deactivation reaction from S* to S is mediated by the phosphatase E′. The enzyme-substrate reaction dynamics can be represented by the rate equations for the variation of concentrations of S, ES and P as shown at the top of the figure.

ing parameters in each branch assigned identical values. The relative increase in the activity of $MAP2K_B$** and $MAPK_B$** on blocking $MAPK_A$ activation are measured at five different signal concentrations between [S]=10^{-12}M and 10^{-8}M (a total of 65500 random realizations). The robustness of the system response is measured as a function of the degree of variation in its parameter values quantified by the Total Parameter Variation (TPV) [111]:

$$TPV = \sum_{i=1}^{n} \left| \log_{10} \left(\frac{p_i}{p_{i,HF}} \right) \right|,$$

where p_i denotes the value of the i^{th} parameter in a given realization and $p_{i,HF}$ is the reference value of the parameter as given in the Huang-Ferrell model ($i = 1, \ldots, n$). We observe that even relatively large deviations from the HF parameter values produce results qualitatively similar to that reported in this chapter.

2.3 Results

We study in detail the model of a branched network motif shown in Fig. 2.3 and Fig. 2.4 (A): a MAP3K–$MAP2K_{A,B}$–$MAPK_{A,B}$ cascade where the MAP3K, upon activation by an input signal (stimulus) S, phosphorylates two different types of MAP2K (designated $MAP2K_A$ and $MAP2K_B$ respectively). The doubly phosphorylated MAP2Ks in turn act as enzymes for the phosphorylation of the respective MAPK molecules designated as $MAPK_A$ and $MAPK_B$ respectively. When the product formation rate of $MAPK_A$, k_8^A is suppressed, we observe noticeable changes in the phosphorylation levels of the other kinases in the system even though the affected molecule is at the downstream terminus of the cascade. This is somewhat counter-intuitive as we normally expect information to only flow "down" the cascade from MAP3K to MAPK. In contrast, here were see that information about the suppression of MAPK activity can also travel in the opposite direction, i.e., "up" the cascade from MAPK to MAP3K, a phenomenon that we term as "retrograde" propagation. In experiments, preventing MAPK phosphorylation can be realized by blocking the

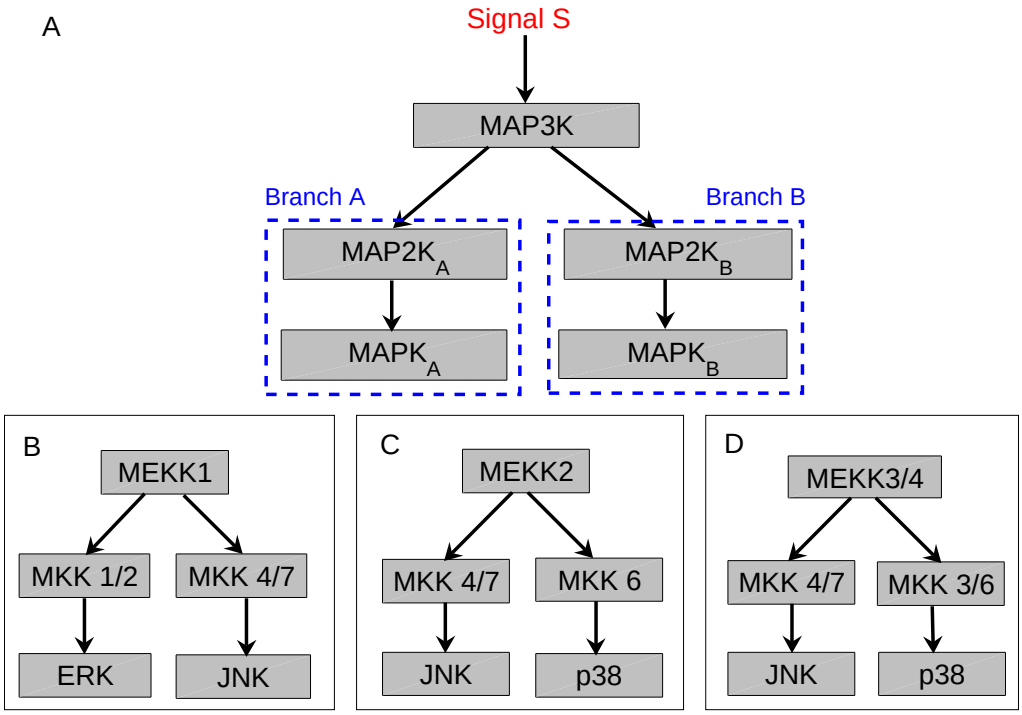

Figure 2.3: **The branched MAPK cascade network motif.** (A) A schematic diagram representing a simple branched cascade with two parallel pathways that is seen in many experimental systems, e.g., the T-cell receptor network [102]. The initial signal S activates a common MAP3K that can phosphorylate two different types of MAP2K molecules, viz., MAP2K$_A$ and MAP2K$_B$. Each MAP2K type activates a particular type of MAP kinase, MAPK$_A$ and MAPK$_B$, respectively. Specific examples of branched MAPK cascade motifs obtained from the experimental literature are shown in (B-D). They correspond to systems with the specific MAP3K (the branching point of the motif) being (B) MEKK1 that activates both MKK1/2 [112] and MKK4/7 [113, 114], (C) MEKK2 that activates both MKK4/7 [115, 116] and MKK6 [116], and (D) MEKK3/4 that activates both MKK4/7 [115–119] and MKK3/6 [116–118].

ATP binding site of the target kinase. Fig. 2.4 (A) shows the relative change in the response of the system as a result of preventing activation of $MAPK_A$ ($k_8^A = 0$), in terms of the relative increase or decrease in the steady-state concentrations of kinases in different branches compared to the unperturbed values. We observe that the perturbation reduces the activation of $MAP2K_A$ but all kinases in the unperturbed branch B, as well as, the MAP3K common to both branches, exhibit significant amplification in their phosphorylation. The modulation of kinase activity in the unperturbed branch in the absence of any direct connection between the molecules in the two branches reveals an implicit cellular mechanism for long-range communication in signaling networks through retrograde propagation. When $k_8^A = 0$, the concentration of free $MAP2K_A^{**}$ is reduced as most of it is trapped in the enzyme-substrate kinase (ESK) complex $MAP2K_A^{**}$–$MAPK_A$ from which it can be released only at the relatively slow rate of complex unbinding (k_{-7}^A). As more $MAP2K_A$ molecules are phosphorylated and end up bound in the above complex, this in turn reduces the concentration of free $MAP2K_A$. The decrease in free $MAP2K_A$ results in $MAP2K_B$ gaining relatively more access to MAP3K that enhances the phosphorylation of branch B kinases.

As the phosphorylation of the MAP3K at the branch-point is modulated by the strength of the external signal S while the dephosphorylation of $MAP3K^*$ is mediated by the concentration of its phosphatase MAP3K-Pase, we examine how different combinations of signal dose and phosphatase concentrations affect the extent of retrograde propagation in the system. Fig. 2.4 (B) shows the coupled effect of the strength of the input stimuli S and the concentration of [MAP3K-Pase] on the relative change in activity of MAP3K when the activation of $MAPK_A$ is blocked ($k_8^A = 0$). The enhanced concentration of $MAP3K^*$ results in a corresponding increase in activity of the B branch kinases, $MAP2K_B$ (Fig. 2.3 D) and $MAPK_B$ (Fig. 2.4 F). The amount of total MAP3K available also affects the relative increase in activity of the unperturbed branch as a result of the blocking of $MAPK_A$ phosphorylation. As seen in Figs. 2.4 (C) and 2.4 (E), there is an optimal range of total MAP3K concentration, [MAP3K], where the largest relative increase in the activity of $MAP2K_B$

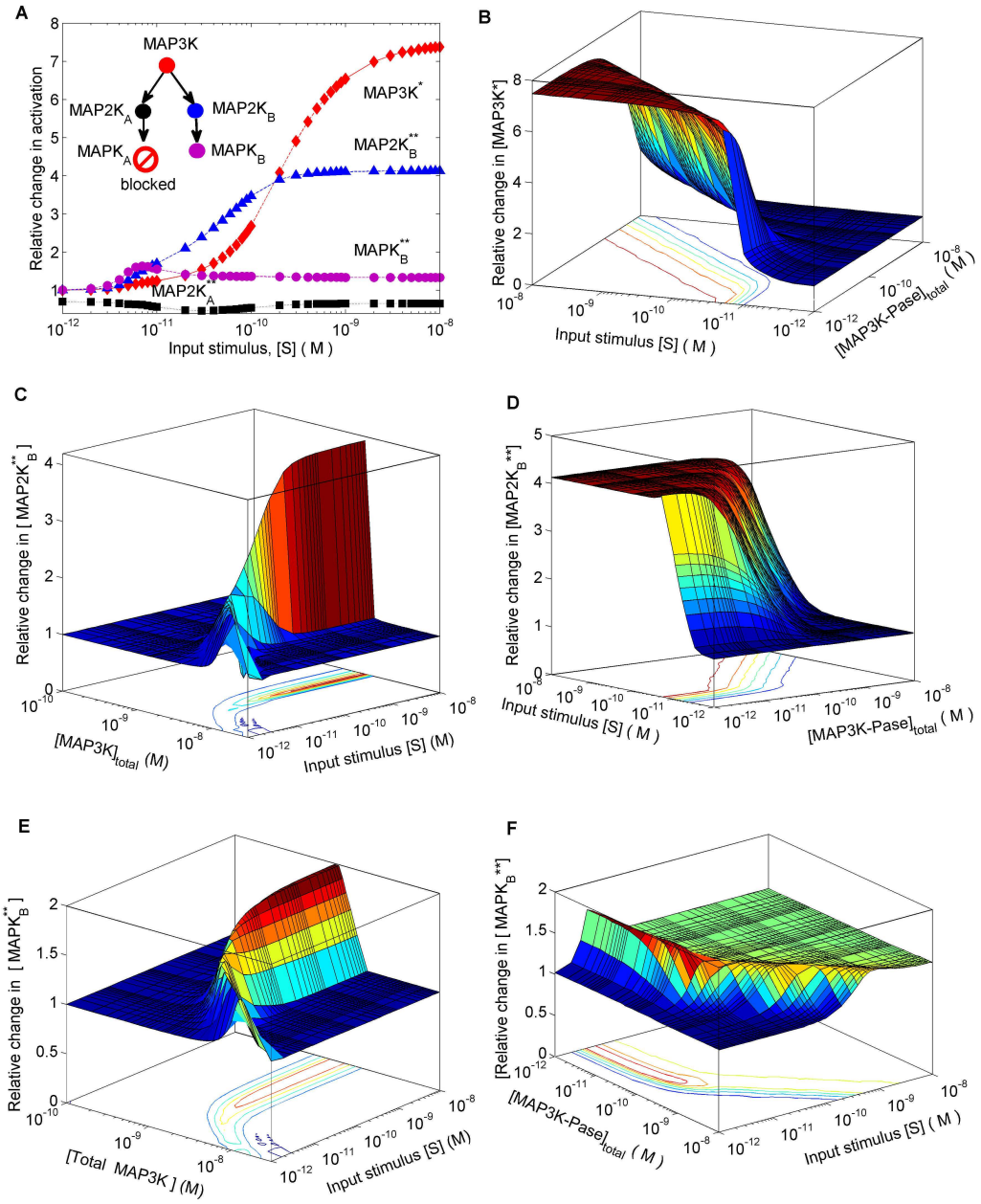

Figure 2.4: **Amplification of response in unperturbed branch through retrograde propagation of information in a branched motif.** (A) Relative change in the response of different molecular species in branches A and B as a function of signal strength, when phosphorylation of terminal kinase $MAPK_A$ in branch A is inhibited. Inset shows a schematic diagram of the branches, with the "stop" sign indicating blocking of activation of MAPK in branch A. (B–F) The perturbation results in amplification (relative to the unperturbed condition) of the concentrations of (B) $MAP3K^*$ (D) $MAP2K_B^{**}$ and (F) $MAPK_B^{**}$, shown as a function of signal strength and concentration of the phosphatase MAP3K-Pase. Relative increase in concentrations of (C) $MAP2K_B^{**}$ and (E) $MAPK_B^{**}$ as a result of the perturbation are also shown as a function of signal strength and the total concentration of MAP3K.

and $MAPK_B$ is observed for a given signal strength. In particular, when [S] is varied over 10^{-12} – 10^{-8} M, i.e., the physiologically plausible range of values, the largest relative change in [$MAP2K_B{}^{**}$] and [$MAPK_B{}^{**}$] occurs between between [MAP3K]= 10^{-9} – 10^{-8} M.

The perturbation we have discussed above corresponds to complete blocking of $MAPK_A$ phosphorylation. We also examine the effect of a *graded* perturbation where the product formation rate of the ESK complex (k_8^A) is decreased but remains finite (> 0). As the magnitude of this perturbation is increased (i.e., k_8^A is decreased to even lower values), it will approach the situation described earlier: complete absence of $MAPK_A$ activity. We note that another class of perturbations may also superficially exhibit a similar nature, viz., gradually reducing the total concentration of $MAPK_A$ which will affect the reaction flux by reducing the formation of the ESK bound complex. For this case also, as the magnitude of perturbation is increased, the concentration of activated $MAPK_A$ steadily approaches zero. These two types of interventions correspond respectively to (i) using an ATP inhibitory agent targeting the ATP binding site, and (ii) using siRNA to block the expression for the gene coding for $MAPK_A$, with both interventions having less than 100% efficiency.

Although decrease in concentration of $MAPK_A$ (Fig. 2.4 A) and lowering the product formation rate in the reaction of $MAPK_A$ with $MAP2K_A{}^{**}$ (Fig. 2.5 B) functionally serve the same purpose, viz., inhibiting the production of $MAPK_A{}^{**}$, the resulting effect on other kinases in the system as a result of retrograde information propagation are fundamentally different. While the increase (decrease) in [$MAPK_A$] leads to increase (decrease) in [$MAP2K_B{}^{**}$] and [$MAPK_B{}^{**}$] respectively (Fig. 2.5 A), the exact opposite effect is observed in the B branch kinases when product formation rate is increased (decreased) (see Fig. 2.5 B). This reveals a reciprocal response machinery of the branched cascade when it is subjected to different inhibitor types with apparently similar function (i.e., inhibiting activation of $MAPK_A$), viz., the siRNA-type inhibitor which inhibits a fraction of total

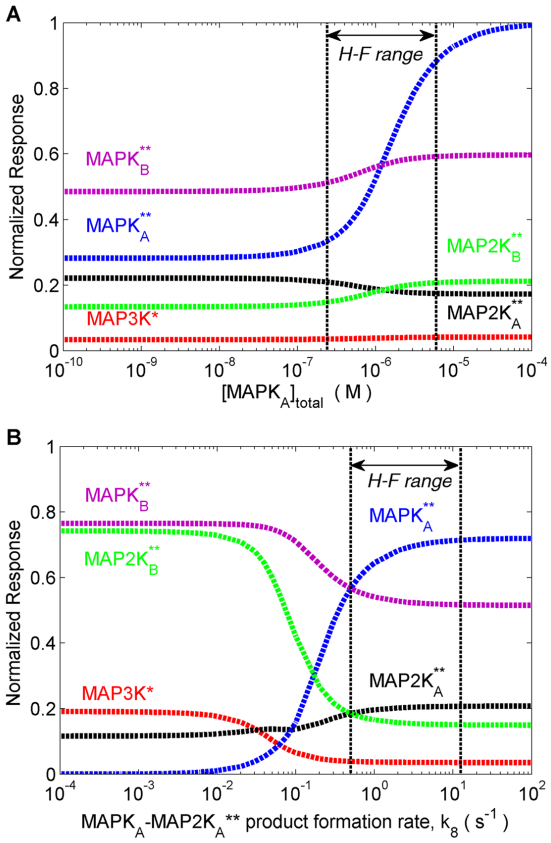

Figure 2.5: **The branched MAPK cascade responds differently to to distinct perturbations which inhibit the activation of a terminal kinase in one branch.** The normalized response, i.e., the ratio of activated to total kinase concentration, of different molecular species in the reaction cascade when (A) the total concentration of MAPK_A is varied and (B) when the product formation rate (k_8^A) of MAPK_A is varied shows opposing behavior in the activity of molecules in the unperturbed B branch, although both types of perturbations have the same functional goal of decreasing the activation of MAPK_A. As $[\text{MAPK}_A]_{total}$ is decreased, both MAPK_B and MAP2K_B decrease in activity. However, when the product formation rate of MAPK_A is decreased, the activity of both MAPK_B and MAP2K_B are increased. The activity of the kinase MAP3K which forms the branch-point also shows different response to the two perturbations unlike the result for single-branch cascades: in (A), the activity is unchanged, whereas in (B), the activity of MAP3K increases, on decreasing the activation of MAPK_A. Note that the curves corresponding to MAPK_A and MAPK_B intersect in (A) when $[\text{MAPK}_A]_{total} = [\text{MAPK}_B]_{total} = 1.2\ \mu$M and they intersect in (B) when the product formation rates for MAPK_A and MAPK_B, i.e., k_8^A and k_8^B respectively, are both 0.5 s^{-1}. Except for $[\text{MAPK}_A]_{total}$ and k_8^A, all other total molecular concentrations and reaction rates are kept fixed at the HF values [109]. The broken lines indicate the physiologically plausible range of values for $[\text{MAPK}_A]_{total}$ in (A) and k_8^A in (B), as used in the Huang-Ferrell model [109].

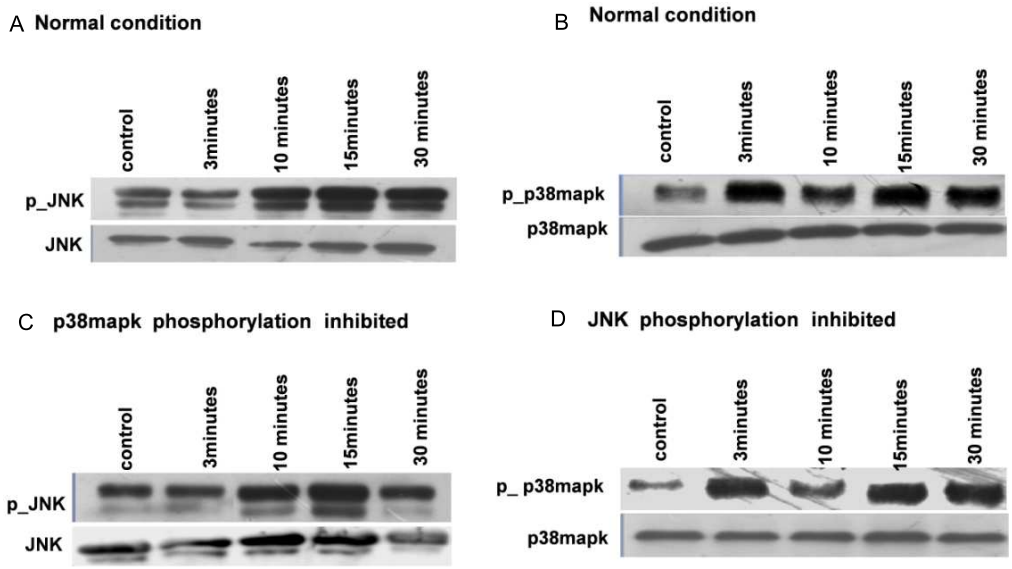

Figure 2.6: **Experimental validation of the amplification of activity in one branch on inhibiting the activity of the terminal kinase in the other branch.** Western blots of JNK and p38MAPK phosphorylation in primary macrophages when subjected to CD40 stimuli of strength 3 μg/ml are shown under normal (JNK: A, p38MAPK: B) and perturbed (JNK: C, p38MAPK: D) conditions. Perturbation is applied by inhibiting the phosphorylation of either p38MAPK (C) or JNK (D) by using pharmaceutical agents. In each figure, the upper and lower panels show the phosphorylated and total concentrations of the different molecular species (represented by “p_ JNK" and “JNK", and “p_ p38" and “p38" for JNK and p38MAPK respectively). The **control** condition shows an unstimulated system, while the times 3, 10, 15 and 30 minutes refer to observations after the system has been exposed to stimuli of the corresponding duration. All experiments have been carried out by our collaborators at NCCS in triplicate and a representative set of blots are shown. The densitometric analysis of the blots is indicated in Table 2.1.

concentration of $MAPK_A$ and the pharmaceutical inhibitor which blocks the ATP binding site of a fraction of total $MAPK_A$. The simulations reveal that while concentration depletion of $MAPK_A$ results in depletion of [$MAP2K_B{}^{**}$] and [$MAPK_B{}^{**}$] (Fig. 2.5 A), suppressing the phosphorylation of $MAPK_A$ (Fig. 2.5 B) amplified the concentrations of the B branch kinases. Subsequently, experiments have been conducted by our collaborators at NCCS on a two-branch cascade comprising JNK and p38MAPK molecules, observing their phosphorylation in unperturbed and perturbed conditions. The JNK and p38MAPK are the terminal MAPK molecules belonging to two distinct pathways which share a common branch-point MAP3K molecule MEKK3/4 [116, 117]. In the experiments, stimulation of CD40 receptor (that acts as input to both JNK and p38MAPK pathways [120]) in macrophage cells from BALB/c mice by a ligand dose leads to JNK (Fig. 2.6 A) and p38MAPK phosphorylation (Fig. 2.6 B) in comparable magnitudes. The system is next subjected to ATP blockers that inhibit phosphorylation of either p38MAPK (Fig. 2.6 C) or JNK (Fig. 2.6 D). The densitometric analysis of the western blots (Figs. 2.6 A–D) in Table 2.1 shows the relative increase in phosphorylation of MAPK for the perturbed and unperturbed conditions.

Table 2.1: **Densitometric analysis of western blots for JNK and p38MAPK under normal and perturbed conditions at different time points.**

	Normal Condition						p38MAPK phosphorylation inhibited			JNK phosphorylation inhibited		
	$[JNK^{*}] + [JNK^{**}]$	$[JNK]_{total}$	r_{JNK}	$[p38^{*}] + [p38^{**}]$	$[p38]_{total}$	r_{p38}	$[JNK^{*}] + [JNK^{**}]$	$[JNK]_{total}$	r_{JNK}	$[p38^{*}] + [p38^{**}]$	$[p38]_{total}$	r_{p38}
Control	1	1	1	1	1	1	1	1	1	1	1	1
3 min	0.79	1.2	0.66	1.92	1.05	1.84	1.25	0.87	1.43	4.11	0.87	4.7
10 min	1.44	0.74	1.95	1.46	1.07	1.36	1.64	1.26	1.3	2.27	0.97	2.34
15 min	1.62	1.28	1.27	1.84	0.98	1.89	2.08	1.17	1.78	4.12	1.24	3.34
30 min	1.52	1.51	1.01	1.71	0.9	1.91	1.12	0.71	1.57	4.56	1.05	4.34

The observations show the change in activity as a function of exposure to stimuli applied for different durations. The perturbations correspond to inhibiting p38MAPK phosphorylation by a pharmaceutical agent followed by measurement of JNK activity, and conversely, inhibiting JNK phosphorylation by a pharmaceutical agent followed by measurement of p38 activity. The column headings $[JNK]_{total}$ and $[p38]_{total}$ refer to the total concentrations of the respective kinase molecules, $[JNK^{*}]+[JNK^{**}]$ and $[p38^{*}]+[p38^{**}]$ correspond to the concentrations of the respective (single or double) phosphorylated forms, while r_{JNK} and r_{p38} represent the ratio of phosphorylated kinase to total kinase concentrations. *Control* refers to the unstimulated system, while the different times (viz., 3 minutes, 10 minutes, 15 minutes and 30 minutes) correspond to the duration for which the stimulus is applied.

The experiments validate the key prediction from our model: blocking the phosphorylation of $MAPK_A$ by targeting the ATP binding site enhances the phosphorylation of $MAPK_B$ (Table 2.1). As there is no experimental evidence of explicit feedback regulation between JNK and p38MAPK, it is extremely likely that the observed modulation of phosphorylation amplitudes of JNK and p38MAPK emerges as a result of long-range coordination through retrograde information propagation. The experiments additionally show that under perturbation, the retrograde propagation from JNK to p38MAPK is significantly higher than from p38MAPK to JNK (Table 2.1). Plausible reasons for this could lie in the asymmetry of the molecular concentration values and/or the reaction parameters in the two branches. Thus, we next analyze the effect of such asymmetry on the retrograde propagation of information.

We observe that when the ESK complex binding rates (k_3^A, k_5^A and k_7^A) in the perturbed branch A are much higher (viz., by a factor of 10) than those in the unperturbed branch B (k_3^B, k_5^B and k_7^B), there is a remarkable increase in the activity of the latter branch, e.g., by three orders of magnitude in [$MAPK_B{}^{**}$], over a certain range of signal strength S (Fig. 2.7 A) as compared to the opposite situation, i.e., when the binding rates of the unperturbed branch are much higher than those of the perturbed branch (Fig. 2.7 B; see also Fig. 2.8). Also, when the total concentrations of MAPK and MAP2K in the perturbed branch A are much larger (viz., by a factor of 10) than the corresponding quantities in the unperturbed branch B, on blocking activation of $MAPK_A$ there is an observable increase in the activity of the unperturbed branch, e.g., about 20-fold increase in [$MAPK_B{}^{**}$] over a range of input stimuli strength (Fig. 2.7 C), compared to the case when the total concentrations of MAPK and MAP2K in the unperturbed branch are higher than the perturbed branch (Fig. 2.7 D). These results indicate that the extent of retrograde information propagation seen in a signaling cascade will depend on the magnitude of forward reaction flux (i.e., the strength of the flow "down" the cascade from MAP3K to MAPK) in the different branches of the system. If the branch having a larger downstream flux (resulting from relatively higher binding rates or larger total concentrations of the molecules belonging

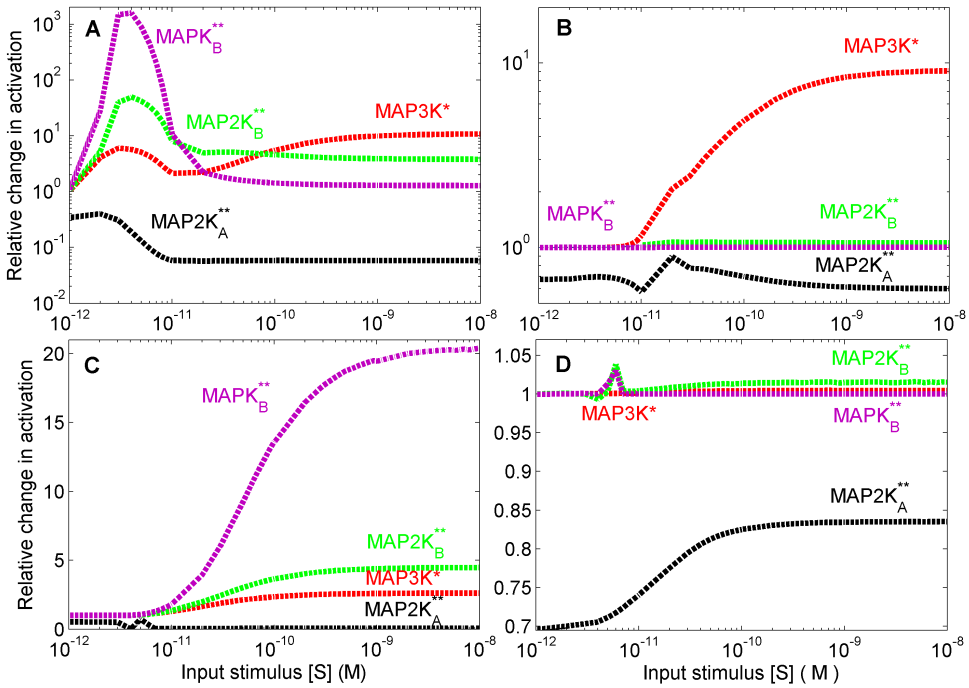

Figure 2.7: **Role of branch asymmetry.** Relative increase in response as a function of signal concentration on blocking MAPK_A activation when the binding reaction rates for (A) branch A (k_3^A, k_5^A and k_7^A) or (B) branch B (k_3^B, k_5^B and k_7^B) are 10 times higher than those in the other branch, and, when the total concentrations of MAPK and MAP2K in (C) branch A or (D) branch B are 10 times higher than the corresponding values for the other branch (= mean value in the Huang-Ferrell range). Note that in (A) the MAPK_B activation can increase by more than 1000 times for a particular range of signal strength. In contrast, there is relatively little change in the activity of the two branches when the activation of the terminal kinase for the branch having lower values of reaction rates is blocked.

to that branch) is perturbed by inhibiting the activation of its terminal kinase, this will result in a much larger proportional increase in access to MAP3K for molecules in the unperturbed branch. On the other hand, if the branch having a lower magnitude of forward reaction flux is perturbed, the resulting increase in the activation of the unperturbed branch will be marginal as compared to the already higher activation levels of this branch in the control condition. This is further corroborated by the effect of asymmetry in other parameters of the perturbed and unperturbed branches, viz., (i) the product formation rates in the activation enzyme-substrate reactions, (ii) the binding rates with phosphatases in the deactivation reactions and (iii) the total phosphatase concentrations (Fig. 2.9). Thus, the experimentally observed asymmetry in the response of JNK and p38MAPK (Fig. 2.6) can be explained as a result of the differences in the reaction parameters or total concentrations of components of the two pathways.

Previous demonstrations of implicit feedback in linear signaling pathways had identified sequestration as the key mechanism [107, 121]. However, in the complex branched network motif investigated here, there are additional effects contributing to the retrograde propagation of information. In particular, we note the presence of competitive inhibition, e.g., through competition between singly phosphorylated and unphosphorylated forms of a substrate molecule (e.g., $MAPK_A{}^*$ and $MAPK_A$) for the common kinase enzyme (e.g., $MAP2K_A{}^{**}$). We have investigated the contribution of such competition in producing retrograde propagation by comparing the system with an artificial cascade model that allows only single phosphorylation of $MAPK_{A,B}$ and $MAP2K_{A,B}$ so that competitive inhibition is absent. We observe that the magnitude of retrograde propagation (and hence the amplification of kinase activation in the unperturbed branch) for the system with singly phosphorylated MAP2K and MAPK is significantly reduced (Fig. 2.10 A) compared to the original branched motif where the corresponding kinase molecules are doubly phosphorylated (Fig. 2.4 A). Retrograde propagation is also perceptibly weaker in model systems with reduced competitive inhibition, viz., where either (i) only MAP2K is singly phosphorylated while MAPK is doubly phosphorylated (Fig. 2.10 B), or (ii) only MAPK is

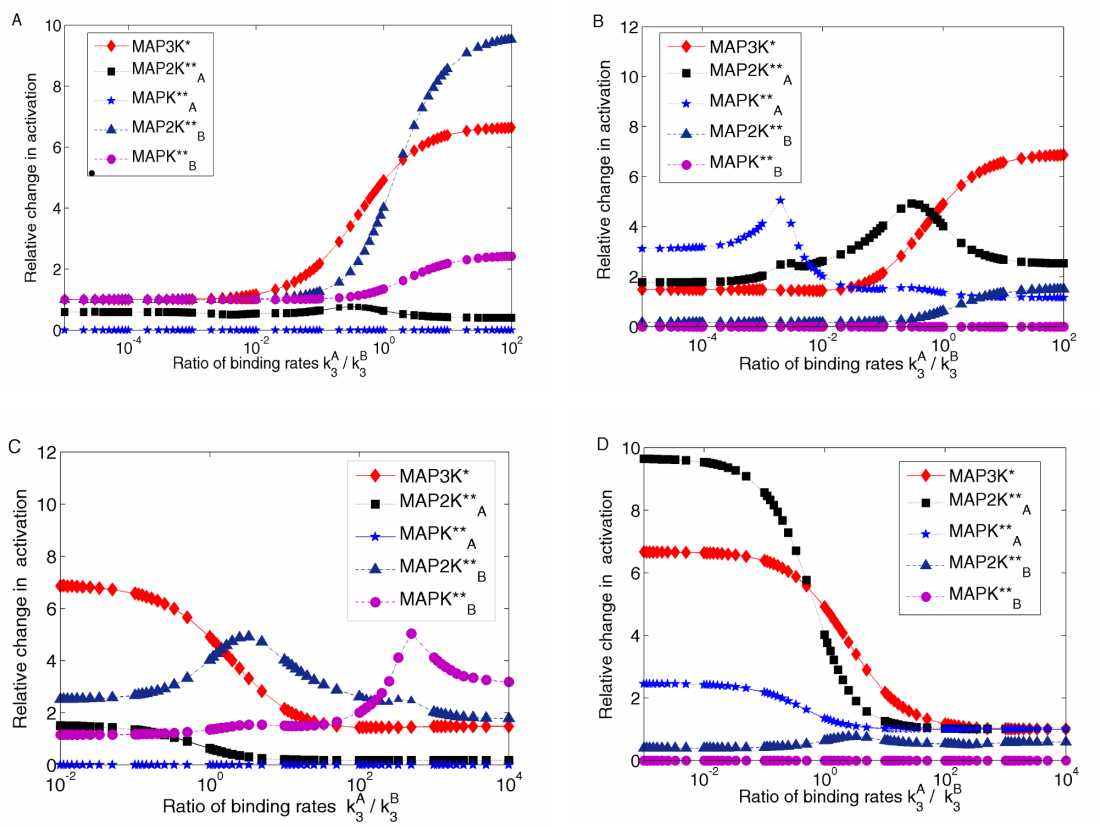

Figure 2.8: **Role of competition between the two branches in binding to MAP3K*.** Relative increase of response as a function of the ratio of the binding rates of MAP2K$_A$ ($k_3^A = 1.67\times10^7$ held constant) and MAP2K$_B$ (k_3^B) with MAP3K* on blocking (A) MAPK$_A$ activation and (B) MAPK$_B$ activation. Relative increase of response as a function of the ratio of the binding rates of MAP2K$_A$ (k_3^A) and MAP2K$_B$ ($k_3^B = 1.67 \times 10^7$ held constant) with MAP3K* on blocking (C) MAPK$_A$ activation and (D) MAPK$_B$ activation.

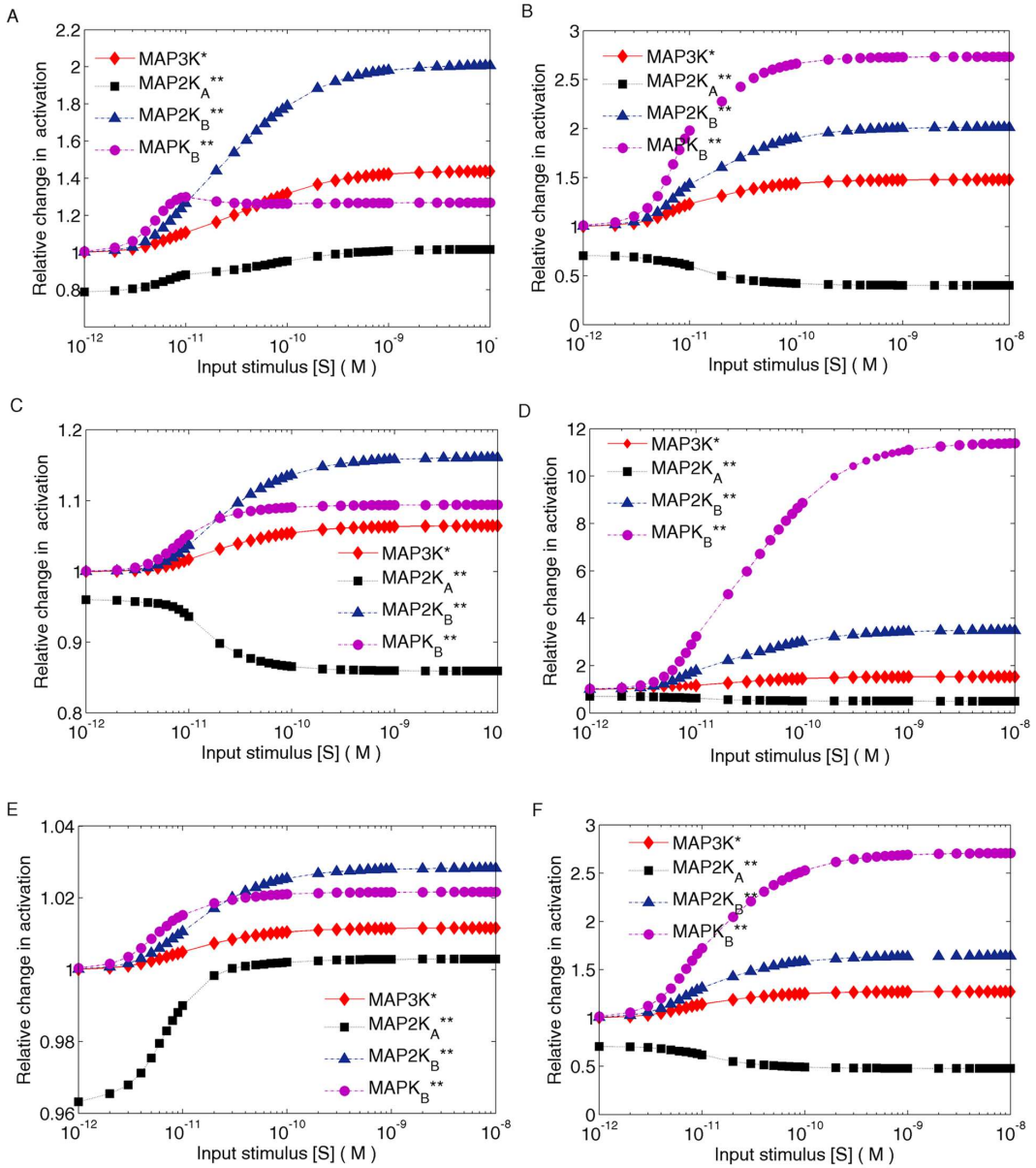

Figure 2.9: **Role of asymmetry for reaction parameters in the two branches.** Relative increase of response as a function of the signal on blocking $MAPK_A$ activation when (A) the product formation rates for branch A (k_4^A and k_6^A) are 5 times higher than those in branch B and (B) the product formation rates for branch B (k_4^B and k_6^B) are 5 times higher than those in branch A; (C) the binding reaction rates for branch A (kp_3^A and kp_5^A) are 10 times higher than those in branch B; (D) the binding reaction rates for branch B (kp_3^B and kp_5^B) are 10 times higher than those in branch A; (E) the total concentrations of phosphatase of $MAP2K^*$ and $MAP2K^{**}$ of branch A are 10 times larger than the corresponding values for branch B (=mean value in the Huang-Ferrell range) and (F) the total concentrations of phosphatase of $MAP2K^*$ and $MAP2K^{**}$ of branch B are 10 times larger than the corresponding values for branch A (=mean value in the Huang-Ferrell range).

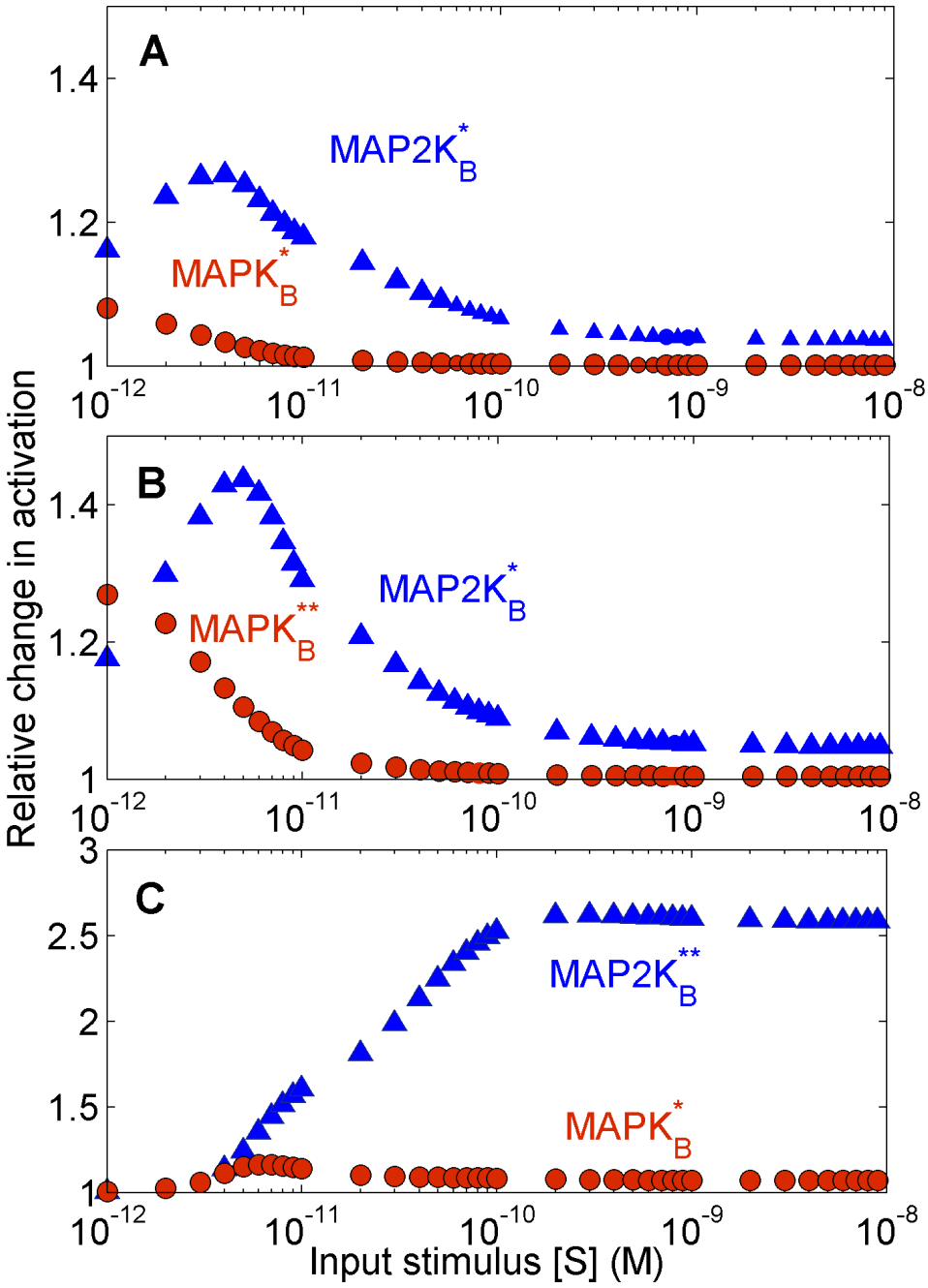

Figure 2.10: **Role of competitive inhibition.** Relative increase with stimulus strength of the steady-state response of different molecular species in the unperturbed branch B ($MAP2K_B$ and $MAPK_B$) shown as a function of signal concentration, when phosphorylation of $MAPK_A$ is prevented. In (A) both MAP2K and MAPK are *singly* phosphorylated, while in (B) MAP2K are *singly* phosphorylated but MAPK are *doubly* phosphorylated. Note that there is a small increase in the steady-state response of $MAP2K_B$ and $MAPK_B$ in (B) compared to (A). (C) When MAP2K are *doubly* phosphorylated whereas MAPK are *singly* phosphorylated, the relative increase in steady-state activity of $MAP2K_B$ and $MAPK_B$ is more prominent.

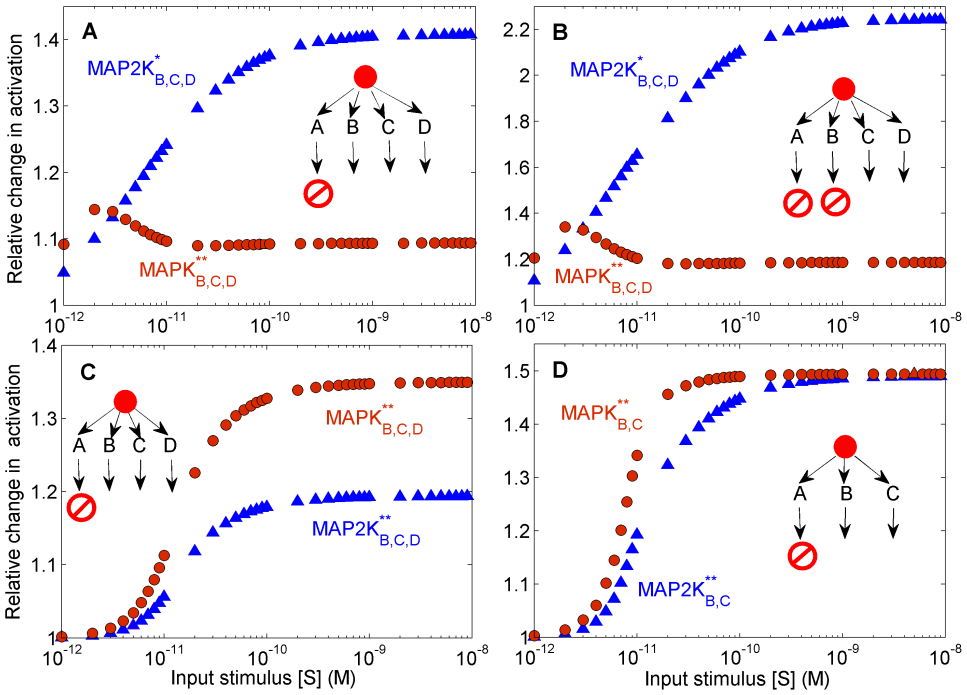

Figure 2.11: **Effect of multiple branches.** (A–B) Relative increase in steady-state response in a *four*-branch cascade which allows only *single* phosphorylation of MAP2K while MAPK is *doubly* phosphorylated as a function of stimulus strength for the cases (A) when $MAPK_A$ phosphorylation is prevented and (B) when phosphorylation of in two branches, i.e., of both $MAPK_A$ and $MAPK_B$, are blocked. The effect of having four branches but with singly phosphorylated MAP2K is similar to having two branches with double phosphorylated MAP2K but with a lower relative change in the response. This suggests that competitive inhibition is playing a role but is not solely responsible for the retrograde propagation in the branched cascade model. (C–D) Relative increase in steady-state response of *doubly* phosphorylated MAP2K and MAPK in the unperturbed branches as a function of signal strength when phosphorylation of $MAPK_A$ is prevented in a (C) a *four*-branch and (D) a *three*-branch cascade. As the number of branches are increased, the relative change in MAP2K activity on perturbation decreases faster than that of MAPK.

singly phosphorylated while MAP2K is doubly phosphorylated (Fig. 2.10 C); however both exhibit higher degree of amplification of kinase activity in the unperturbed branch as compared to the situation when there is no competitive inhibition (Fig. 2.10 A).

Thus, sequestration effects inherent to the cascade reaction mechanism, as well as, competition between multiple substrates (e.g., $MAP2K_A$, $MAP2K_B$, $MAP2K_A{}^*$ and $MAP2K_B{}^*$) for the same enzyme ($MAP3K^*$), both contribute to the magnitude of retrograde propagation seen in a branched network. If these are the only factors affecting the degree of amplification of kinase activity in the unperturbed branch then the results of the original two-

branch system should be reproducible in a hypothetical four-branch model network with doubly phosphorylated MAPK but only *singly* phosphorylated MAP2K (Figs. 2.11 A–B). When a single branch in such a system is perturbed (e.g., by inhibiting the phosphorylation of $MAPK_A$), the effect on the unperturbed branches is relatively less compared to the original system where the MAP2K molecules are doubly phosphorylated (Fig. 2.4 A). In the former situation, the four competing substrates ($MAP2K_{A,B,C,D}$) are all present initially, whereas, in the original network model (as well as in the experimental system comprising JNK and p38MAPK pathways), two of the competing substrates ($MAP2K^{*}_{A,B}$) are the reaction products of the other two competing substrates ($MAP2K_{A,B}$) and are not present initially. Thus, in the latter case, the rise in concentrations of (and hence, the resulting competition from) two of the competing substrates has a time-delay with respect to the remaining two, making this system fundamentally different from the four-branch cascade. Further, double phosphorylation of $MAP2K_{A,B}$ introduces an additional delay in the activation dynamics of the downstream MAPKs.

Based on these results, the extent of retrograde propagation of information in a branched network structure is seen to depend on multiple factors, viz.,: (i) the competition between branches for a common enzyme at the branch-point, (ii) sequestration of a kinase through binding in an ESK complex [121] and (iii) competitive inhibition between the un-phosphorylated and singly phosphorylated forms of the same kinase (capable of double phosphorylation) for its enzyme [122]. Fig. 2.11 (B) shows the relative increase of response in a four-branch network where the MAP2K molecules are phosphorylated only at a single site while the MAPK molecules are doubly phosphorylated, when the activation of MAPK in two branches (A and B) is prevented. For a range of strengths of the input stimulus, it is seen that the retrograde flow of information on blocking activation in two branches is higher as compared to blocking only one branch (compare Figs. 2.11 A and B). On the other hand, we observe that allowing double phosphorylation of MAP2K molecules (Fig. 2.11 C) results in a stronger response (compared to allowing only single phosphorylation of MAP2K as in Fig. 2.11 A) in the unperturbed branches (B,C,D) when

activation of MAPK in one branch (A) is blocked. Note that in a four-branch cascade with double phosphorylation of MAPK and MAP2K, the relative increase in activity resulting from the perturbation is higher for MAPK as compared to MAP2K in the unperturbed branches (which is opposite to the situation observed in a two-branch network). However, when the number of branches is reduced from four to three (Fig. 2.11 D) the relative increase of activity for MAPK and MAP2K in the unperturbed branches (B,C) become comparable. Thus, comparing Figs. 2.11 (C–D) and 2.4 (A), we note that as the number of branches increase, on inhibiting the activation of the terminal molecule (MAPK) in one of the branches, the resulting relative increase in activity of the MAP2K molecules in the unperturbed branches is reduced, while that of the corresponding MAPK molecules remains comparatively unchanged.

In a biological setting, the concentrations of kinases and phosphatases are usually of comparable magnitude and the systems are exposed to a wide range of signal strengths [87, 121]; our results are valid under these realistic conditions. We also stress that our results have been obtained by simulating the full dynamical model without using quasi-steady-state assumptions that focus exclusively on steady-state behavior. Such approximations ignore rapid transient changes in the concentrations of signaling molecules and do not reproduce the effect of feedback interactions [110, 123]. We also note that critical biological properties should be robust against parameter changes [111], caused by variations in the environment, polymorphisms or mutations [124] that can influence not only a single parameter but many of them simultaneously [106]. Thus, if retrograde propagation of information is indeed expected to play a significant role in intracellular signaling it should be robust. We have established the robustness of our observations by verifying that the results are not sensitively dependent on system parameters (Fig. 2.12).

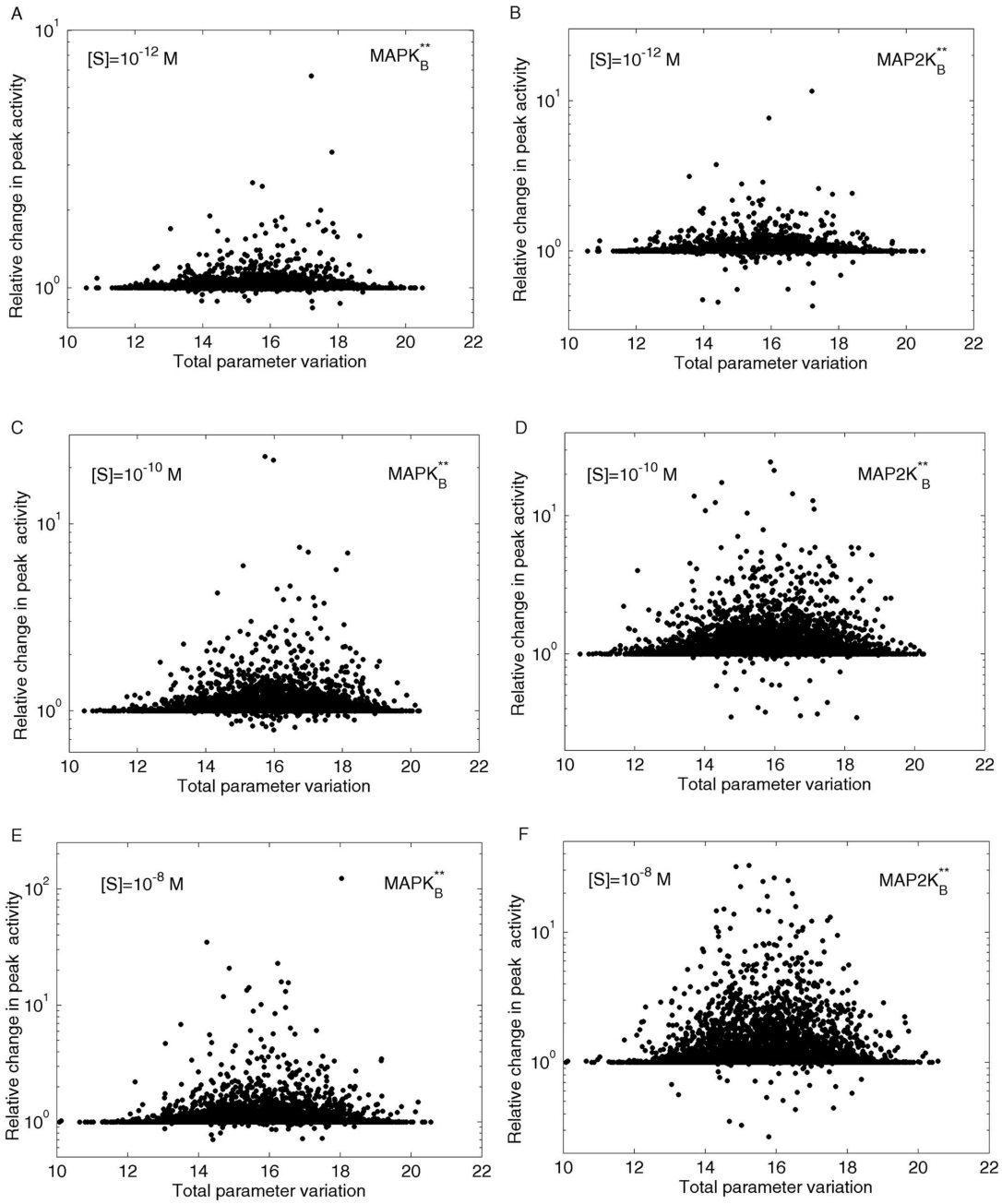

Figure 2.12: **Robustness of the effect of retrograde propagation in branched motifs with respect to parameter variation.** The variation in the relative increase in response, i.e., concentrations of $MAPK_B^{**}$ and $MAP2K_B^{**}$ on blocking the phosphorylation of $MAPK_A$, measured in terms of Total Parameter Variation (TPV) on randomly varying the 38 parameters in the model. The dots in each figure indicate the individual values obtained from 10^4 realizations for three different signal strengths: (a-b) $[S_0] = 10^{-12}$ M, (c-d) $[S_0] = 10^{-10}$ M and (e-f) $[S_0] = 10^{-8}$ M. The 38 parameters, which include total concentrations and reaction rates for all kinases and phosphatases at a given branch (the corresponding values for the other branch are taken to be the same) are randomly chosen from uniform distributions bounded between physiologically plausible minimum and maximum values for the parameters.

2.4 Discussion and Conclusion

Our demonstration that local intervention in a signaling network can have remarkable non-local consequences has implications for understanding how the intricate machinery of information processing functions in the cell. In particular, reciprocal inhibition between parallel pathways can occur without the involved agents directly reacting with each other. This allows a high level of adaptability in control without a concomitant increase in the wiring complexity in the network, leading to a more efficient system design. Long-range effects also assume importance in light of the current experimental paradigm in systems biology where the observation of up- or down-regulation of activity for a molecule as a result of perturbing another molecule is assumed to indicate the existence of a direct interaction between them [125]. While interconnections in a signaling network are often inferred on the basis of such observations, our results show that dynamical correlations between molecular activities may have a fundamentally different origin. Note that although we have used the MAPK module to illustrate the mechanism of retrograde propagation, it is conceivable that branching in the signaling network can occur upstream of MAP3K resulting in indirect communication over even larger distances in the system.

The results reported here have potential significance for designing drugs against systems-level diseases such as cancer that proliferate through complex orchestration of multiple signaling molecules [86, 126, 127]. Such diseases are multi-factorial and may have multiple possible targets for pharmaceutical intervention. Conversely, many drugs that are used for treatment may be working through as yet undetermined mechanisms, so that even if they have a known target, there can be potentially undesirable 'off-target' effects [128]. From the point of view of drug design for systems-level diseases, the crucial implication of the results reported here is that the effect of competitive blocking of a ATP binding site with a pharmaceutical agent may be completely different from the effect of suppressing the kinase expression by siRNA mediated inhibition. This is critical when one considers the 'off-target' effect of such intervention on the activity of other molecules in the sys-

tem which may also be playing a crucial role in the disease. In fact, these two methods are classically assumed to have the same effect on the system and siRNA experiments are often used to validate the model driven hypothesis suggested by experiments involving pharmaceutical inhibitors. Thus, the physiological impact of retrograde information propagation is realized with dramatic effect in branched signaling cascades as the multiple pathways show strikingly different response to apparently similar interventions.

We can, for example, consider the proteins JNK and p38MAPK, both of which are now established to be intimately involved in the proliferation of cancer and its cure. However, the respective functions of these two molecules in cancer development are not well-understood and their contribution to genesis and propagation of cancer may sometimes appear to be contradictory [85]. Certain cells use these signalling pathways to oppose cell proliferation and morphological transformation, whereas cancer cells can subvert these pathways to facilitate proliferation, survival and invasion. For example, while the JNK and p38MAPK can both act as pro-apoptotic pathways that may help cure cancer, they have also been found to function as oncoproteins that help cancer cells survive [85]. Hence depending on the cellular context, the JNK and p38MAPK pathways could be used by the cell either to deliver complementary outputs, or, to trigger antagonistic cells fates [85]. Under these circumstances, drug design for such diseases should take into consideration the system-level consequence of inhibiting a particular kinase on the activity of other kinases. This can have consequences for drug therapy as one can control the activity of a molecular species with a pharmaceutical agent that interacts with a different molecule, provided the two species are indirectly related by the long-range control mechanism described here.

Complexity of a signaling network [48, 104] may be increased further through inter-modular cross-talk [105]. In principle, there can be multiple inputs impinging on a signaling motif, as well as, interference between signals traveling through different pathways activated by the same receptor. For example, it is observed that the kinase Cot can activate

the ERK MAPK independent of the corresponding MAP3K (Raf) [129]. This implies that there can be additional inputs to MAP2K, apart from its usual MAP3K. For such a situation in the branched module discussed here, retrograde propagation will affect all the inputs, its magnitude varying according to the strength of crosstalk between the additional input and the branched MAPK cascade. Thus, the perturbation of a terminal kinase will not only affect the immediate module of which it is part, but through the interaction of the module with other inputs the retrograde propagation of information can connect apparently remote and unrelated modules of the network, e.g., through Cot the perturbation of MAPK may eventually affect regulation of IkB kinase [130].

We have shown here that different types of perturbations having the same objective (e.g., both siRNA inhibitor as well as ATP binding blocker aims at reducing MAPK activity), while having similar consequences in a linear signaling cascade, can give rise to very different results in a branched network. Such distinct responses to apparently similar perturbations may be crucial when dealing with co-regulated diseases. For example, JNK signaling is enhanced and p38MAPK signaling is abrogated in different cancers [85,131]. On the other hand, both JNK and p38MAPK signaling promotes diabetes by negatively regulating the function of insulin receptor [132,133]. Co-diagnosis of both cancer and diabetes in the same individual is not uncommon, although the system level causality behind such occurrence is not understood [134]. More importantly, the protein IRS1 (insulin receptor substrate 1) which has a critical role in insulin-signaling pathways and whose mutation is known to result in genetic risk for type 2 diabetes [135], has been identified as a drug target for cancer [136]. We can qualitatively argue from our analysis that pharmaceutical intervention designed to inhibit JNK phosphorylation for suppressing cancer might result in prolonged enhancement of p38MAPK signaling which could consequently exceed a cellular threshold, worsening any pre-existing diabetic condition. Thus, while an ATP blocker-type drug can be used to inhibit JNK phosphorylation, if the resulting increase in p38MAPK phosphorylation is undesirable, a better option is to use an inhibitor reducing JNK expression. Such insights on the differential effects of drugs designed for

the same disease is an important outcome of studying how the local dynamical properties of modules (such as the branched cascade motif) can affect the function of the larger signaling networks in which they are embedded [123].

Our results may also be used to understand the evolutionary advantage of intracellular pathogens which only target molecules in one branch of parallel MAPK pathways but which can nevertheless modulate the cellular response to its own advantage. These pathogens often target the MAPK module as the upstream signaling intermediates converge to MAPK for final integration of the signal deciding the cellular responses. As a result, the pathogens do not need to devise extremely complicated interception strategies targeting many types of molecules in order to survive within the host. As the signal strength cannot be adjusted beyond the MAPK module, the pathogen strategy would be a winning one. The retrograde propagation capability of a branched motif described here can explain how host cell signaling can still be adjusted to maintain the MAPK-dependent cellular functions.

3

Mesoscopic organization of cancer gene network

3.1 Introduction

Cancer is the collective name of a group of diseases characterized by unconstrained cell growth which have significant mortality and high public health cost. In developed countries where life expectancy has increased and "diseases of poverty" such as tuberculosis have largely been controlled, cancer is one of the leading causes of death [137]. Despite years of sustained research efforts cancer is still untamed [138]. This is at least in part because cancer is a "systems disease" [139], i.e., it cannot be completely understood without considering the network of interactions between a number of different elements. It is therefore unlikely that it can be cured by treating a single cause.

The difficulty of investigating cancer as a network disease is in the large number of elements that are involved and the myriad ways in which they interact. A possible approach to this complex disease is to compartmentalize the entire system of interactions into sub-networks that are easier to analyze by exploiting the modular nature of biologi-

cal networks [28, 123]. By focusing on the modules of networks related to cancer and the interactions between them, it is possible to use a mesoscopic approach for understanding the system-level aspect of cancer without getting mired in the complexity of the large number of molecules and interactions involved.

In this chapter we have reconstructed a cancer gene network and a cancer category/tumor type network using a comprehensive database relating different categories of cancers and types of tumors with mutations of specific genes. This is done by taking projections from the bipartite network that connects nodes representing genes with the nodes representing the types of tumor in which mutations of those genes have been implicated. Using community detection algorithms, we identified several modules in these networks. By classifying genes in terms of their connectivity to members of their own module and to members of different modules, we identify their functional importance in the cancer network. We show that genes playing the role of connector hubs and global hubs, i.e., having high connectivity with other modules in addition to genes belonging to their own module, have a disproportionately high representation in the human signaling pathways related to cancer. Therefore, these genes can be identified as potential targets for therapeutic intervention. The importance of connector and global hubs is further underlined by observing that nodes having these roles in the protein-protein interaction network have an extremely high probability of being related to cancer compared to the corresponding probability for a randomly chosen protein. Finally, we show that genes that are connector hubs are associated with diseases that have a much lower survival rate than others, pointing to the critical positions they occupy in the cancer network.

3.2 Materials and Methods

Connectivity data

For constructing the networks analyzed here we have used information on the associa-

tion of 927 cancer-related genes with 35 different cancer categories and 146 tumor types, obtained from the F-Census or functional census database of human cancer genes [140]. These are derived from high-throughput mutational screens of cancer genomes collected from various sources including Cancer Gene Census (CGC) [35] and Online Mendelian Inheritance in Man (OMIM) [141]. For constructing the protein-protein interaction network we have used data for interactions between different proteins obtained from the Human Protein Reference Database (HPRD) [142]. The data set we consider comprise 9645 proteins whose inter-connections have been identified using yeast two-hybrid analysis, in vitro or in vivo methods [143].

Network Randomization

Ensembles of randomized versions of the empirical networks has been constructed using two different methods: (i) by degree preserved randomization of links [144] (10^6 links randomly swapped) with link weights remaining associated with the swapped connections and (ii) strength preserved link randomization [145] of links (10^6 randomizations) by adjusting weights according to $w_{ij} \rightarrow w_{ij} + \delta_i(\frac{w_{ij}}{\sum_j w_{ij}})$ where $\delta_i = s_i - \sum_j w_{ij}$ in order to balance s_i and $\sum_j w_{ij}$. Note that while the first method preserves the empirical distribution of degree weights (although it alters the strength of each node, i.e., the sum of link weights associated with it), the second method does not (Fig. 3.1).

Modular spectra

We analyze the relation between different tumor types (and cancer categories) by using a decomposition in terms of the overlap of their associated genes with the different modules of the gene network. Let the set of all genes be optimally partitioned into M modules. We then define an overlap matrix O, whose rows correspond to the different groups of genes associated with specific tumor types or cancer categories, and the columns correspond to the different modules of the cancer gene network. An element of this overlap matrix O_{ij} is the number of genes in group i that are from the module j. Thus, the decomposition of

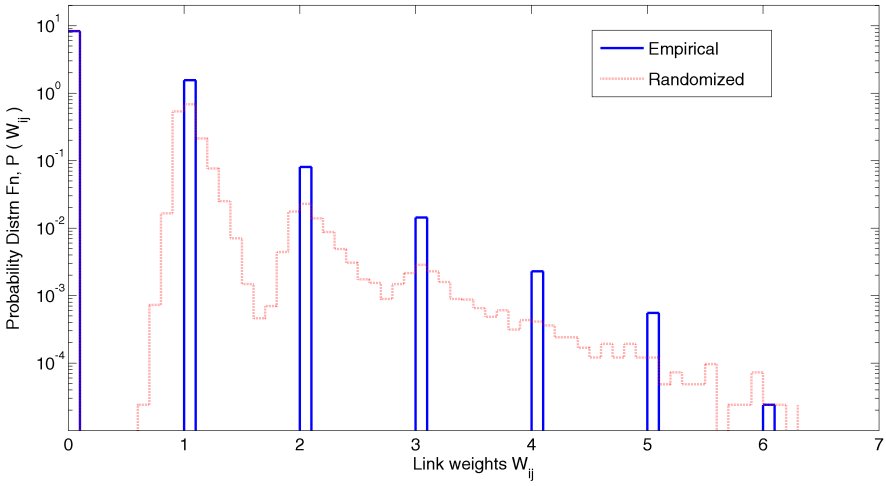

Figure 3.1: Comparison of weight distributions of empirical network with that of strength-preserved randomized networks.

the ith group in the abstract M dimensional basis space formed by the modules is

$$\left\{\frac{O_{i1}}{N_i}, \frac{O_{i2}}{N_i}, \ldots, \frac{O_{iM}}{N_i}\right\},$$

where $N_i = \sum_{k=1}^{M} O_{ik}$ is the total number of genes in the i-th group. The distance between two groups i and j in this "modular" space is defined as

$$d_{i,j}^{modular} = \sqrt{\sum_k \left[\frac{O_{ik}}{N_i} - \frac{O_{jk}}{N_j}\right]^2}.$$

This measure can be used as a metric for closeness or proximity between different tumor types or cancer categories. For visualization of the relation between different groups, a dendrogram is constructed where the ordinate represents the closeness between a pair of groups.

Determining the intra- and inter-modular role of a gene

The role played by each gene in terms of its connectivity within its own module and in the entire network is determined according to two properties [146]: (i) the relative within module degree, z, and (ii) the participation coefficient, P.

The z-score of the within module degree distinguishes nodes that are hubs of their com-

munities from those that are non-hubs. It is defined as

$$z_i = \frac{\kappa^i_{c_i} - \langle \kappa^j_{c_i} \rangle_{j \in c_i}}{\sqrt{\langle (\kappa^j_{c_i})^2 \rangle_{j \in c_i} - \langle \kappa^j_{c_i} \rangle^2_{j \in c_i}}}, \tag{3.1}$$

where κ^i_c is the number of links of node i to other nodes in its community c and $\langle \cdots \rangle_{j \in c}$ are taken over all nodes in module c. The within-community degree z-score measures how well-connected node i is to other nodes in the community.

The nodes are also distinguished based on their connectivity profile over the entire network, in particular, their connections to nodes in other communities. Two nodes with same within module degree z-score can play different roles, if one of them has significantly higher inter-modular connections compared to the other. This is measured by the participation coefficient P_i of node i, defined as

$$P_i = 1 - \sum_{c=1}^{M} \left(\frac{\kappa^i_c}{k_i} \right)^2, \tag{3.2}$$

where M is the total number of communities, κ^i_c is the number of links from node i to other nodes in its community c and $k_i = \sum_c \kappa^i_c$ is the total degree of node i. Therefore, the participation coefficient of a node is close to 1, if its links are uniformly distributed among all the communities, and is 0, if all its links are within its own community.

3.3 Results

Modular structure of the network of cancer-related genes

We first construct a bipartite network consisting of two types of nodes, viz., cancer-related genes (G) and tumor types (TT)[alternatively, we also use cancer categories (CC)]. Nodes of different types are connected based on association between genes and tumor types (or cancer categories) related to them according to information obtained from the F-Census database [140]. From this bipartite network, we produce a tumor type cancer

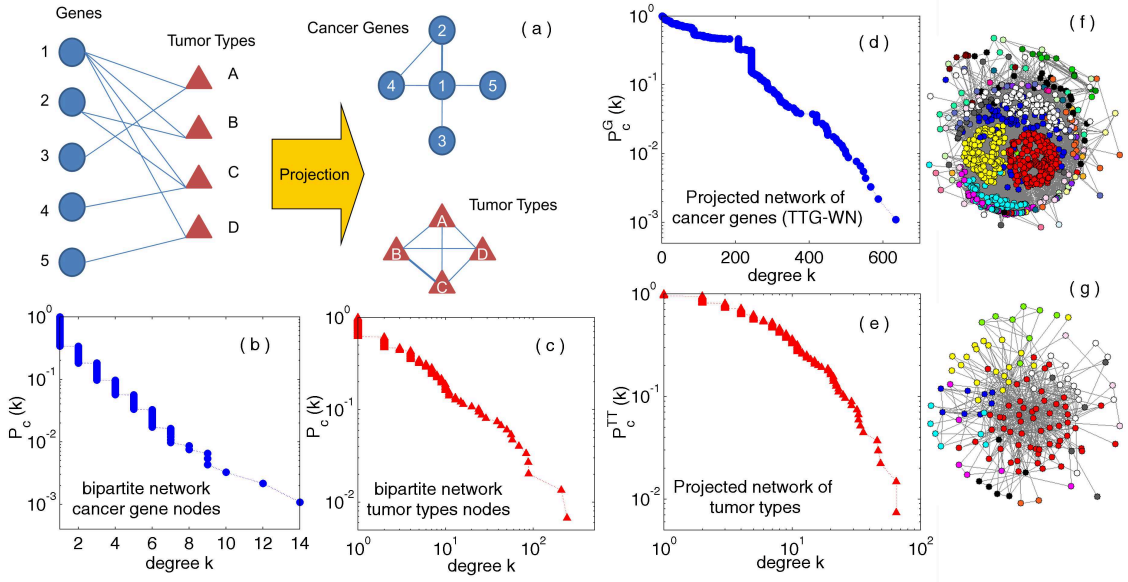

Figure 3.2: Networks of cancer genes and tumor types. (a) Schematic diagram showing the bipartite network comprising genes (represented by circles) and tumor types (triangles). A gene is connected to a tumor type if mutations in the gene result in a tumor of that specific type. The *tumor type-cancer gene network* (TT-GN) is obtained from a projection of the bipartite network, where two genes are connected if there is a tumor type that can be related to mutations in either gene. In the *tumor type-cancer gene weighted network* (TT-GWN), a link between two cancer related genes (e.g., 1 and 2) in the projected network is weighted by the number of tumor types (B,C, ... etc) to which both nodes (viz., 1 and 2) are connected in the bipartite network. The other possible projection yields the tumor type network (TTN) where two tumor types are connected if either can result from mutations in the same gene. Each link can be weighted in proportion to the number of common genes with which the two tumor types are associated. (b-d) The cumulative degree distribution $P_c(k)$, i.e., the probability that a nodes will have k or more links is shown for the (b) genes and (c) tumor types of the bipartite network, as well as for the two networks obtained by projection, viz., (d) the network of cancer genes and (e) the network of tumor types. (f-g) Pictorial representation of (f) the cancer gene network comprising 910 nodes and (g) the network of 135 tumor types.

gene network (TT-GN) and tumor type network (TTN) by using the method of projections [Fig. 3.2 (a)]. According to this technique, from a bipartite network consisting of two categories of nodes, Type I and Type II respectively, one can construct two networks, one comprising only Type I nodes (obtained by connecting any pair of Type I nodes that share as a common neighbor a Type II node in the bipartite network) and the other only Type II nodes (connecting pairs of Type II nodes that share a common Type I node as neighbor) [32]. In the *tumor type-cancer gene network* (TT-GN), the nodes represent different cancer-related genes. Two genes are connected to each other if they have at least one tumor type that they are associated with in common. In the *tumor type-cancer gene weighted network* (TT-GWN), the links are weighted in proportion to the number of common tumor types associated with any pair of connected genes. In the *tumor type network* (TTN), the nodes represent tumor types and two tumor types are connected if there is at least one common element in the set of genes that each is related to. The weight associated with a link is proportional to the number of genes that appear in common for both tumor types. The cumulative degree distribution $P_c(k)$ for cancer genes in the bipartite network shows a rapidly decreasing exponential nature [Fig. 3.2 (b)] while that of the nodes corresponding to tumor types decays more slowly, resembling a power-law (as indicated by the approximately linear nature in double logarithmic scale) [Fig. 3.2 (c)]. However, the projected networks of cancer genes and tumor types both have rapidly decaying tails in the cumulative degree distribution [Fig. 3.2 (d-e)]. The representation of the two projected networks [Fig. 3.2 (f-g)] appears to suggest that they both have a densely connected core surrounded by a periphery of sparsely connected set of nodes.

We have analyzed the mesoscopic organization of TT-GWN by first identifying its modular arrangement using the Infomap method [147]. The basic principle of this community finding algorithm is that optimal compression of network topology uses the regularities in network structure, in particular, the occurrence of modules [95]. Based on its performance in several benchmark tests, the Infomap method has recently emerged as one of the most efficient algorithms for partitioning a network into communities [88, 91]. Figure 3.3 (a)

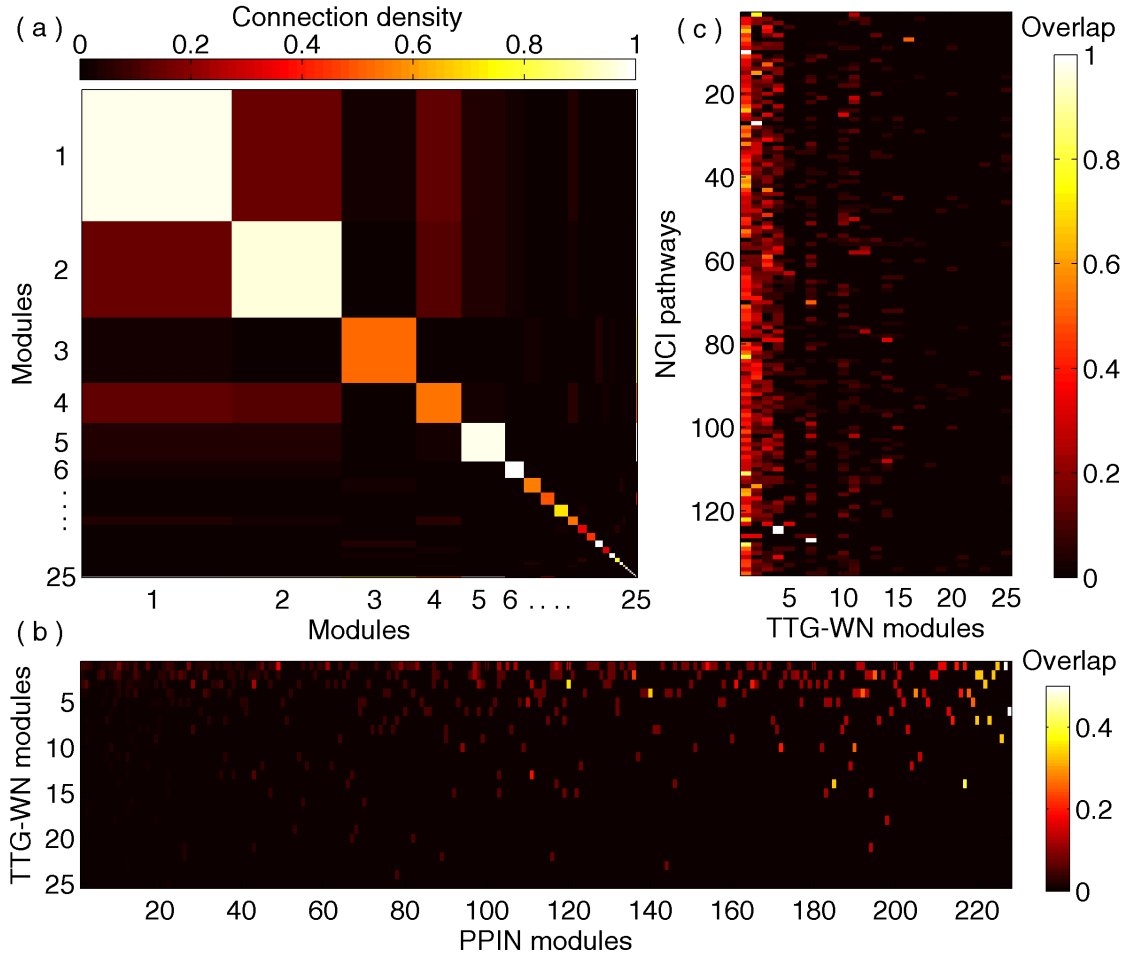

Figure 3.3: Modular interconnectivity in the tumor types-gene network. (a) Matrix representing the average connection density between genes occurring within modules and those in different modules of the TT-GWN. Note that the genes within a module are not only much more densely interconnected compared to the overall connectivity of the network, but modules 1,2,5 and 6 show almost complete intra-connectivity within the genes belonging to them. (b) The overlap between the modules of TT-GWN and those of the Protein-Protein Interaction Network (PPIN) with the modules arranged according to their decreasing size. Several of the smaller PPIN modules have a high degree of overlap with the larger modules of the TT-GWN implying that some of the latter modules contain groups of genes encoding mutually interacting proteins. (c) The overlap between modules of TT-GWN and genes present in different human signaling pathways related to cancer obtained from the National Cancer Institute (NCI) database.

shows the clustering of the network into 25 communities using this method suggesting that the network has a strong modular organization. The modules are of heterogeneous sizes, the largest having 246 genes while the smallest has only 1.

Next we perform a modular decomposition of the largest connected component of the human protein-protein interaction network(PPIN) comprising 9270 proteins which yields 542 modules using the Infomap method [95]. Fig. 3.3 (b) shows that several of the smaller PPIN modules have large overlap with the larger TT-GWN modules implying that the latter modules contain genes that code for mutually interacting proteins. The overlap between modules of TT-GWN and the genes present in the National Cancer Institute (NCI) pathway interaction database [Fig. 3.3 (c)] indicate that many of the genes in the larger modules belong to different human signaling pathways related to cancer.

We also do a k-core decomposition of TT-GN inspired by the apparent core-periphery structure of the gene network suggested by Fig. 3.2 (f). In this method, we recursively remove all nodes having degree less than k to obtain the core of order k. Fig. 3.4 (a-b) show that while members of inner cores have strong inter-connectivity those in the outermost cores have sparse connections, providing quantitative support to the impression given by Fig. 3.2 (f). Fig. 3.4 (c) implies that this is a straightforward consequence of the modular organization of the network as there is significant overlap between the modules of TT-GWN and the shells of different orders (a k-shell is defined as comprising nodes belonging to the kth order core but not that of $(k+1)$-th order). As the core decomposition technique can strictly be applied only to unweighted networks, we use a threshold ω to obtain a sequence of such networks from the TT-GWN 3.4 (d). In these networks a pair of nodes are considered to be connected only if their link weight $> \omega$. We then carry out core-decomposition on each of these networks. The highest link weight in TT-GWN is 6 (between genes ABCA1 and ABCA3) and we note that identifying the innermost core members for networks obtained at large value of ω may provide targets for clinical treatment of cancer as they will be associated with several types of tumor.

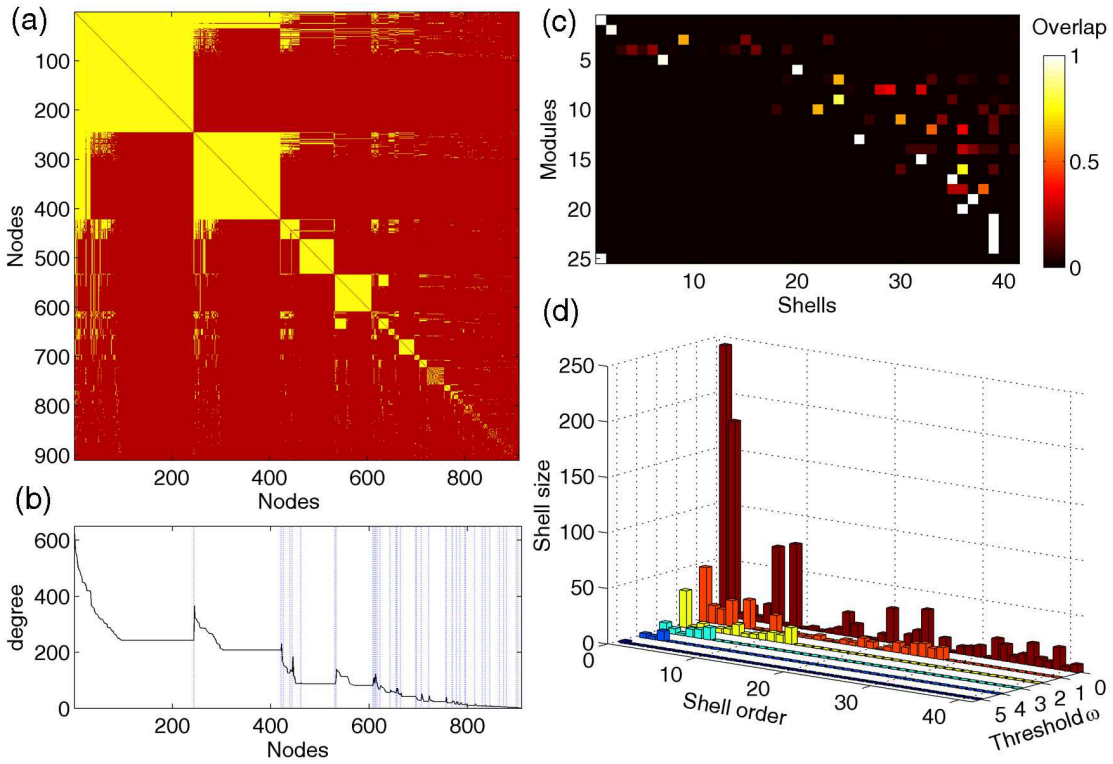

Figure 3.4: Core-periphery analysis of TT-GN. (a) Matrix representing the connections between different genes, which are arranged according to the highest order k-core they belong to (innermost core to the left, outermost or periphery to the right) and then by degree as indicated in (b). The dotted vertical lines represent the boundary of each k-shell, comprising of elements which belong to the k-core but not the $(k+1)$-th core. (c) The overlap between modules of TT-GWN and the different k-shells indicating that most of the shells can be identified with one or more specific modules. (d) The core-periphery structure of TT-GWN using a threshold ω. A sequence of unweighted networks are generated for different values of ω from TT-GWN by considering a pair of nodes to be connected only if their link weight $> \omega$. The corresponding unweighted networks are then analyzed using k-core decomposition.

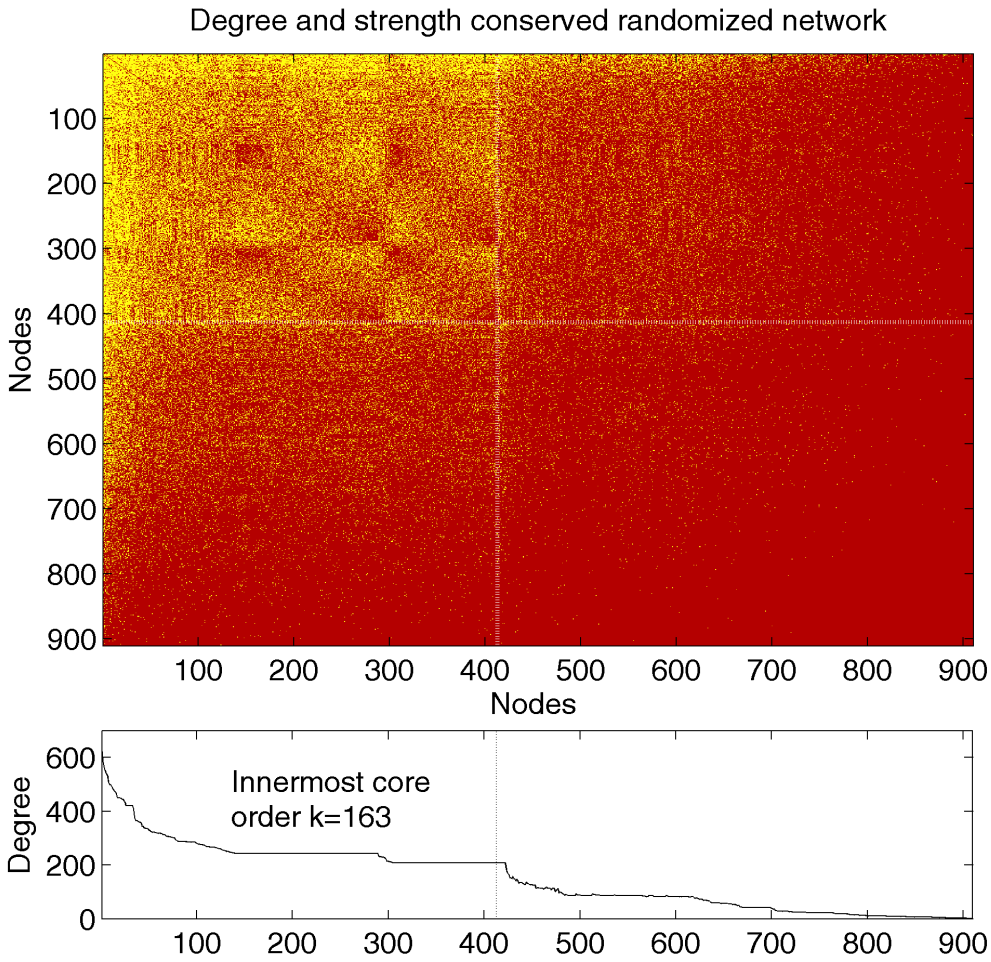

Figure 3.5: Core-periphery analysis of degree and strength-preserved randomized TT-GWN. There is absence of any core-periphery organization similar to that seen in the empirical network.

In order to test the significance of the core-periphery organization of the empirical network, we have performed the same analysis on degree-preserved randomized networks. The randomization of the TT-GWN results in a homogeneous network that does not have any apparent modular or core-periphery organization (Fig. 3.5), suggesting that the observed mesoscopic structure of the cancer-related gene network is highly significant.

Modules, Cancer Categories and Gene Ontology

We have analyzed the composition of the different modules in terms of gene ontology. Fig. 3.6 shows all the TT-GWN modules represented as circles and the inter-connections between them represented as lines whose thickness is related to the total number of con-

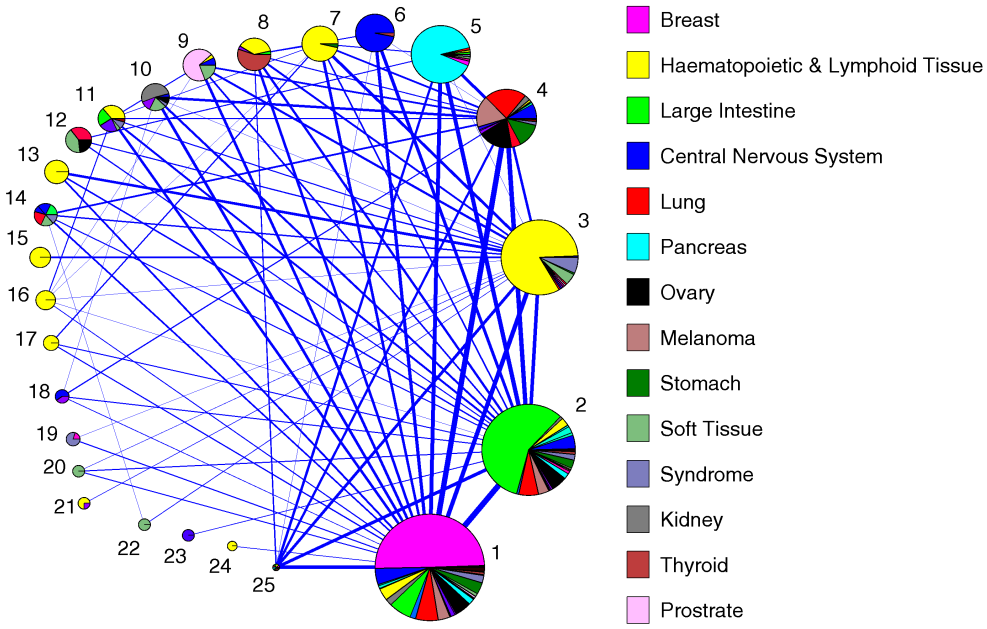

Figure 3.6: The composition of the modules of TT-GWN in terms of different cancer categories. Each circle represents a module of TT-GWN with its size being proportional to the number of genes in that module. The thickness of a line representing a link between a pair of modules is related to total number of connections that exist between the genes of the two modules.

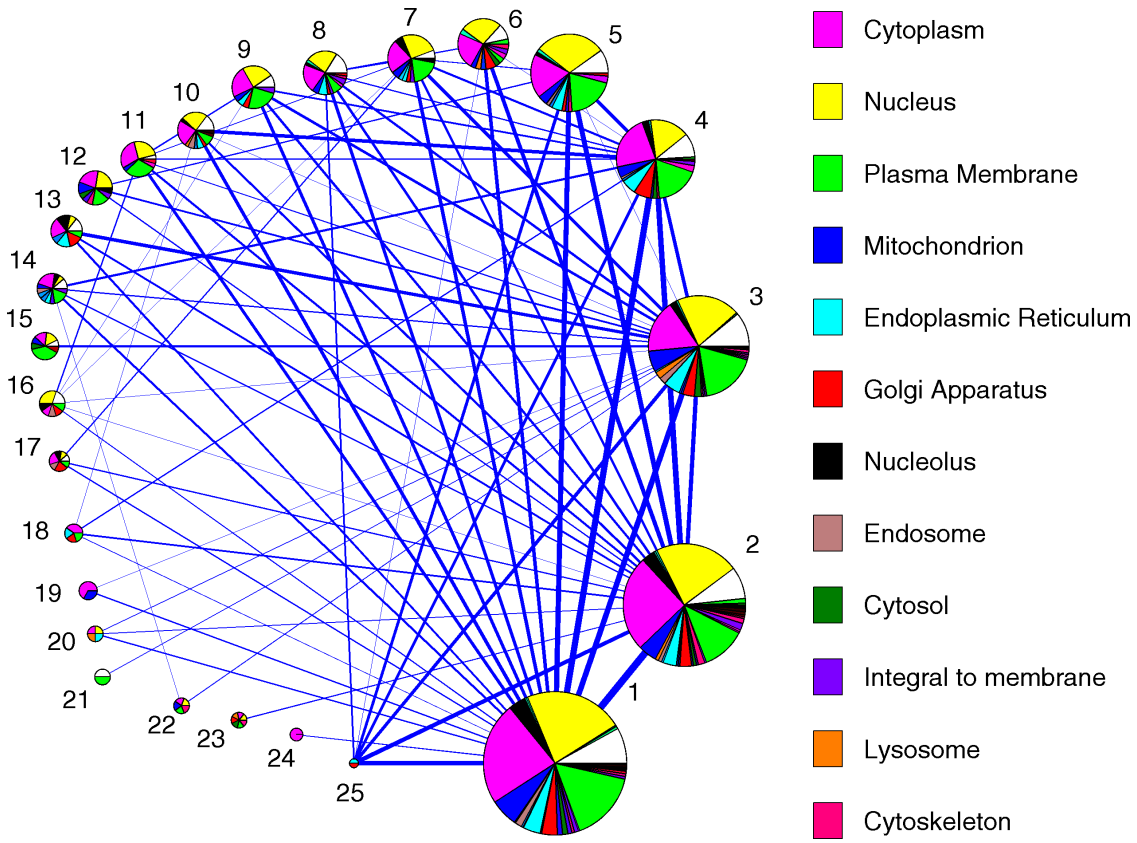

Figure 3.7: The composition of the modules of TT-GWN in terms of different cellular components.

nections between genes belonging in module with genes in the other module. We observe that most of the important cancer categories dominate a particular module. For example, Breast cancer related genes comprise about half of the members of module 1, genes related to cancers of Large Intestine are responsible for more than half the genes belonging to module 2, cancers of the Pancreas and of Central Nervous System dominate modules 5 and 6 respectively, etc. More importantly, a major fraction of genes in eight of the modules are related to Haematopoietic and Lymphoid tissue cancers, making this category the most prolific in terms of dominating the mesoscopic organization of the cancer gene network, even though fewer genes (237) are associated with it than breast cancer(244 genes).

Figs. 3.7 and 3.8 show the dominance of different ontology domains, viz., cellular components and biological processes, in the different modules of the TT-GWN. Apart from

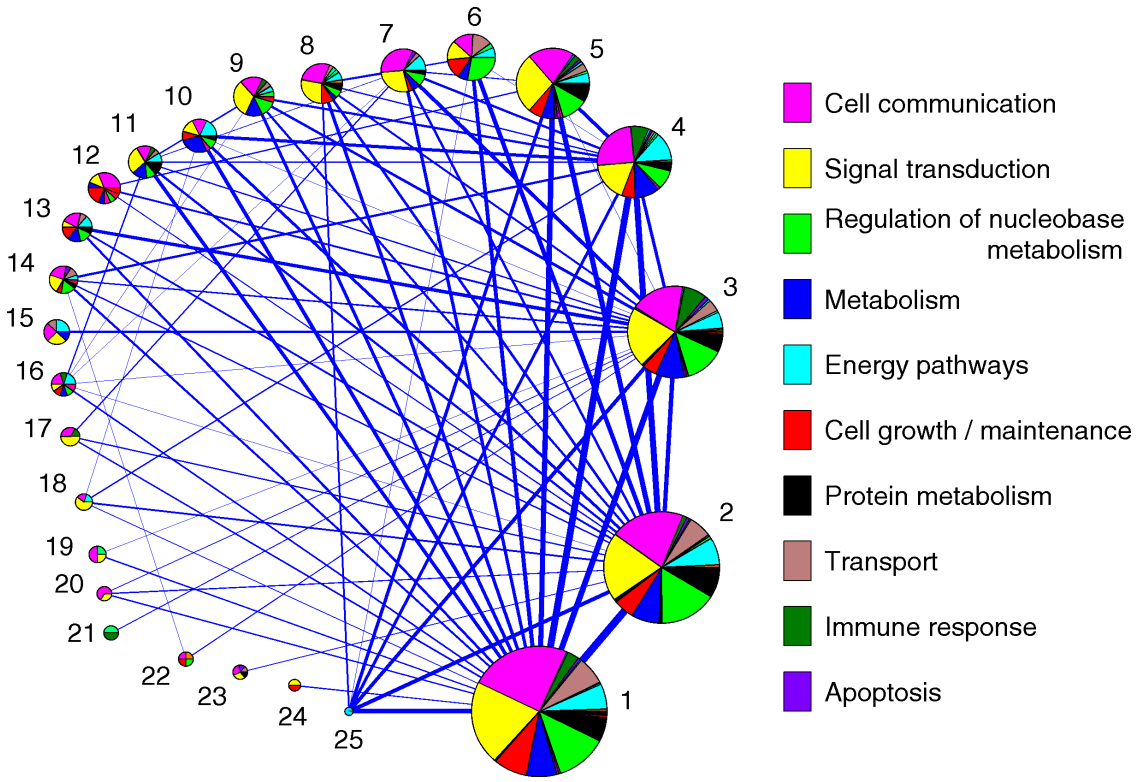

Figure 3.8: The composition of the modules of TT-GWN in terms of different biological processes.

these we have also considered the molecular functions (figure not shown). Unlike the case of cancer categories (Figs. 3.6), none of the modules of TT-GWN can be considered to be related to a specific cellular component or biological processes or molecular function. These appear to have almost similar distribution in the different modules, e.g., the genes belonging to cellular locations corresponding to the cytoplasm, the nucleus and the plasma membrane dominate most of the modules, while in the case of biological processes, genes responsible for cell communication, signal transduction or regulation of nucleobase metabolism contribute the majority of elements in most modules. This result can be understood in light of the fact that cancer is a complex group of multi-factorial diseases involving several genes in multiple cellular locations and responsible for different biological processes and molecular functions.

Closeness between different cancer categories and tumor types

The relation between different cancer categories or tumor types can be understood in terms

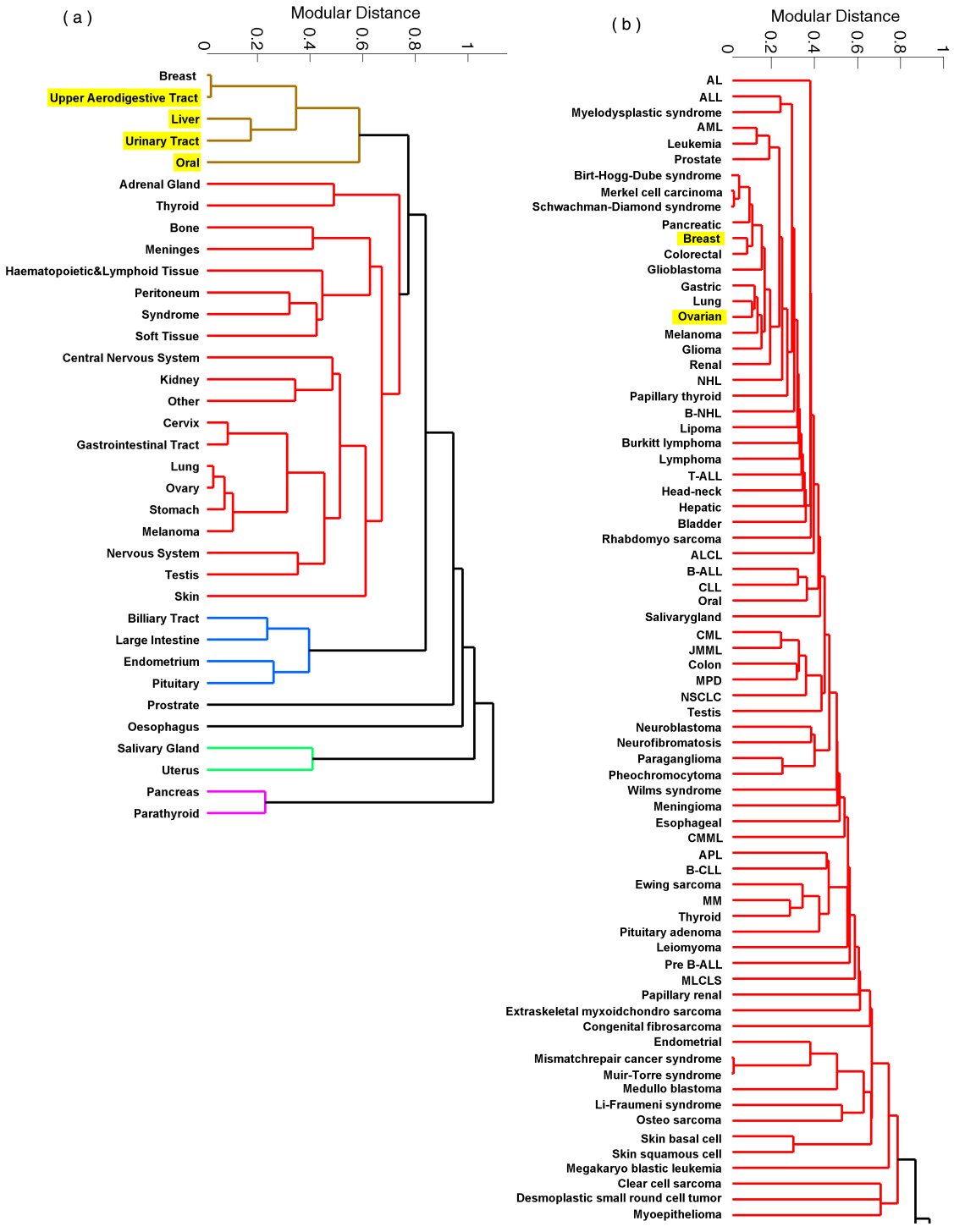

Figure 3.9: (a) Dendrogram of cancer categories obtained by projecting CC gene classes over the space of modules of TT-GWN. Closely connected cancer categories related via environmental factors are highlighted. (b) Dendrogram of tumor types obtained by projecting TT gene classes over the space of PPIN modules (only a section of the entire tree is shown). The closeness of breast and ovarian tumor types, which are related by hormones, hereditary linkages and clinical treatments is indicated in the figure.

of the degree of overlap of the genes associated with the different modules of TT-GWN or PPIN. For this purpose we cluster the cancers or tumors in terms of the similarity in their modular spectra (see Materials and Methods). Relations between the different cancer categories (CC) are represented in terms of a dendrogram shown in Fig. 3.9 (a) where the CC gene classes are projected on the space of modules of TT-GWN. Closely connected cancer categories that are related through environmental factors, viz., oral cancers, cancer of upper aerodigestive tract, liver cancer and urinary tract cancers, have been highlighted. By performing a similar decomposition of the different tumor types in modular space we observe the closeness between breast and ovarian tumors which are related by hormones, hereditary link and clinical treatment [Fig. 3.9 (b)].

Functional roles of cancer genes

We now investigate the importance of individual genes in terms of their connectivity. This is revealed by a comparison between the localization of their connections within their own community and their global connectivity profile over the entire network. In order to do this, we focus on (i) the degree of a node within its module, z, that indicates the number of connections a node i has to other members of its module, and (ii) its participation coefficient, P, which measures how dispersed the connections of a node are among the different modules [146]. A node having low within-module degree is called a non-hub ($z < 1$) which can be further classified according to their fraction of connections with other modules. Following Ref. [146], these are classified as (R1) ultra-peripheral nodes ($P \leq 0.05$), having connections only within their module, (R2) peripheral nodes ($0.05 < P \geq 0.62$), which have a majority of their links within their module, (R3) nonhub connectors ($0.62 < P \geq 0.8$), with many links connecting nodes outside their modules, and (R4) kinless nodes ($P > 0.8$), which form links uniformly across the network. Hubs, i.e., nodes having relatively large number of connections within their module($z \geq 1$), are also divided according to their participation coefficient into (R5) provincial hubs ($P \leq 0.62$), with most connections within their module, (R6) connector hubs ($0.62 < P \leq 0.8$), with a significant fraction of links distributed among many modules, and (R7) global hubs

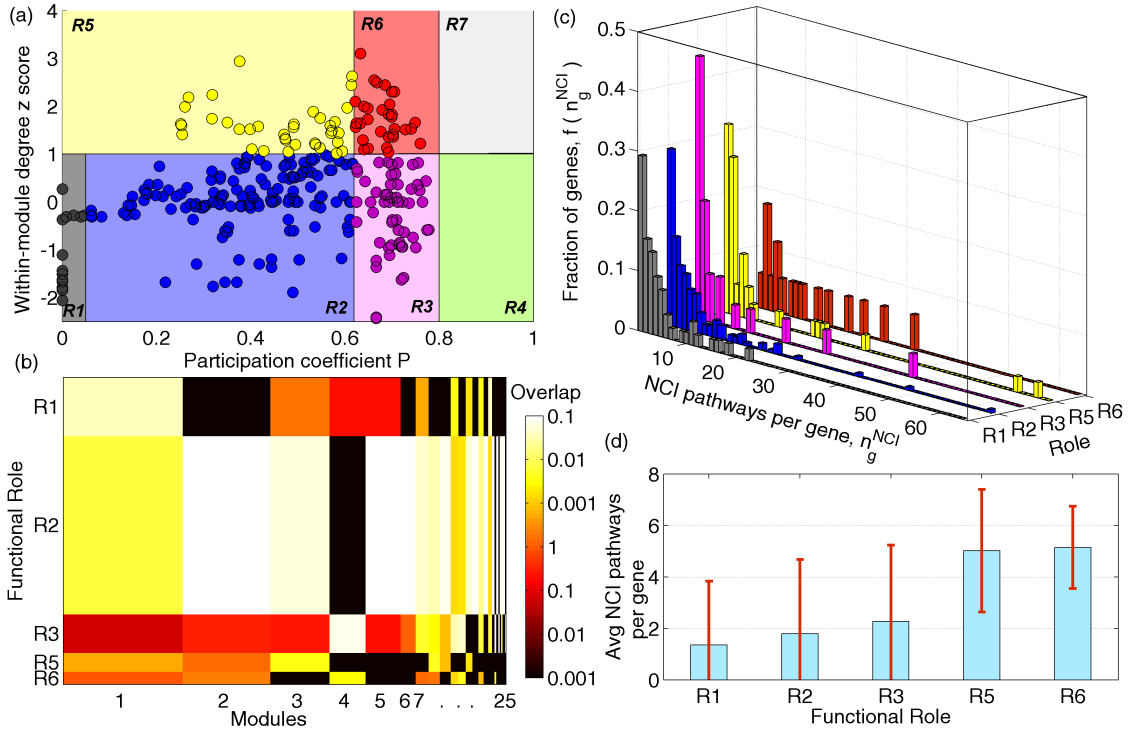

Figure 3.10: Classification of genes in terms of their functional role according to intra- and inter-modular connectivity in TT-GWN. (a) The within-module degree z score of each gene in TT-GWN is shown against the corresponding participation coefficient P. The within module degree measures the connectivity of a node to other nodes within its own module, while the participation coefficient measures its connectivity with nodes in the entire network. Nodes in different regions in the $P - z$ space are categorized as R1: ultra-peripheral nodes, i.e., nodes with all their links within the module, R2: peripheral nodes, i.e., nodes with most of their links within their module, R3: nonhub connector nodes, i.e., nodes with many links to other modules, R4: nonhub kinless nodes, i.e., nodes with links homogeneously distributed among all modules, R5: provincial hubs, i.e., hub nodes with the vast majority of links within their module, R6: connector hubs, i.e., hubs with many links to most of the other modules and R7: global hubs, i.e., hubs with links homogeneously distributed among all modules. (b) Matrix representing the overlap between modules of TT-GWN and the functional roles of their constituent elements (modules are arranged in terms of decreasing size). (c) The fraction of genes with a particular functional role associated with a specified number of human signaling pathways related to cancer (NCI database). (d) The number of signaling pathways in the NCI database that a gene with a specific functional role is associated with on average.

($P > 0.8$), which connect homogeneously to all modules. This classification allows us to distinguish nodes according to their different roles as brought out by their intra-modular and inter-modular connectivity patterns.

We now use this classification method on the nodes of TT-GWN in order to identify genes that play a vital role in cancer through coordinating the behavior of the network either locally within their community or globally over the entire system [Figure 3.10 (a)]. Our analysis reveals that while the network does not have any global hubs, there are several connector hubs - e.g., MAPK14, TP53, BCL10, etc. (see Table 3.1) - that can be potential targets for therapeutic intervention, e.g., through pharmaceutical drugs. Fig. 3.10 (b) shows the overlap between the modules and the functional role of the genes belonging to them. The overlap is measured in terms of the fraction of genes in a specific module that has a particular functional role. Next, we investigate the significance of the functional role of a gene determined from its intra- and inter-modular connectivity by looking at its association with the probability that the gene is connected to one or more human signaling pathways related to cancer. For this purpose we use the Pathway Interaction Database(PID) [148] maintained by the U.S. National Cancer Institute (NCI) and Nature Publishing Group [149]. This is a highly-structured collection of 137 curated and peer-reviewed human signaling pathways. These have been assembled from 9248 known human biomolecular interactions and key cellular processes. Fig. 3.10 (c) shows the fraction of genes with a particular functional role (R1-R6) associated with a specified number of human signaling pathways related to cancer. The distribution clearly shows that there are many more pathways associated with connector hubs (R6) than with genes having other functional roles. Further, genes which are hub nodes (R5 and R6) have a much higher number of signaling pathways associated with them, on average [Fig. 3.10 (d)]. Thus, this supports our earlier conclusion that connector hub genes can be potential therapeutic targets.

Functional roles of proteins in the protein-protein interaction network (PPIN)

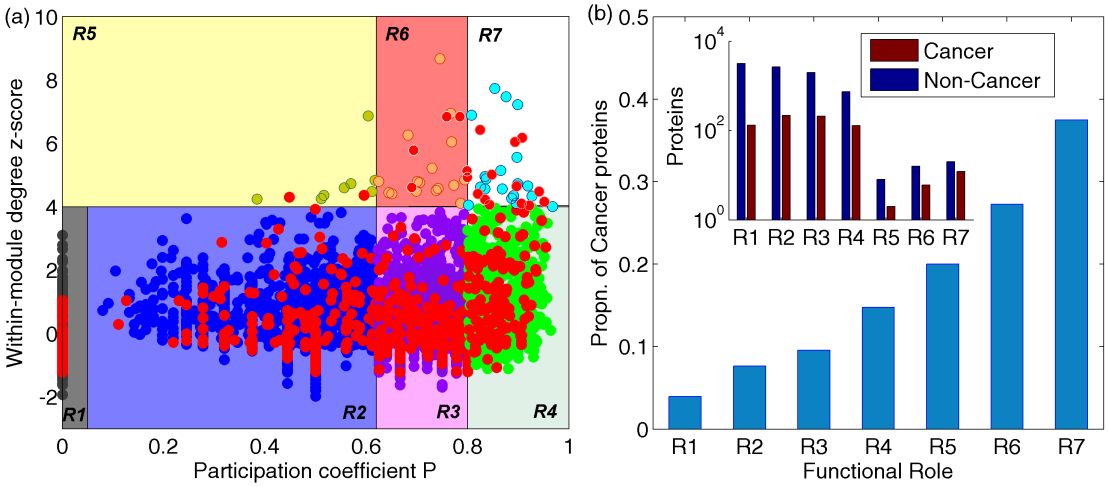

Figure 3.11: The role of individual proteins according to their intra- and inter-modular connectivity in the Protein-Protein Interaction Network (PPIN). (a) The within-module degree z-score of each protein in the PPIN is shown against the corresponding participation coefficient P. The red color circles represent the cancer proteins. The probability that a global hub (R7) or connector hub (R6) protein is related to cancer is extremely high (0.38 and 0.27, respectively) compared to the corresponding average probability for any node in the PPIN (=0.07). (b) The proportion of cancer proteins (and the total number of cancer and non-cancer proteins, inset) in the population of proteins with each functional role.

The basis of all biological functions in the cell in health and disease are the interactions between proteins. Therefore, to further support our hypothesis regarding the importance of network elements having functional roles R6 and R7 we have also analyzed the largest connected component of the protein-protein interaction network(PPIN) comprising 9270 proteins. Classification of proteins into different functional roles in terms of their intra- and inter-modular connectivity shows a preponderance of cancer genes among the connector hubs and global hubs [Fig. 3.11 (a)]. Compared to the probability of a randomly chosen element of the PPIN being related to cancer (0.07), the probability that a global hub (R7) or connector hub (R6) is related to cancer is seen to be extremely high (0.38 and 0.27, respectively) [Fig. 3.11 (b)].

Our mesoscopic structural study of the PPIN reveals several global hubs of which 12 are known to be cancer genes. The 20 other genes which have also been identified as being global hubs in our analysis may have previously unsuspected roles in the genesis and

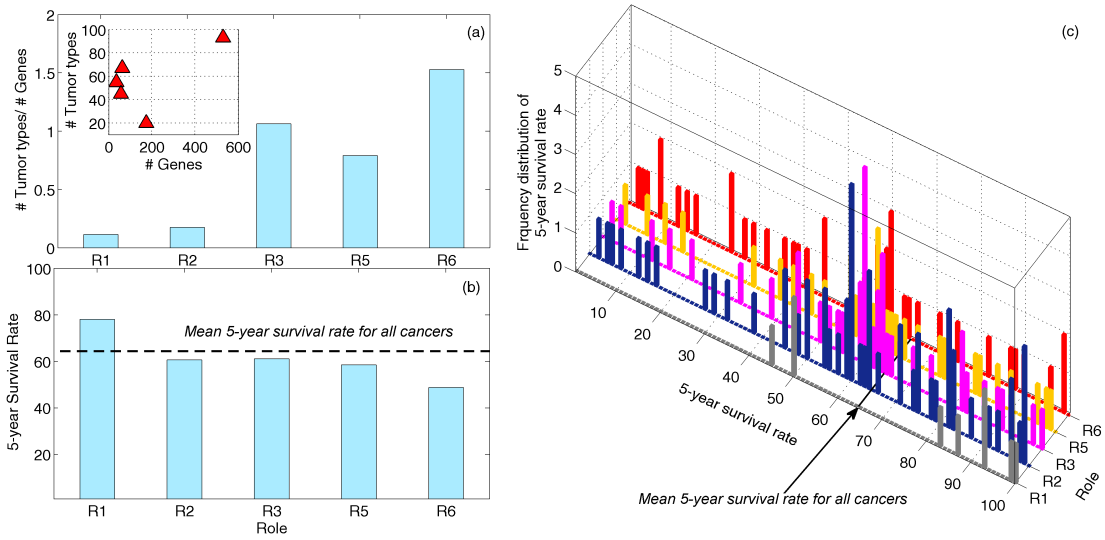

Figure 3.12: Distribution of cancer survival rates associated with genes having specific functional roles in TT-GWN. (a) The ratio of number of tumor types to genes for each functional role category (R1-R6) of genes in TT-GWN. (b) The mean 5-year survival rates for different tumor types corresponding to genes having different roles (R1-R6). The broken line represents the mean 5-year survival rate for all cancers. The data is for US population of cancer patients obtained from the Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute [150]. The rates become progressively low from R1 to R6 signifying that genes that are connector hubs (R6) have relatively high risk than others. (c) The frequency distribution of 5-year survival rates for tumor types associated with genes having different functional roles (R1-R6). It explicitly shows that tumor types associated with connector hub genes of TT-GWN have lower 5-year survival rates than other genes.

treatment of cancer (Table 3.1).

Relating functional role of cancer gene and patient survivability

It is well known that survival probability of a cancer patient depends on the tumor type or cancer category. For instance, the 5-year survival rate for breast or prostate cancer patients is significantly higher than patients diagnosed with brain or lung cancer. Therefore, we have also investigated the relation between tumor types associated with genes that have specific functional roles (R1-R6) and the 5-year survivability rates for patients having these types of tumors. For this purpose, we have used 5-year survival statistics from the Surveillance Epidemiology and End Results (SEER) Program database [150], compiled by the National Cancer Institute as a service to researchers and physicians. The survival rates for 73 tumor types is available from the SEER data (see Table 3.2) is compared to

the classification into 135 tumor types we have used for constructing TT-GWN.

Fig. 3.12 (a) shows the ratio of the number of tumor types to genes for each functional role category of genes in TT-GWN. This shows that connector hub genes are associated on average with a much larger number of tumor types than genes having other functional roles. Fig. 3.12 (b) shows that the 5-year survival rates for tumor types associated with connector hub genes are lower than those associated with genes having other functional roles. For comparison, we show the average 5-year survival rate for all tumor types (broken line). The importance of connector hub genes is even more clearly brought out in Fig. 3.12 (c) which shows the frequency distribution of 5-year survival rates for tumor types associated with genes having different functional roles. Thus, genes which act as connector hubs in TT-GWN are associated with tumors having higher mortality and should be preferentially targeted for therapeutic intervention.

3.4 Discussion and Conclusion

Despite being one of the leading causes of death in the developed world, cancer is yet to be tamed owing to the complex, heterogeneous nature of the disease. Despite the understanding that cancer is a systems-level disease and cannot be treated by targeting a single factor, the large number of elements involved and the dense set of interactions between them have prevented a major breakthrough in this area. In this chapter we have adopted a mesoscopic approach by identifying structural modules in the network of cancer-related genes. This has helped us identifying several genes that have the important functional role of connecting members in their own module with members of other modules. Thus, these genes help coordinate the behavior of the entire network in health and disease - and play a vital role in the origin and treatment of cancer. We validate our hypothesis by showing that tumors associated with these genes are involved in many human signaling pathways related to cancer. More importantly, we show that patients suffering for tumors

involving these genes have a much lower survival rate than those suffering from other types of tumors. The integrated knowledge of cancer network gained by assembling and evaluating the functional roles of the different genes and proteins associated with many tumor types and cancer categories may provide new insights towards understanding the inter-connectedness of key players in the genesis and treatment of the disease. This may have implications for enhancing the efficacy of multiple drug action and proper drug administration, as well as in discovery of novel drug targets.

Table 3.1: Identities of genes that are connector hubs (R6) and global hubs (R7) in TT-GWN and PPIN.

Connector Hubs (R6) of TT-GWN			
APC	INSRR	RPS6KA2	MELK
FAS	MARK1	MAP2K4	KIF1B
ATM	NRAS	TFE3	RAD54B
BRAF	ROR1	TGFBR2	TRIM33
MAPK14	PCM1	TP53	TEX14
EPHB1	PDGFRA	TTN	WNK4
FRAP1	PRCC	TRRAP	ALPK2
FYN	PTCH1	BCL10	NEK10
IGH@	ROS1	AATK	NEK8
Connector Hubs (R6) of PPIN			
CREBBP	GNAI1	SKP1	CCDC85B
DLG1	JUN	SRC	C1orf103
DLG4	SMAD1	TGFBR1	KRTAP4-12
ESR1	MDFI	TRIP13	SFRS12
FN1	PCNA	SLC9A3R1	
FYN	SHBG	SETDB1	
Global Hubs (R7) of PPIN			
ACTA1	EWSR1	PRKACA	YWHAB
ACTB	HDAC1	PRKCA	YWHAG
AR	HRAS	RAC1	GFI1B
CDC42	SMAD2	RB1	NDRG1
MAPK14	SMAD4	ATXN1	PRPF40A
CTNNB1	SMAD9	STX1A	ATF7IP
ATN1	MAGEA11	TP53	UBQLN4
EP300	PPP1CA	TRAF2	SUMO4

Table 3.2: **5-year survival rates (5YSR) for different tumor types obtained from SEER program database [150]**

Tumor Type	5YSR	Tumor Type	5YSR	Tumor Type	5YSR
Acute leukemia	5.8	Esophageal	13.6	Oligo dendro glioma	68.2
Anaplastic large-cell lymphoma	53.9	Ewingsarcoma	48.4	Oral	59.4
Acute lymphocytic leukemia	62.2	Extra skeletal myxoidchondro sarcoma	91	Osteo sarcoma	59.2
Acute myelogenous leukemia	16.5	Gastro intestinal	27.5	Ovarian	53.8
Acute promyelocytic leukemia	60	Glio blastoma	2.9	Pancreatic	4.8
Adrenal	38.7	Glioma	45.2	Papillary thyroid	98.7
Adreno cortical	41.2	Head-neck	57.1	Paraganglioma	65.1
B-cell Non-Hodgkin Lymphoma	50.4	Hepatic	8	Parathyroid	93.1
Basal cell carcinoma	99.4	Hyper parathyroidism-jawtumor syndrome	93.1	Pheochromo cytoma	60.3
Bladder	81.9	Leiomyomata	51.9	Pilocyticastro cytoma	35.8
Brain	23.6	Leukemia	55	Pituitary adenoma	63.8
Breast	87.1	Lipoma	82.8	Prostate	97.6
Burkitt lymphoma	45.4	Lymphocytic leukemia	79.5	Renal	60.2
Chronic lymphatic leukemia	74.9	Lymphoma	70.6	Retinoblastoma	93.5
Chronic myelomonocytic leukemia	37.7	Multiple myeloma	29.4	Rhabdomyo sarcoma	64
CNS	69.5	Myelo proliferative disorder	31.7	Salivarygland	73.9
Cervical	71.5	Medullary thyroid	82.1	Schwannomatosis	99
Cholangio carcinoma	4.5	Medullo blastoma	66.4	Sezary syndrome	88.4
Chondro sarcoma	81.6	Melanoma	90.2	Stomach	21
Clear cell sarcoma	83.4	Meningioma	60	T-cell acute lymphoblastic leukemia	24.3
Colon	64	Merkel cell carcinoma	62.8	Testis	96
Colorectal	62.6	Mesothelioma	8.2	Thyroid	96
Difuse large B-cell lymphoma	50.4	Non-Hodgkin lymphoma	60	Wilms syndrome	78.1
Dermatofibrosarcoma protuberans	99.9	Non-small cell lung cancer	12.1		
Endometrial	74.6	Nasopharyngeal	56.6		

4

Epidemiological dynamics of the 2009 influenza A(H1N1)v outbreak in India

4.1 Introduction

In the preceding chapters we have seen how analysis of networks in the cellular level, in particular, those related to intra-cellular signaling and gene-tumor associations, can help us in understanding certain aspects of different diseases. Networks play an equally important role in the macro-scale dynamics of disease, viz., in the spreading of epidemics in populations of individuals. In this and subsequent chapters we will investigate different aspects of epidemic dynamics and their modeling. While the mathematical investigation of the propagation of biological contagia has received much attention, it is only recently that the role of complex networks has been looked at in detail.

One of the simplifying assumptions often used in mathematical epidemiology is that populations are "well-mixed", i.e., any individual in a population has a equal probability of contact with (and hence, contracting an infection from) any other randomly chosen individual. In reality, people tend to have a higher likelihood of interaction with members

of their immediate social circle, and this property can be used to define a social contact network along which pathogens can spread. Such considerations become important in reality when one is estimating crucial parameters that govern the growth of an epidemic from disease incidence data. For example, when investigating the propagation of an infectious disease in a spatially extended setting, e.g., over an entire country, one may need to obtain a overall growth rate for the epidemic, although it is clear that residents of a particular location (e.g., a town or a city) are much more likely to be infected by people living in the same location than from people staying elsewhere. However, the ease with which people move between geographically dispersed regions at present (through a highly connected transport network) suggests that rapid communication between different locations would allow the "well-mixed" assumption - at least at a coarse level - to be still valid. In this chapter, we analyze empirical data from the 2009 outbreak of influenza A(H1N1) in India and estimate the basic reproduction number R_0 for the epidemic. We observe that despite small regional variations in this parameter, one can estimate a reasonable value for R_0 valid for the entire country.

Influenza is a viral disease which has outbreaks occurring somewhere around the world in most years [151]. Several pandemics of influenza have been reported in the past three centuries, emerging at 10 to 50 year intervals [152]. The pandemic of 1918-19 ("Spanish flu") traveled across the world in three separate waves and resulted in the death of about 50 million people [153], making it one of the worst natural disasters in the history of mankind. Pandemics result from new influenza virus subtypes arising through reassortment of different virus strains [152]. It is important to understand the dynamics of the initial stages of such pandemics in order to come up with possible control strategies. In 2009, a novel influenza strain termed influenza A(H1N1)v, that was first identified in Mexico in March, rapidly spread to different countries and became the predominant influenza virus in circulation worldwide [154, 155]. By April 11, 2010 it caused at least 17798 deaths in 214 countries [156]. The first confirmed case in India, a passenger arriving from the USA, was detected on May 16, 2009 in Hyderabad. The initial cases were

passengers arriving by international flights. However, towards the end of July, the infections appeared to have spread into the resident population with an increasing number of cases being reported for people who had not been abroad. By April 11, 2010, there were 30352 laboratory confirmed cases in India (out of 132796 tested) and 1472 deaths were reported, i.e., 5 % of the cases which had tested positive for influenza A(H1N1)v were fatal [157].

To devise effective strategies for combating the spread of pandemic influenza A(H1N1), it is essential to estimate the transmissibility of this disease in a reliable manner. This is generally characterized by the reproductive rate R of the epidemic which is defined as the average number of secondary infections resulting from a single (primary) infection. A special case is the *basic reproduction number* R_0, which is the value of R measured when the overall population is susceptible to the infection as is the case at the initial stage of an epidemic. Estimate of the basic reproduction number for influenza A(H1N1)v in reports published from data obtained for different countries have varied widely. For example, R_0 has been variously estimated to be between 2.2 to 3.0 for Mexico [158], 1.72 for Mexico City [159], between 1.4 and 1.6 for La Gloria in Mexico [160], between 1.3 to 1.7 for the United States [161] and 2.4 for Victoria State in Australia [162]. The divergence in the estimates for the basic reproduction number may be a result of under-reporting in the early stages of the epidemic or due to climatic variations. They may also possibly reflect the effect of different control strategies used in different regions, ranging from social distancing such as school closures and confinement to antiviral treatments.

Here we estimate the basic reproduction number for the infections using the time-series of infections in India extracted from reported data. By assuming an exponential rise in the number of infected cases $I(t)$ during the initial stage of the epidemic when most of the population is susceptible, we can express the basic reproduction number as $R_0 = 1 + \lambda\tau$ (see, e.g., Ref. [57], p. 19), where λ is the rate of exponential growth in the number of infections, and τ is the mean generation interval[1], which is approximately equal to 3

[1]Mean generation interval is a measure of the time period between the occurrence of a primary infec-

days [159]. Using the time-series data we obtain the slope λ of the exponential growth using several different statistical techniques. Our results show that this quantity has a value of around 0.15, corresponding to $R_0 \simeq 1.45$.

4.2 Materials and Methods

We have used data from the daily situation updates for influenza A(H1N1) available from the website of the Ministry of Health and Family Welfare, Government of India [163]. In our analysis, data up to September 30, 2009 was used, corresponding to a total of 10078 positive cases. Note that, after September 30, 2009, patients exhibiting mild flu like symptoms (classified as categories A and B) were no longer tested for the presence of the influenza A(H1N1) virus.

As the data exhibit very large fluctuations, with some days not showing a single case while the following days show extremely large number of cases, it is necessary to smooth the data using a moving window average. We have used an n-day moving average ($n = 2-10$), which removes large fluctuations while remaining faithful to the overall trend.

4.3 Results

From the incidence data for the 2009 pandemic influenza in India it appears that the disease has been largely confined to the urban areas of the country. Indeed, 6 of the 7 largest metropolitan areas of India (which together accommodate about 5 % of the Indian population [164]) account for 7139 infected cases up to September 30, 2009, i.e., 70.8 % of the data-set we have used. However, it is possible that this is partly a result of bias

tion and a secondary case caused by the primary infected individual (i.e., "the average time taken for the secondary cases to be infected by a primary case", as defined in Ref. [57]). It has alternatively been defined as the “sum of the average latent and the average infectious period” where latent period refers to the time that an individual has been infected but is not yet infectious [57].

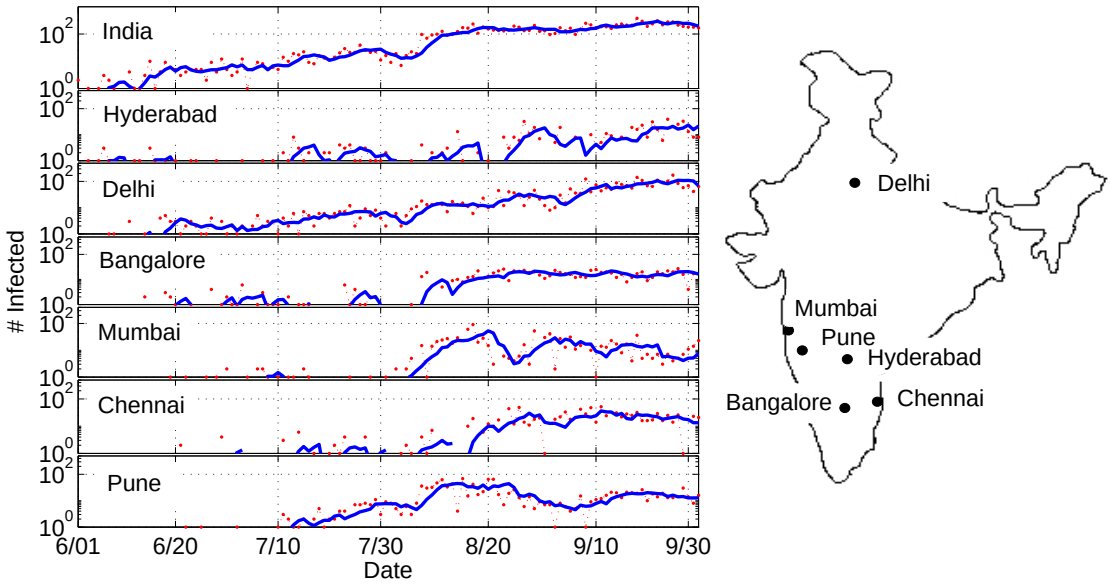

Figure 4.1: Time-series of the number of infected cases, #Infected, of influenza A(H1N1)v showing the daily data (dotted) as well as the 5-day moving average (solid line) for India and the six metropolitan areas with the highest number of infections (whose geographic locations are shown in the adjoining map). The period shown is from June 1 to September 30, 2009. At the beginning of this period most of the infected people were arriving from abroad, while at the end of it the infection was entrenched in the local population. The data shows that almost all the cities showed a simultaneous increase in the number of infections towards the end of July and the beginning of August. This is manifested as a sudden rise in #Infected for India as a whole (note the semilogarithmic scale), and can be taken as the period in which the infection started spreading in the resident population.

introduced by the easier accessibility to testing facilities for urban populations.

Figure 4.1 shows the daily number of confirmed infected cases, as well as, the 5-day moving average from June 1 to September 30, 2009, for the country as a whole and the six major metropolitan areas which showed the highest incidence of the disease: Hyderabad, Delhi, Bangalore, Mumbai, Chennai and Pune. The adjoining map shows the geographic locations of these six cities. In the period up to July 2009, infections were largely reported in people arriving from abroad. There is a marked increase in the number of infections towards the end of July and the beginning of August 2009 in all of these cities (note that the ordinate is in logarithmic scale). This is manifested as a sudden rise in the number of infected cases for the country as a whole, implying that the infection started spreading in

the resident population between July 28 and August 12.

Figure 4.2 (a) shows the exponential slope λ estimated in the following way. The time-series of the number of infections is first smoothed by taking a 5-day moving average. The resulting smoothed time-series is then used to estimate λ by a regression procedure applied to the logarithm of the number of infected cases [log(#infected)] across a moving window of length Δt days. The origin of the window is varied across the period 1st June to 20th August (in steps of 1 day). We then repeat the procedure by varying the length of the window over the range of 7 days to 36 days. To quantify the quality of regression we calculate the correlation coefficient r [Fig. 4.2 (b)] between log (#Infected) and time (in days), and its measure of significance p [Fig. 4.2 (c)]. The correlation coefficient r is bounded between -1 and 1, with a value closer to 1 indicating a good fit of the data to an exponential increase in the number of infections. The measure of significance of the fitting is expressed by the corresponding p-value, which expresses the probability of obtaining the same correlation by random chance from uncorrelated data. The average of the estimated exponential slope λ is obtained by taking the mean of all values of λ obtained for windows originating between July 28-Aug 12 and of various sizes, for which the correlation coefficient $r > r_{cutoff}$ (we consider $0.75 < r_{cutoff} < 1$ in our analysis) and the measure of significance $p < 0.01$. For comparison, we show again in Figure 4.2 (d) the number of infected cases of H1N1 in India (dotted) together with its 5-day moving average (solid line). The horizontal broken lines running across the figure indicate the period between July 28 and August 12 which exhibited the highest increase in number of infections within the period under study (from 1st June to 30th September) .

Figure 4.3 shows the average exponential slope $\langle\lambda\rangle$ as a function of r_{cutoff}, calculated for the original data and for different periods n over which the moving average is taken ($n = 2, 3, 4, 5$ and 10). For n = 3-5, the data show a similar profile indicating the robustness of the estimate of the average exponential slope $\langle\lambda\rangle$ with respect to different values of n. The sudden increase in $\langle\lambda\rangle$ around $r_{cutoff} \simeq 0.9$ implies that beyond this region the

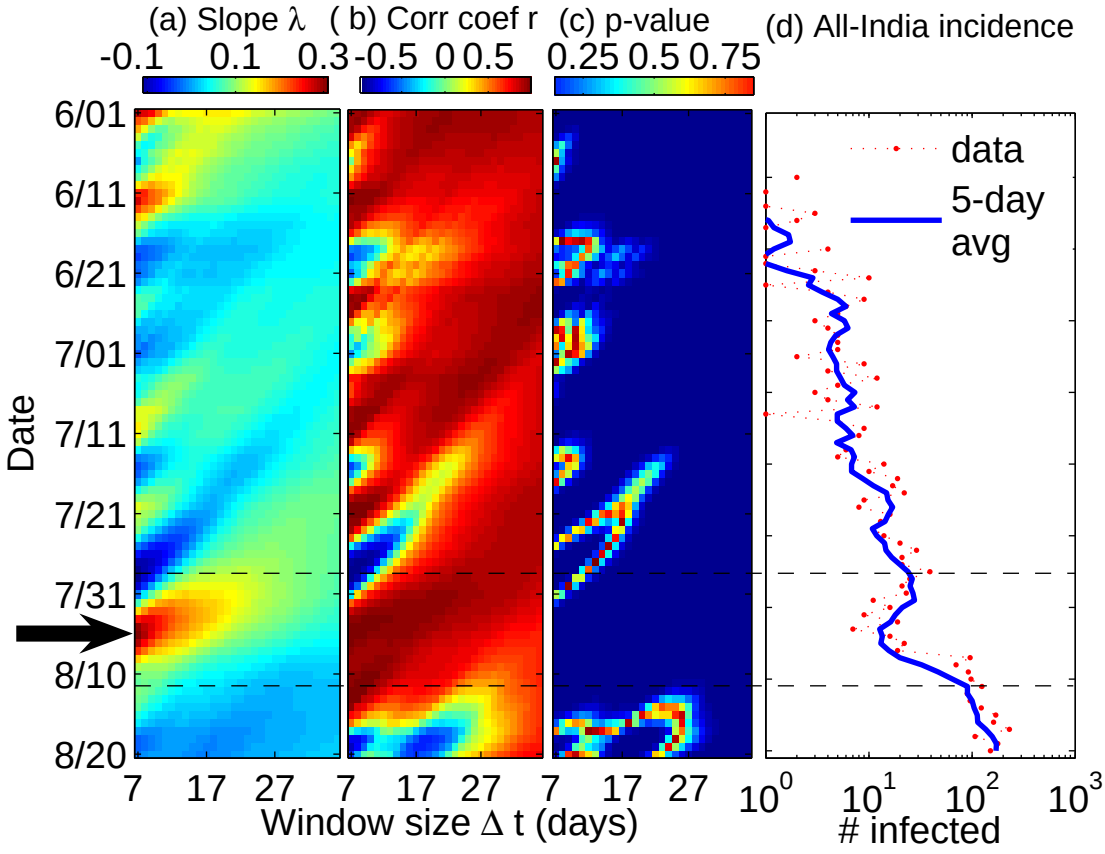

Figure 4.2: (a) The exponential slope λ estimated from the time-series data of number of infected cases, #Infected, averaged over a 5-day period to smoothen the fluctuations (d, solid curve). The slope λ is calculated by considering the number of infected cases over a moving window having different sizes (Δt), ranging between 7 days and 36 days. By moving the starting point of the window across the period 1st June-20th August (in steps of 1 day) and calculating the best fit linear slope of the data on a semi-logarithmic scale (i.e., time in normal axis, number of infections in logarithmic axis) we obtain an estimate of λ. The arrow indicates the region between July 28-August 12 (region within the broken lines), which shows the largest increase in number of infections within the period under study, corresponding to the period when the epidemic broke out in the resident population. Over this time-interval, the average of λ is calculated for the set of starting dates and window sizes over which (b) the correlation coefficient r between log(#Infected) and t, is greater than r_{cutoff} (we consider $0.75 < r_{cutoff} < 1$ in our analysis) and (c), the measure of significance for the correlation $p < 0.01$.

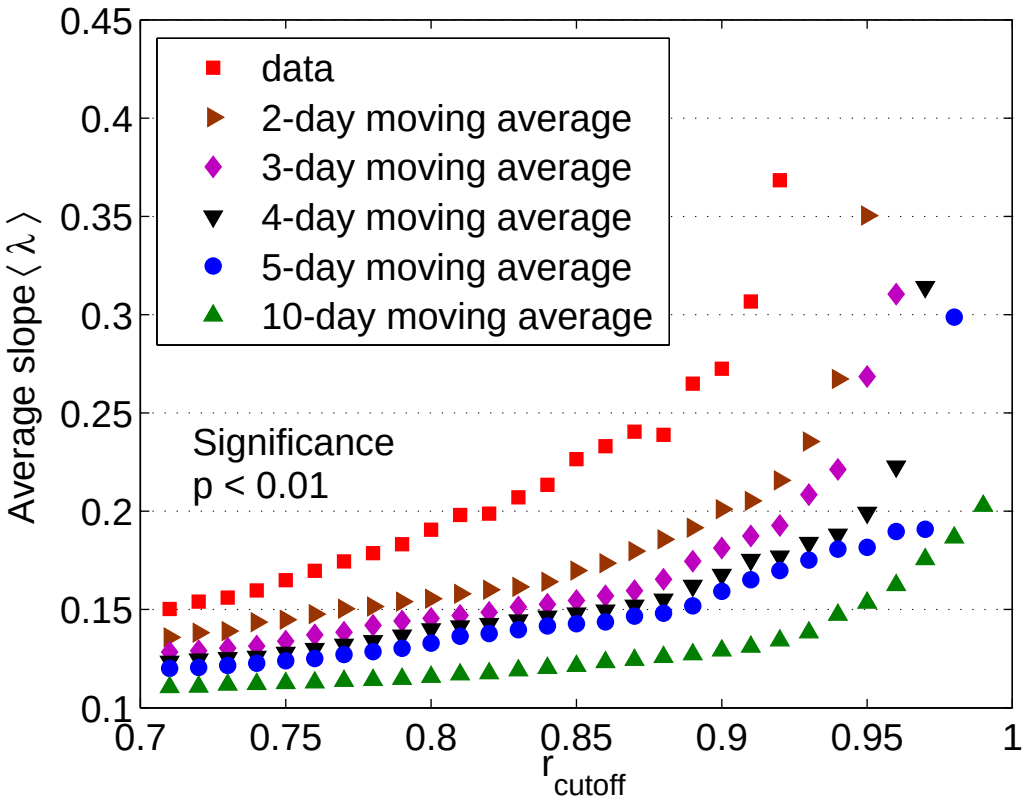

Figure 4.3: Average slope $\langle\lambda\rangle$ of the variation in log(#Infected) with time t, as a function of the threshold of correlation coefficient, r_{cutoff}, used to filter the data. The averaging is performed for infections occurring within the period July 28-August 12 (for details see caption to Fig. 4.2). Different symbols indicate the actual daily time-series data (squares) and the data smoothed over a moving n-day period, with $n = 2$ (right-pointed triangle), 3 (diamond), 4 (inverted triangle), 5 (circle) and 10 (triangle). The significance of the correlation between log(#Infected) with time t, $p < 0.01$ for all data points used in performing the average. Note that for $n = 3, 4, 5$ the data show very similar profiles for variation of $\langle\lambda\rangle$ with r_{cutoff}, indicating the robustness of the estimate with respect to different values of n used. The sudden increase in the value of the average slope around $r_{cutoff} \simeq 0.9$ implies that beyond this region the slope depends sensitively on the cutoff value. Considering the region where the variation is more gradual gives us an approximate value of the slope $\lambda \sim 0.15$, corresponding to a basic reproduction number $R_0 \simeq 1.45$.

slope depends sensitively on the cutoff value. Considering the region where the variation is smoother gives an approximate value $\lambda \sim 0.15$, corresponding to a basic reproduction number for the epidemic $R_0 = 1 + \lambda\tau \simeq 1.45$, assuming the mean generation interval, $\tau = 3$ days.

We compute the confidence bounds for the estimate of R_0 from the 5-day moving average time-series by using the *confint* function of the scientific software MATLAB [165]. This function generates the goodness of fit statistics using the solution of the least squares fitting of log(#Infected) to a linear function. It results in a mean value $\langle\lambda\rangle = 0.16$, with the corresponding 95 % confidence intervals calculated as [0.116, 0.206], consistent with our previous estimate of $R_0 \simeq 1.45$.

We have also used bootstrap methods to estimate the exponential slope, λ. This involves selecting random samples with replacement from the data such that the sample size equals the size of the actual data-set. The same analysis that was performed on the empirical data is then repeated on each of these samples. The range of the estimated values λ' calculated from the random samples allows determination of the uncertainty in estimation of λ. Fig. 4.4 (a) shows the average, $\langle\lambda'\rangle$, calculated for different periods (with abscissa indicating the starting date and the symbol indicating the duration of the period) from the 5-day moving average time-series data of infected cases. The curves corresponding to the periods of different durations (14-16 days) intersect around July 31, 2009, indicating that the value of the average exponential slope is relatively robust with respect to the choice of the period about this date. The average value of the bootstrap estimates λ' at the intersection of the three curves is 0.15, in agreement with our earlier calculations of λ.

Fig. 4.4 (b) shows the distribution of the bootstrap estimates of the exponential slope for a particular period, July 31 to August 15, 2009. The average slope $\langle\lambda'\rangle$ obtained from 1000 bootstrap samples for this period is 0.166 with a standard deviation of 0.024, which indicates that the spread of values around the average estimate of $\langle\lambda'\rangle = 0.15$ is not large. This confirms the reliability of the estimated value of the exponential slope, and hence of

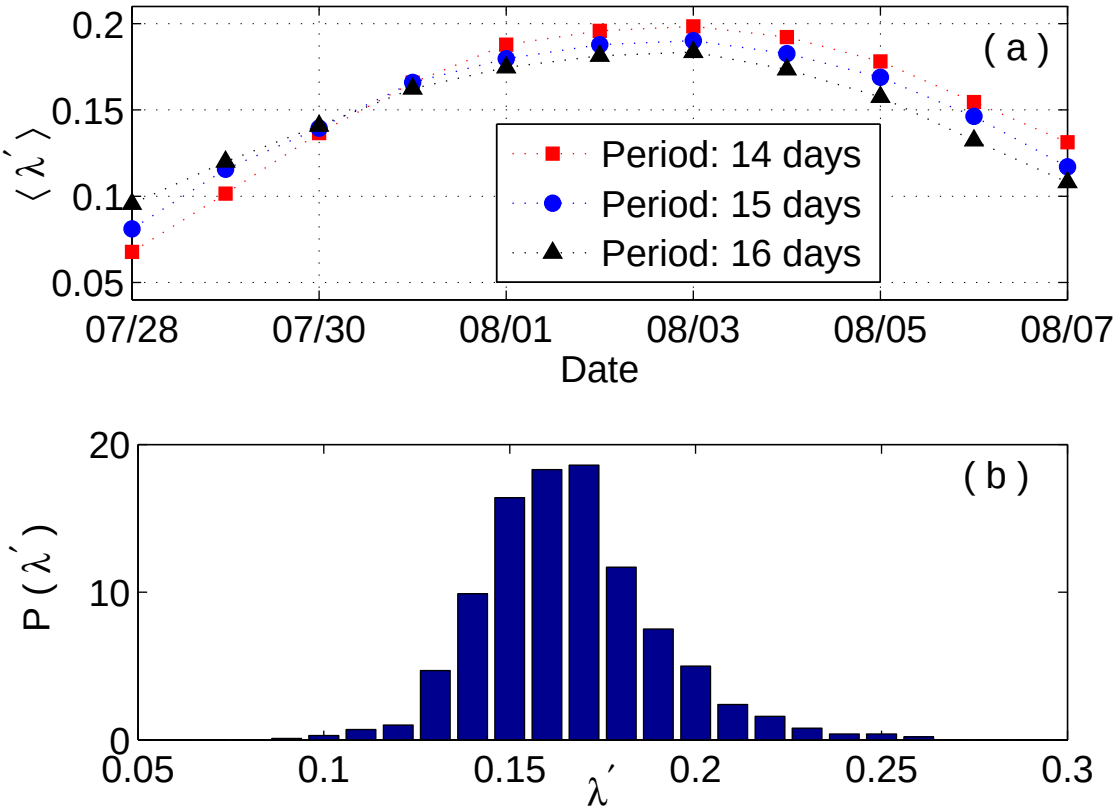

Figure 4.4: (a) The averages of the bootstrap estimates for the exponential slope, λ', calculated for different periods (with the abscissa indicating the starting date and the symbol indicating the duration) from the 5-day moving average time-series data of infected cases in India. The curves corresponding to the periods of different durations (14-16 days) intersect around July 31, 2009, indicating that the value of the average exponential slope is relatively robust with respect to the choice of the period about this date. (b) The distribution of bootstrap estimates of the exponential slope for the period July 31 to August 15, 2009. The average slope $\langle \lambda' \rangle$ obtained from 1000 bootstrap samples is 0.166 with a standard deviation of 0.024, which agrees with the approximate value of $\lambda = 0.15$ (corresponding to $R_0 = 1.45$) calculated in Fig. 4.3.

Table 4.1: **Regional variation of basic reproduction number for 2009 Influenza A(H1N1)v epidemic in India.** R_0 is estimated by the method of exponential curve fitting from 5-day moving averages of incidence data for different regions/cities. In each case, bootstrap estimates yielded similar values. For each region/city, the time interval over which R_0 is determined (*Period*) is chosen on the basis of exhibiting the highest rise in disease incidence. Note that the *Southern Region* comprises the cities of Bangalore, Chennai and Hyderabad.

Region or City	Period	$\langle\lambda\rangle$	R_0
Pune	30/07-14/08	0.25 ± 0.04	1.74 ± 0.14
Mumbai	05/08-20/08	0.22 ± 0.06	1.65 ± 0.18
Delhi	13/08-28/08	0.12 ± 0.02	1.36 ± 0.06
Southern Region	15/08-30/08	0.11 ± 0.02	1.34 ± 0.05

our calculation of the basic reproduction number.

In addition to estimating R_0 for the entire country, we have also separately evaluated the basic reproduction number for the different regions in which the epidemic occurred (Table 4.1).

4.4 Discussion and Conclusion

It may appear surprising that there was a very high number of infections in Pune (1238 positive cases up to September 30), despite it being less well-connected to the other major metropolitan cities of India, in comparison to urban centers that did not show a high incidence of the disease. For example, the Kolkata metropolitan area, which has a population around three times the population of the Pune metropolitan area [164], had only 113 positive cases up to September 30. This could possibly reflect the role of local climatic conditions: Pune, located at a relatively higher altitude, has a generally cooler climate than most Indian cities. In addition, the close proximity of Pune to Mumbai and the high volume of road traffic between these two cities could have helped in the transmission of the disease. Another feature pointing to the role of local climate is the fact that in Chennai, most infected cases were visitors from outside the city, while in Pune, the majority of the cases were from the local population, even though the total number of infected cases

listed for the two cities in our data-set are comparable (928 in Chennai and 1213 in Pune). This suggests the possibility that the incidence of the disease in Pune could have been aided by its cool climate, in contrast to the hotter climate of the coastal city of Chennai. The rapid spread of the disease in Pune may also have originated in transmission amongst the large crowds of people who had gathered in the H1N1 testing centers, given that the numbers appearing for testing here were much larger than elsewhere.

The calculation of R_0 for India assumes well-mixing of the population (i.e., homogeneity of the contact structure) among the major cities in India. Given the rapidity of travel between the different metropolitan areas via air and rail, this may not be an unreasonable assumption. However, some local variation in the development of the epidemic in different regions can indeed be seen (Fig. 4.1). Around the end of July, almost all the cities under investigation showed a marked increase in the number of infected cases - indicating spread of the epidemic in the local population. This justifies our assumption of well-mixing in the urban population over the entire country for calculating the basic reproduction number.

To conclude, we stress the implications of our finding that the basic reproduction number for pandemic influenza A(H1N1)v in India lies towards the lower end of the values reported for other affected countries. This suggests that season-to-season and country-to-country variations need to be taken into account in order to formulate strategies for countering the spread of the disease. Evaluation of the reproductive rate, once control measures have been initiated, is vital in determining the future pattern of spread of the disease.

5

Persistence of epidemics in networks with modular organization

5.1 Introduction

Infectious diseases continue to be one of the major causes of death for human populations around the world. Especially in developing countries they pose a large threat to society in terms of individual suffering, economic losses and social tension [56, 166]. Outbreaks of several new and re-emerging infectious diseases (such as avian influenza, SARS, Ebola, tuberculosis, etc.) have been reported in the past decade. It is important to understand how the transmission processes and effects of such diseases can change as a result of climate change, variations in economic patterns, growing human population, increased international travel, environmental degradation and spread of human and animal populations to new ecosystems [166, 176]. The increasing prevalence of antibiotic-resistant pathogens, which strategically adapt and evolve continuously, have also introduced new factors in the design of new drugs and chemicals that target infectious agents and their vectors [166, 176].

Mathematical models and computer simulations have become important tools for understanding and analyzing the transmission patterns of infectious diseases [56,57]. They can also be used to quantitatively test the effectiveness of control measures. The mathematical formulation of infectious disease dynamics models clarifies assumptions, identifies the key variables and parameters related to the spreading process and provide important results such as the calculation of the basic reproduction R_0 for various epidemics. These have the potential to provide valuable ideas and methods to control the spread and severity of infectious diseases. They also enable governmental and other agencies to plan and implement control or eradication programs [56, 154, 156, 157].

In this chapter we use modular network models on which we implement a well-known epidemic dynamical model in order to identify the role of mesoscopic structural organization of contact networks in making a highly infectious disease persistent. We show that while for epidemics with less degree of infectiousness (quantified by the basic reproduction rate R_0) the disease can persist even in homogeneous networks, for higher levels of infectiousness the epidemic persists only in networks with high degree of modular organization. This has obvious implications in the determination of critical community sizes below which a disease cannot become endemic.

5.2 Materials and Methods

The contact network model we have used to investigate the role of community organization in the long-term dynamics of epidemics is constructed such that the N nodes (representing agents) comprising the system are arranged into M modules with $n(= N/M)$ nodes each [63]. The connection probability between a pair of nodes belonging to the same module is ρ_i, while that between nodes belonging to different is ρ_o. The modular nature of the network can be varied continuously by altering the ratio of inter- to intra-modular connectivity, $r = \rho_o/\rho_i \in [0, 1]$, keeping the average degree k constant (Fig. 5.1).

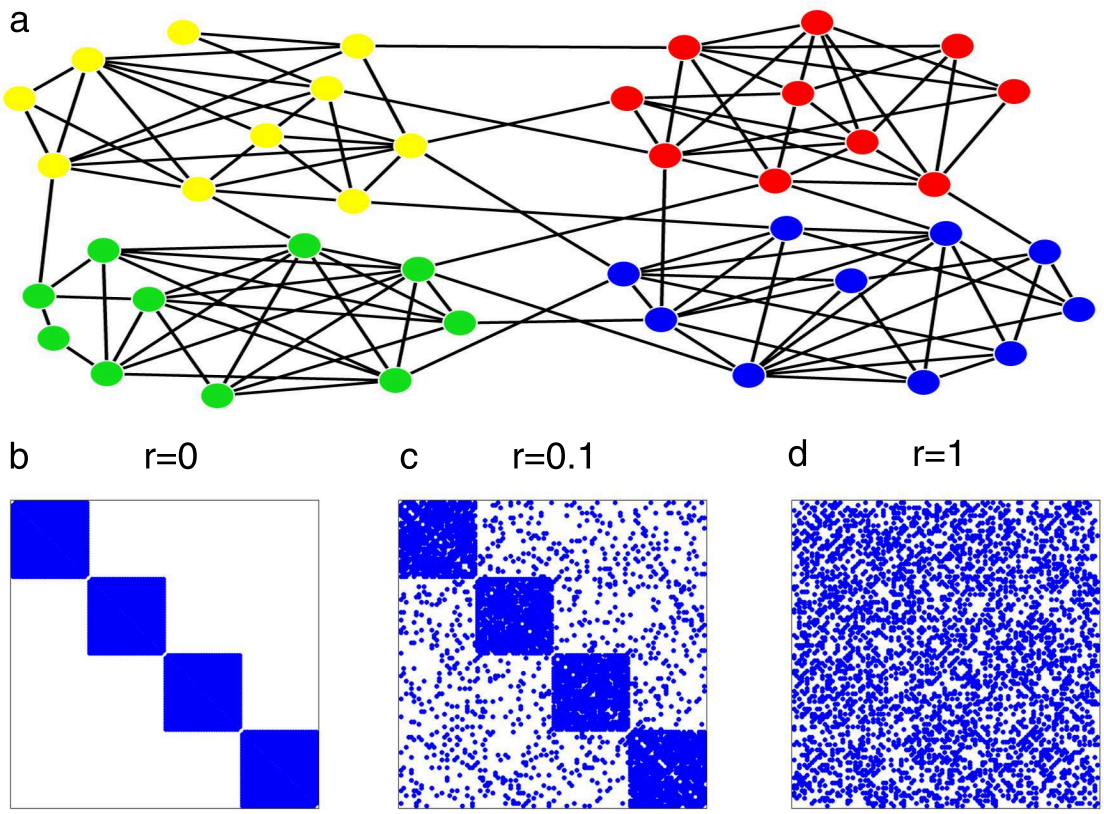

Figure 5.1: (a) Schematic representation of the modular network model having four modules whose members are indicated using different colors. (b-d) Adjacency matrices A defining the network connections at different values of the modularity parameter $r = \rho_{out}/\rho_{in} \in [0, 1]$, the ratio of inter-modular to intra-modular connection density for a fixed average degree k of the nodes. Starting from a collection of isolated clusters ((a), $r = 0$), by increasing r we obtain modular networks ((b), $r = 0.1$) eventually arriving at a homogeneous network ((c), $r = 1$).

The epidemic dynamics model we have used is the well-known SIRS ($Susceptible \rightarrow Infected \rightarrow Recovered \rightarrow Susceptible$) compartmental model(Fig. 5.2). Each node of the network represents an individual in the population that can be in any one of three possible states: susceptible (S), infected (I) or recovered (R). Each link between a pair of nodes is a contact along which an infection can propagate. Susceptible nodes having k_{inf} infected neighbors become infected with the probability of $q = [1 - (1 - \alpha)^{k_{inf}}]$, where α is the rate of infection transmission from an infected to a susceptible individual. Infected individuals recover at the rate β when their state changes recovered. For a epidemic

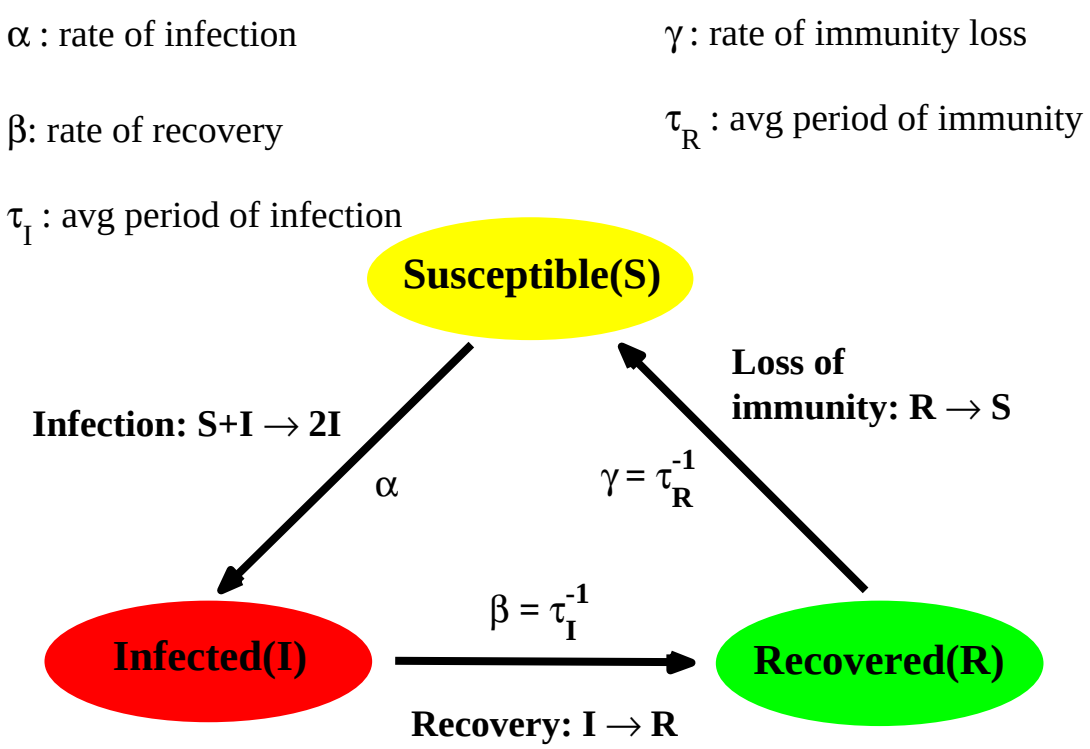

Figure 5.2: Schematic illustration of the dynamics of the SIRS model. The dynamical processes underlying the transition between the different states, viz., Susceptible (S), Infected (I) and Recovered (R) are shown. Initially most of the population is in the compartment S. Passage of a few individuals to the compartment I results in more and more individuals transferring from S to I over time at a rate α. At the same time, individuals move from compartment I to R as they recover at the rate β (= $1/\tau_I$, the period of infection). Over long times, individuals move from R back to S because of loss of immunity at the rate γ (= $1/\tau_R$, the recovery time).

having a fixed infection period τ_I, $\beta = 1/\tau_I$. A recovered individual cannot be infected owing to immunity provided by previous exposure to the disease. Finally, after a time τ_R, an individual loses immunity and becomes susceptible again. The transition from the recovered state to the susceptible state occurs at the rate $\gamma = 1/\tau_R$.

5.3 Results

We have generated modular networks with different sizes upto $N = 2048$ (most of the results shown here are for $N = 1024$) for different values of the modularity parameter r ranging between 10^{-4} and 1. Initially a randomly chosen 1 % of the nodes are infected while the remaining nodes are susceptible. We consider the long-term behaviour of the epidemic spreading on a modular network by simulating the process for more than 10^4 time steps and calculating the time upto which the disease persists in the system. The epidemic persistence time τ is a central issue in epidemiology [175]. After the outbreak of an epidemic the disease can either become extinct or remain endemic in the community. This has resulted in the definition of a critical community size (CCS), the minimum population required for a disease to become endemic. For populations whose size is lower than CCS the disease is expected to become extinct. As most theoretical studies of CCS have considered homogeneous random mixing, here we study how the presence of a realistic community structure in the population that will result in a modular contact network will affect the CCS for a highly infectious disease. In our simulations we continuous record the number of infected in the population at any given time instant t, $I(t)$, and declare a disease to have become extinct if at any time $I(t) = 0$ (note that $I(t-1), I(t-2), \ldots, I(1) > 0$). The time-step at which this happens is recorded as the persistence time. If the epidemic persists upto the time for which the simulation is carried out (typically 2×10^4 time steps), the epidemic is said to persist indefinitely, as this time is much longer than any of the time-scales of the SIRS epidemic model.

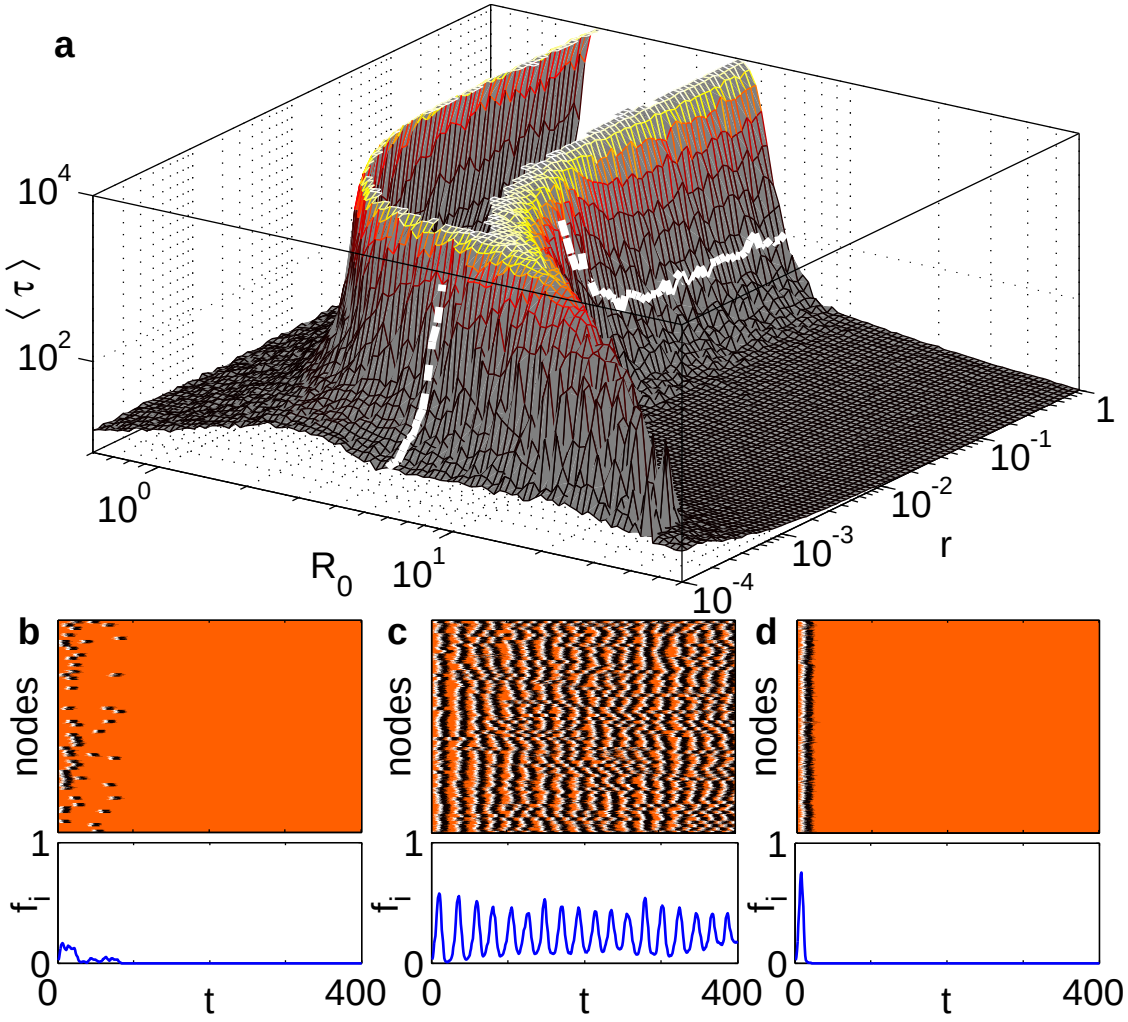

Figure 5.3: Modular organization of the contact network can make highly infectious diseases persistent. (a) The average persistence time of an epidemic in a modular random network shown as a function of the basic reproduction number $R_0(= \alpha k \tau_I)$ of the epidemic and the modularity parameter r. While the epidemic persists indefinitely in the homogeneous network ($r = 1$) for lower values of R_0, for a highly infectious epidemic (e.g., $R_0 = 6$, indicated by the thick broken line) the disease rapidly runs its course through a homogeneous network and becomes extinct. However, if the network is modular, e.g., when $r \sim 10^{-3}$, the epidemic becomes recurrent, persisting in the system for extremely long times. However, as one approaches the limit of the modules becoming almost isolated, the infection is unable to transfer from one module to another and the epidemic again becomes extinct rapidly. This is shown explicitly in the space-time diagrams and the time-series of the infected fraction of population for (b) $r = 2 \times 10^{-4}$, (c) $r = 2 \times 10^{-3}$ and (d) $r = 2 \times 10^{-2}$, for $R_0 = 6$. While both for the case of isolated modules (b) and the relatively homogeneous network (d), the epidemic becomes extinct within 100 time units, for an optimal range of modular organization (c) the epidemic persists for as long as the simulation is continued. For all simulation results shown here we have used a network of $N = 1024$ nodes having $M = 64$ modules of size $n = 16$ each. The nodes have average degree $k = 12$, with $\tau_I = 5$ and $\tau_R = 10$ time units.

Fig. 5.3 shows that modular organization of the contact network can make highly infectious diseases persistent. A disease having $R_0 < 1$ will not be able to initiate an epidemic regardless of the network structure as the number of secondary infections arising through contact are actually less than the number of initially infected individuals. Thus, as the infected fraction of the population $f_i(= I/N)$ rapidly decays to zero, the time for which the disease persists in the population is typically short. For $R_0 \geq 1$, relatively homogeneous networks (i.e., having higher values of r) show a rapid rise in the disease incidence characteristic of an epidemic as the network spreads the infection to a much larger fraction than that which had been infected initially. The infected fraction time-series then settles down to an irregular series of oscillations with the disease persisting in the population for the duration of simulation, provided R_0 is not too high. However, if R_0 is increased indefinitely, we eventually observe a very different long-term behavior where the entire population becomes infected in a short space of time followed by recovery and extinction of the disease. This can be understood in terms of a delay difference equation describing the time-evolution of infections under a mean-field approximation (where we can assume $k_{inf} = kf_i$). For such a case, the fraction of individuals who are infected at a particular time-step $n+1$, $x_{n+1} = [1-(1-\alpha)^{k\sum_{j=0}^{\tau_I-1} x(n-j)}][1-\sum_{j=0}^{\tau_I+\tau_R-1} x(n-j)]$.

When the network becomes modular, the effective threshold for epidemic effectively increases so that we only see the disease spreading through the entire population at values of R_0 much larger than 1. While the epidemic persists indefinitely in the homogeneous network ($r = 1$) for lower values of R_0, for a highly infectious epidemic (e.g., $R_0 = 6$, indicated by the thick broken line) the disease rapidly runs its course through a homogeneous network and becomes extinct. However, if the network is modular, e.g., when $r \sim 10^{-3}$, the epidemic becomes recurrent, persisting in the system for extremely long times. However, as one approaches the limit of the modules becoming almost isolated, the infection is unable to transfer from one module to another and the epidemic again becomes extinct rapidly. This is shown explicitly in the space-time diagrams and the time-series of the infected fraction of population for (b) $r = 2\times10^{-4}$, (c) $r = 2\times10^{-3}$ and (d) $r = 2\times10^{-2}$, for

$R_0 = 6$. While both for the case of isolated modules (b) and the relatively homogeneous network (d), the epidemic becomes extinct within 100 time units, for an optimal range of modular organization (c) the epidemic persists for as long as the simulation is continued.

Fig. 5.4 shows the variation of the probability distribution of persistence time τ with network modularity. The distribution of the time τ (with logarithmic binning) for which an epidemic persists in the network shows a bimodal nature for higher values of the modularity parameter r, with the upper branch diverging as the network becomes more modular. For lower values of r, the distribution is unimodal and the average value of τ decreases rapidly as the modules become effectively isolated. Fig. 5.4 (b) shows the probability that the epidemic persists for more than 10^4 time units as a function of the modularity parameter r for networks having different number of modules M.

In order to understand the simulation results we investigate the Laplacian spectrum of the contact network. The Laplacian matrix L for a network is defined as $L = D - A$, where A is adjacency matrix and D is the degree matrix whose only non-zero entries are along the diagonal that are the degrees of the constituent nodes. For an undirected network as we have considered here, the Laplacian matrix is symmetric and positive semi-definite, with nonnegative real eigenvalues $\lambda_1 = 0 < \lambda_2 \leq \lambda_3 .... \lambda_N$. The second eigenvalue λ_2 is nonzero if and only if the network is connected. The reciprocal of this eigenvalue corresponds to the time-scale at which a dynamical process such as synchronization or diffusion spread over the entire network. Fig. 5.5 (a) shows that this time-scale decreases with the modularity parameter r.

For a network with M modules, the difference of the reciprocals of the eigenvalues λ_M and λ_{M+1} define the spectral gap, which corresponds to the difference in the time-scales of fast intra-modular processes and slow inter-modular processes [63]. The existence of these distinct time-scales is a consequence of modular structure and is often taken as a signature of modular organization in a complex system. Fig. 5.5 (b) shows that the spectral gap decreases with the modularity parameter r. However, networks having

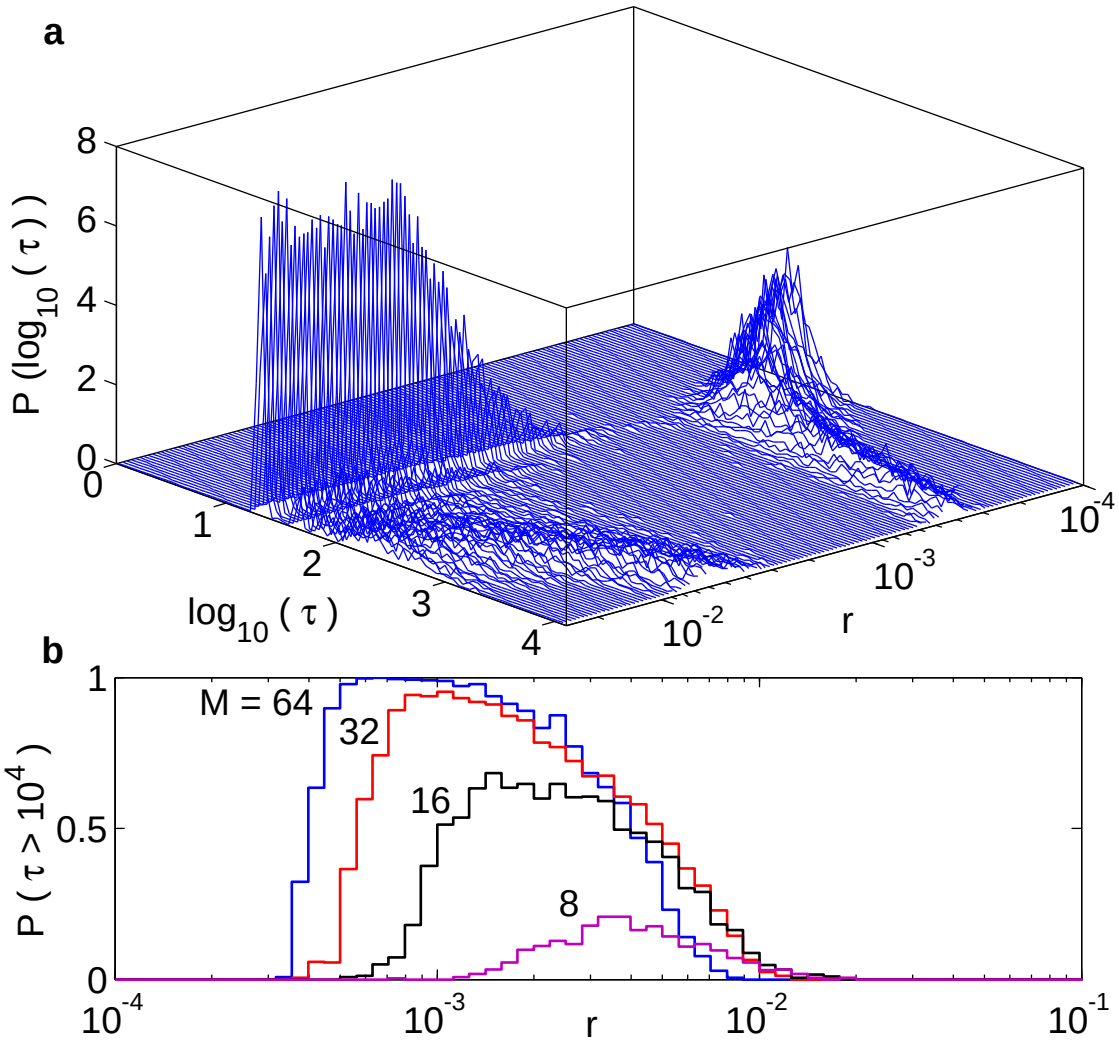

Figure 5.4: The variation of the probability distribution of persistence time τ with network modularity. (a) The distribution of the time τ (with logarithmic binning) for which an epidemic persists in the network shows a bimodal nature for higher values of the modularity parameter r, with the upper branch diverging as the network becomes more modular. For lower values of r, the distribution is unimodal and the average value of τ decreases rapidly as the modules become effectively isolated. Results shown for $N = 1024$ with $M = 64$ modules each with $n = 16$ nodes having average degree $k = 12$. (b) The probability that the epidemic persists for more than 10^4 time units is shown as function of the modularity parameter r for networks having different number of modules M ($N = 1024$ with each node having average degree $k = 12$.) For all results shown here $\alpha = 0.1$, $\tau_I = 5$ and $\tau_R = 10$ time units (i.e., $R_0 = 6$).

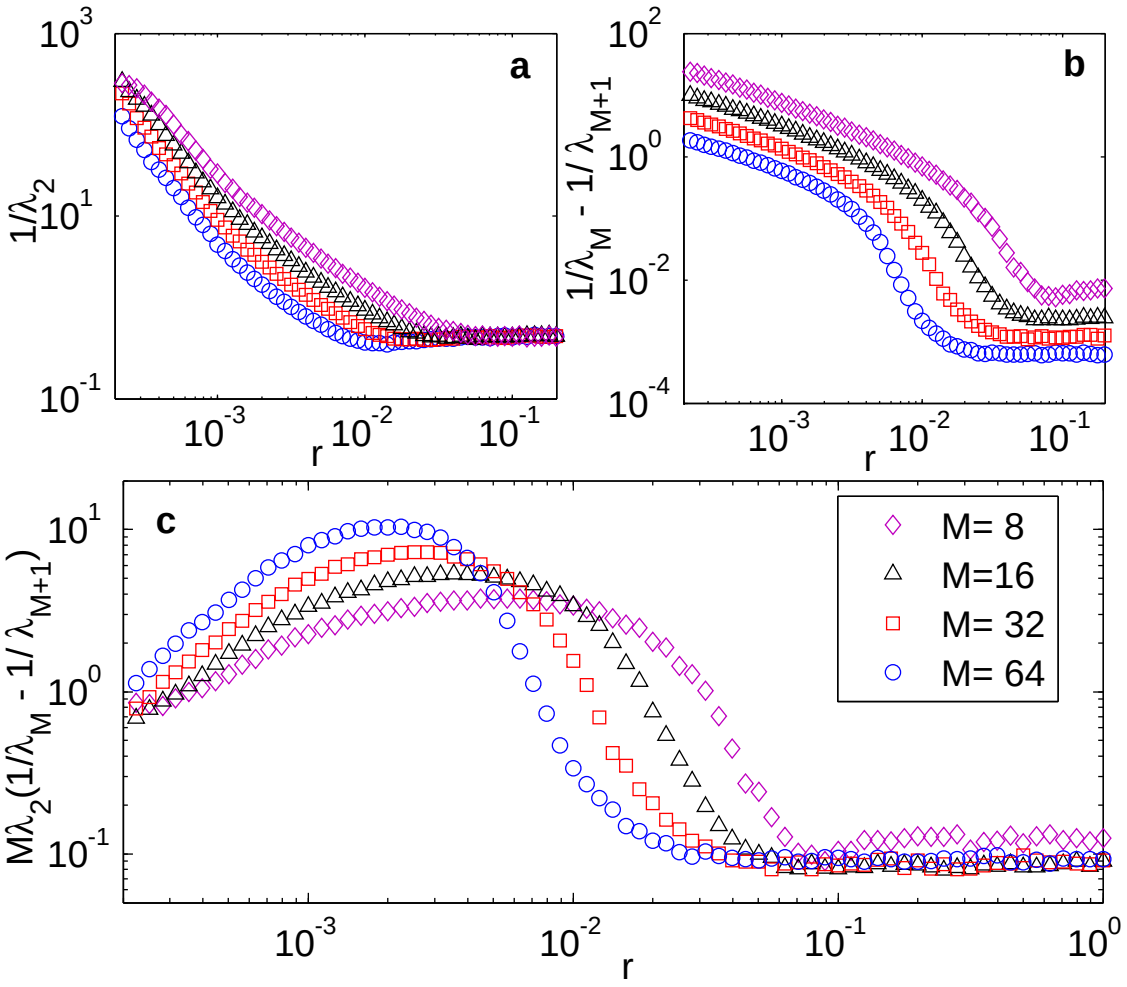

Figure 5.5: The variation of (a) the inverse of the smallest finite eigenvalue of the Laplacian matrix corresponding to the modular network and (b) the Laplacian spectral gap as a function of the modularity parameter r. They represent the global dynamical time-scale (i.e., for a dynamical even like synchronization or diffusion to take place over the entire network) and the time-scale separation between intra- and inter-modular processes respectively. The ratio of these two time-scales with the scaling factor M (number of modules) is shown in (c). The different symbols indicate data for networks having different number of modules explained in the key shown in panel (c). For all results shown here $N = 1024$ and $k = 12$.

different number of modules behave somewhat differently. By scaling with the number of modules M, we observe that the different curves collapse on each other.

Fig. 5.5 (c) shows the ratio of the two time-scales, viz., the global time-scale given by the reciprocal of the smallest finite eigenvalue and the spectral gap. We observe that this peaks at an intermediate value of r, implying that there is an optimal range of modularity where the epidemic propagates fast enough from module to module so as not to die out before spreading; at the same time, the inter-modular passage takes long enough so that when the epidemic returns to a region after cycling around the network, it would have recovered so that the nodes can be infected once more. Thus, the epidemic can persist indefinitely.

5.4 Discussion and Conclusion

In this chapter we have explored the persistence of epidemic dynamics on modular networks where individuals after having recovered from an infection can again becomes susceptible with a certain probability. Our study of SIRS (Susceptible-Infectious-Recovered-Susceptible) dynamics in a modular network suggests that under certain circumstances an epidemic can become persistently recurrent in a population. We show how the probability of persistence of an epidemic depends on the network mesoscopic organization as well as on individual dynamical parameters such as the infection rate. In particular, we show that highly contagious diseases, which quickly die out in a population with homogeneous contact structure, can survive indefinitely (becoming endemic) when there is strong community organization in the population.

6

Spatiotemporal patterns of incidence for a vector-borne infectious disease

6.1 Introduction

Malaria is one of the most important vector-borne infectious diseases in the modern world resulting in the death of more than a million people each year [177]. It is endemic in many developing countries, where the social and economic burden of this disease is considerable [178]. Malaria is caused by infection of parasitic protozoans of the genus *Plasmodium*. Humans and other animals get infected after being bitten by female Anopheles mosquitos that act as the vector. Four species of *Plasmodium* are known to infect human beings, the commonest being *Plasmodium vivax* while the most lethal is *Plasmodium falciparum* [179].

In this chapter we analyze the spatio-temporal patterns of incidence for malaria in a region of northern Bengal, India in order to understand the role of space in the diffusion dynamics of a vector-borne disease. We consider infections of both *Plasmodium vivax* and *Plasmodium falciparum* for a period extending from January 2005 and February 2009. As

epidemiological time-series data are generally noisy and non-stationary, resulting from climatic variations, changes in the socio-economic patterns of the population, large scale infra-structural projects etc., we subject our data to wavelet analysis [185] that reveals a dominant periodicity of corresponding to 1 year. Exploiting the wavelet phase relations between different regions, we identify the epicenters for both *vivax* and *falciparum* infections. While the former appears to have a single source from which the infection spreads out to the entire region, the latter appears to have two different epicenters. Our measurement of the correlation between phase angle difference and the physical distance separating different locations substantiate the spatio-temporal traveling wave nature of the spreading of malaria infections in this area. To account for the annual periodicity of malaria incidence, we look at the correlation with rainfall data and observe that indeed the pattern of infections appear to follow that of rainfall with a lag of 1-2 months for *Plasmodium vivax* and 2-3 months for *Plasmodium falciparum*. Using a spatially detailed malaria transmission model, we show that seasonal variations such as in rainfall that influence the vector emergence rate can indeed result in the incidence data showing the same periodicity as that of the environmental factor. Our results can potentially be used to identify pockets where vector eradication program can be intensified in order to control malaria.

6.2 Materials and Methods

Data

Numbers of *Plasmodium vivax* and *Plasmodium falciparum* cases for every month between January 2005 and February 2009 have been obtained from the Mal Sub-divisional Hospital at Malbazar, Jalpaiguri, West Bengal. Only cases where serological confirmation of malaria has been conducted is included. Data from each of the 51 different health centers (Fig. 6.1) which are responsible for the villages and tea estates in this area allow us to investigate the spatial spreading of the disease.

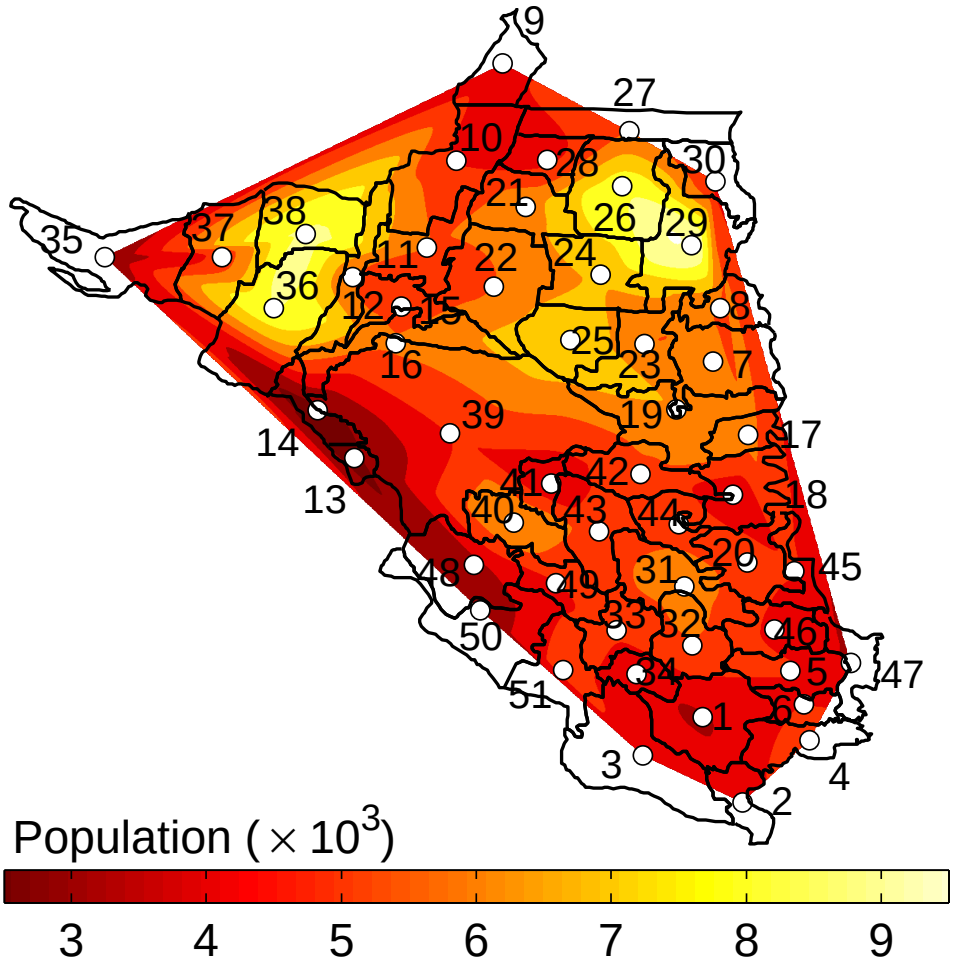

Figure 6.1: Population distribution in locations under different health centers in the rural sub-division. The boundaries of the 51 health centers are indicated and their identities represented by numbers 1-51 (the names of the individual health centers corresponding to these numbers is given in Table 6.1.

To look at the correlation between incidence data and periodic variations in the environment we have used monthly rainfall data (in mm) for Jalpaiguri district obtained from the website of the Indian Meteorological Department, Government of India. In our analysis we used data for the period July 2004 to February 2009.

Wavelet time series analysis All incidence time-series are logarithm transformed (after adding a constant of one) and subsequently scaled to have zero mean and unit variance. We analyze temporal changes in the distribution of power at different scales s (approximately periods) using a Morlet basis function. The Morlet function is essentially a damped complex exponential, which can capture local (in time) cyclical fluctuations in the time series.

The continuous wavelet transform (CWT) of the incidence time series is calculated as the convolution of the incidence data with a scaled and translated version of the Morlet basis function [182, 185]. For this basis, scale is approximately equal to the period obtained from Fourier analysis. Therefore, the lowest scale corresponds to the maximum (Nyquist) frequency of 0.5 periods per time step. The local wavelet power spectrum at a specific time point and a particular scale is given by the square of modulus of the CWT. We have used the MATLAB program made available publicly by Torrence and Compo [180].

The cone of influence shown in Figs. 6.2-6.3 indicates the loss in statistical power near the start and end of the series as a result of edge effects. The width of the cone gives a lower limit on how wide a feature needs to be at a given scale for it to represent genuine cyclical behavior, rather than a spike. Significance tests were done using methods described and discussed in Ref. [180].

Phase relationships between time series

A wavelet transform $W_n(s)$ obtained using the Morlet basis function has a phase angle defined by

$$\Theta(s) = tan^{-1}\left[\frac{\Im\{W_n(s)\}}{\Re\{W_n(s)\}}\right],$$

where $\Re\{W_n(s)\}$ and $\Im\{W_n(s)\}$ are the real and imaginary parts, respectively, of the wavelet transform $W_n(s)$ of the time series. The time series of phase angles Θ associated with the wavelet transform corresponding to the dominant time-scale (=1 year) has been reconstructed for each of the 51 Health Centers for both *Plasmodium vivax* and *Plasmodium falciparum* infections. These series are used to compute phase differences between different regions, restricting them to the range $\pm\pi$.

6.3 Results

The population distribution in the area we have investigated is shown in Fig. 6.1, the color indicating the number of people who fall under the jurisdiction of each of the 51 health centers into which the entire area is divided. It shows that HC-26 (Rangamati tea estate), HC-29 (New Mal Panchayat), HC-36 (Leesh river tea estate) and HC-38 (Chanda company) are relatively high population areas, while HC-39 (Sologhoria) has the least population density, as it is a large forested area. It has neighboring regions with relatively high populations, e.g., HC-40 (SouthHanskhali), HC-41 (Anandapur tea estate), HC-42 (Baroghoria), HC-43 (Dhalabari) and HC-44 (Kodalkati) (see Table 6.1 for identities of the regions indicated by numbers in Fig. 3.2).

We first perform wavelet analysis on the monthly time series of the entire area for both *Plasmodium vivax* (Fig. 6.2) and *Plasmodium falciparum* (Fig. 6.3) cases and identify the dominant periodicity to correspond to 12 months, i.e., 1 year. Next, the time series of *falciparum* and *vivax* for each of the different health centers are subjected to wavelet decomposition. For cycles with a given period, the wavelet analysis generates a phase angle at each time step. From the wavelet decomposition for the 12 month period, the time series of wavelet phase angles is obtained for each of the health centers.

As the number of cases go from a trough to a peak, the phase angle increases from $-\pi$ to 0, and as the number of cases cycles back from a peak to a trough, the phase angle

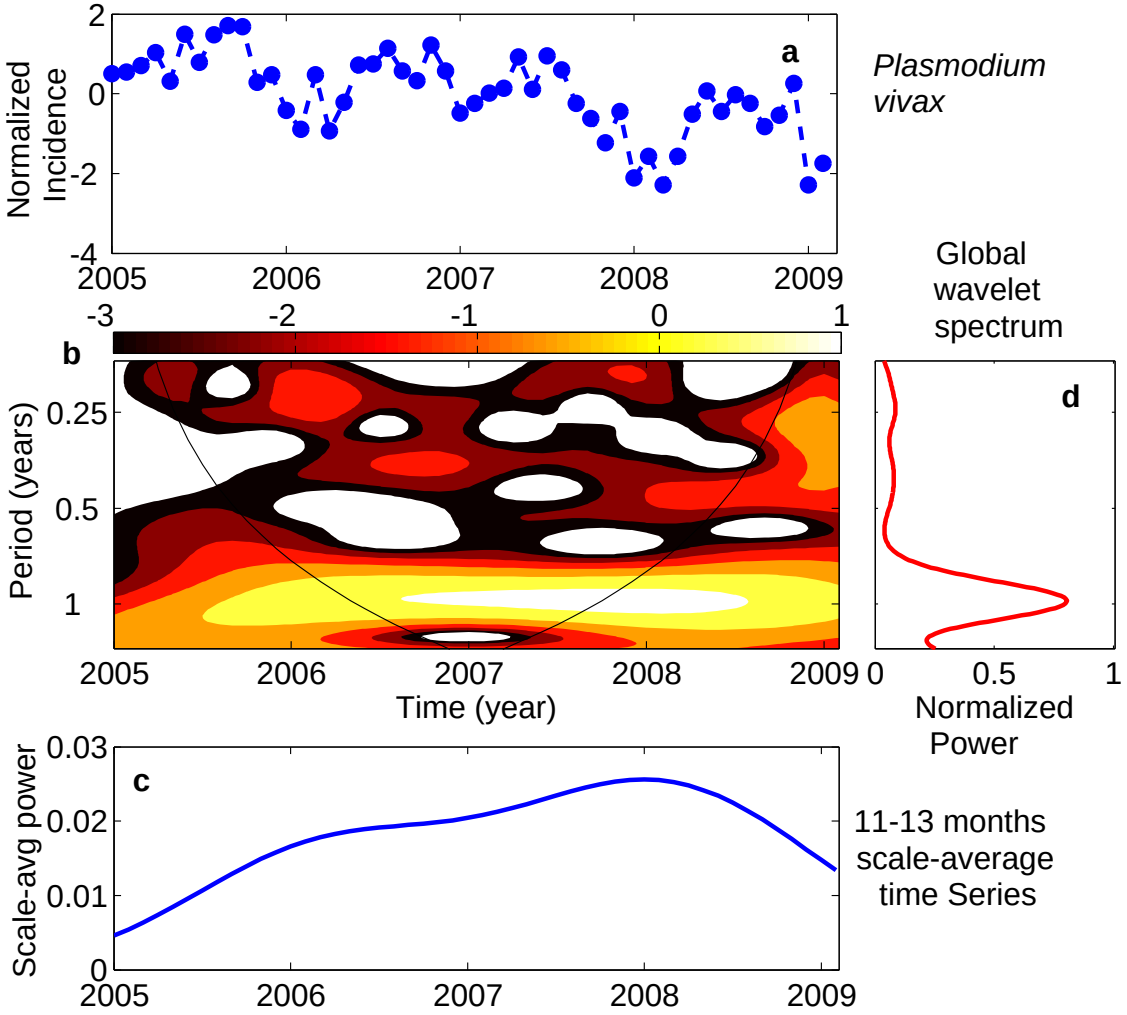

Figure 6.2: Wavelet time series analysis for *Plasmodium vivax* incidence in the sub-division. (a) The time-series of the monthly number of *Plasmodium vivax* cases reported for the entire sub-division during January 2005-February 2009. The incidence data has been first been log-transformed and then scaled to have zero mean and unit variance. (b) Local wavelet power spectrum (LWPS) with the Morlet basis function, normalized by variance (σ^2). The power is indicated by the color coding, the key to which is given above the figure. The ordinate indicates the periodicity (in years). The curve represents the "cone of influence" below which boundary effects cannot be neglected. (c) The global wavelet spectrum over the entire period being considered, indicating a dominant time-scale corresponding to 1 year. (d) The variation of the power with time, averaged over the scales corresponding to 11-13 months.

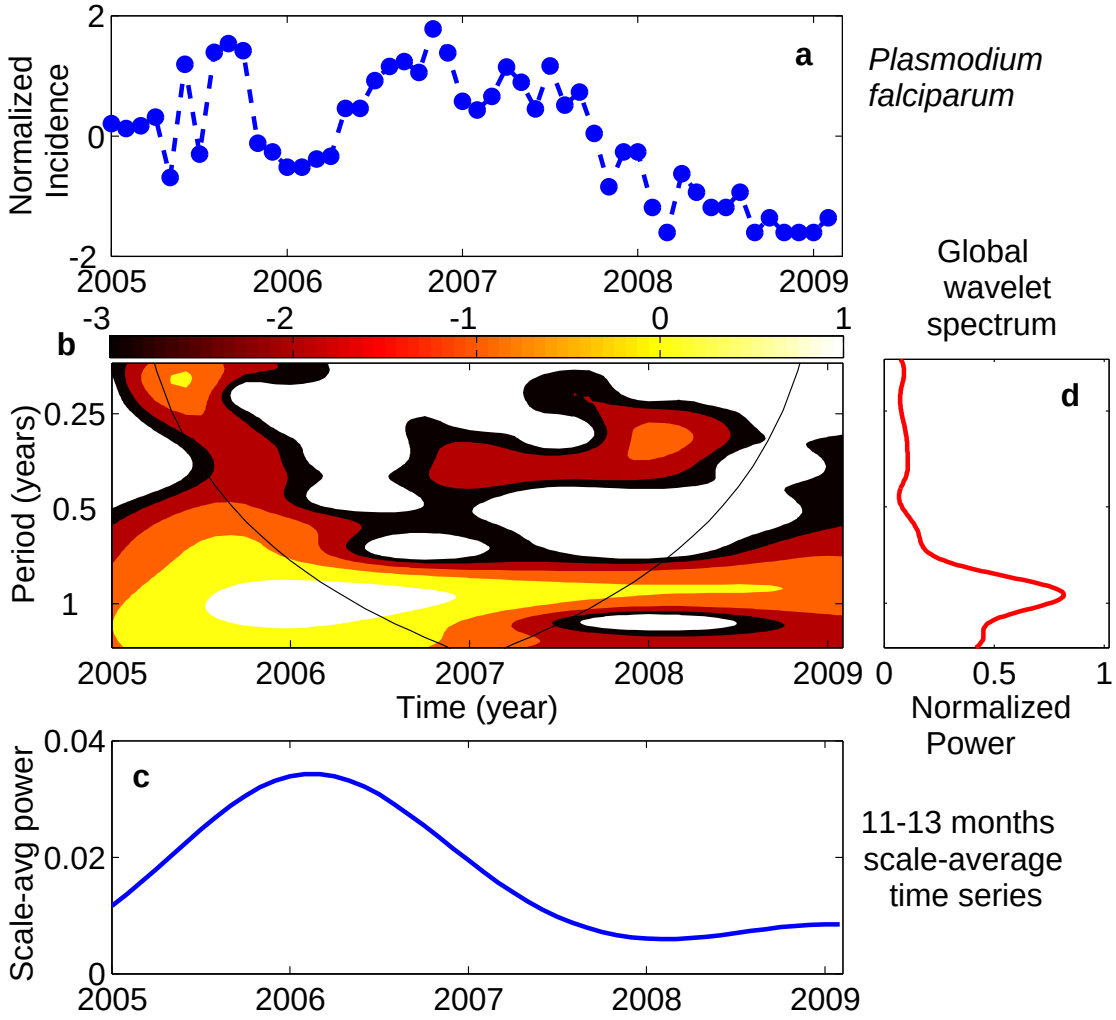

Figure 6.3: Wavelet time series analysis for *Plasmodium falciparum* incidence in the sub-division. (a) The time-series of the monthly number of *Plasmodium falciparum* cases reported for the entire sub-division during January 2005-February 2009. The incidence data has been first been log-transformed and then scaled to have zero mean and unit variance. (b) Local wavelet power spectrum (LWPS) with the Morlet basis function, normalized by variance (σ^2). The power is indicated by the color coding, the key to which is given above the figure. The ordinate indicates the periodicity (in years). The curve represents the "cone of influence" below which boundary effects cannot be neglected. (c) The global wavelet spectrum over the entire period being considered, indicating a dominant time-scale corresponding to 1 year. (d) The variation of the power with time, averaged over the scales corresponding to 11-13 months.

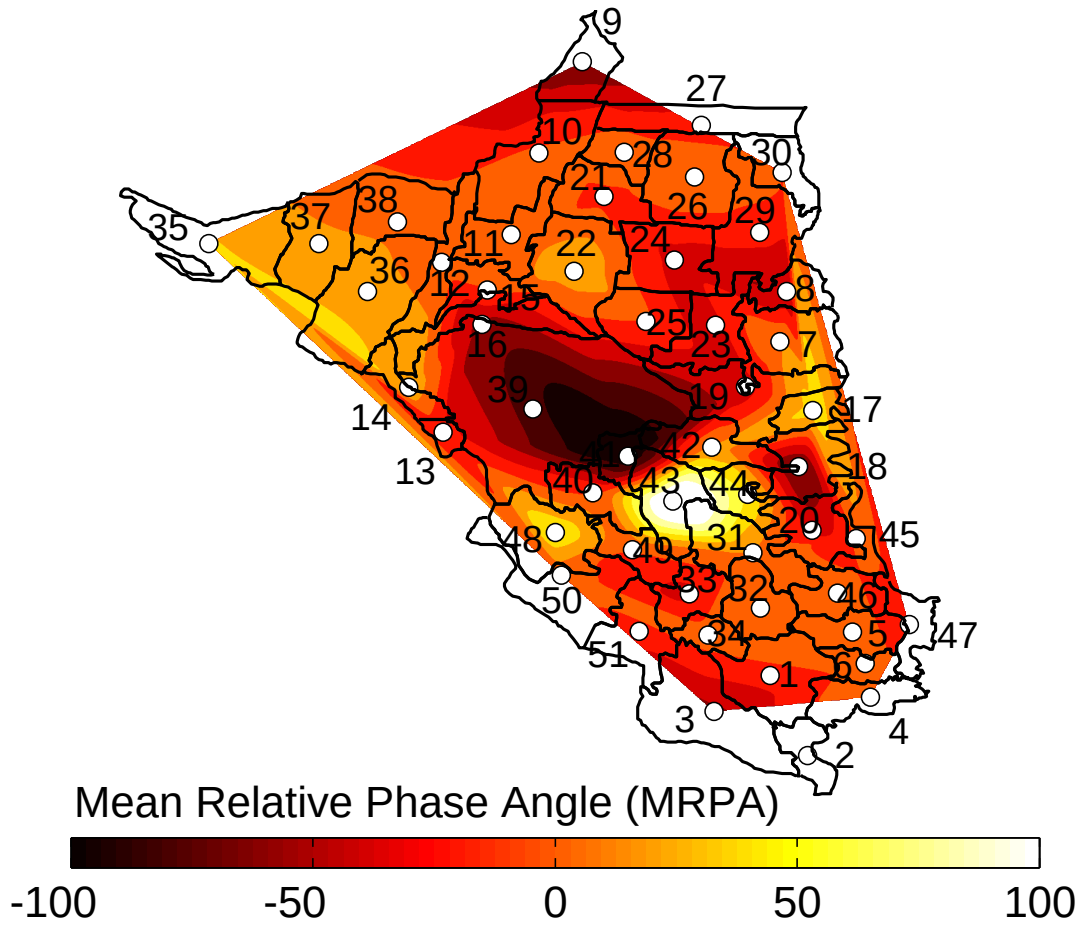

Figure 6.4: Epicenter of *Plasmodium vivax* incidence. The spatial distribution of interpolated mean relative phase angles (MRPA) obtained from a series reconstructed from wavelet spectral components corresponding to the dominant period of 1 year indicates that the infections spread throughout the region from an epicenter located in region under the health center HC-43.

continues to increase from 0 to π. The time-series of relative phase angles are calculated for each health center by subtracting the temporal vector of spatial mean phase angles from the corresponding temporal vector of phase angles. A mean relative phase angle (MRPA) is calculated for each health center by averaging the vector of time-specific relative phase angles. The MRPA is the average phase angle of a given location relative to the spatial average over all regions. Thus, the health center with the highest MRPA represent the locations where a infection outbreak originates (i.e., epicenter), while regions having smaller MRPAs indicate where the infection spread subsequently. Fig. 6.4 shows that the

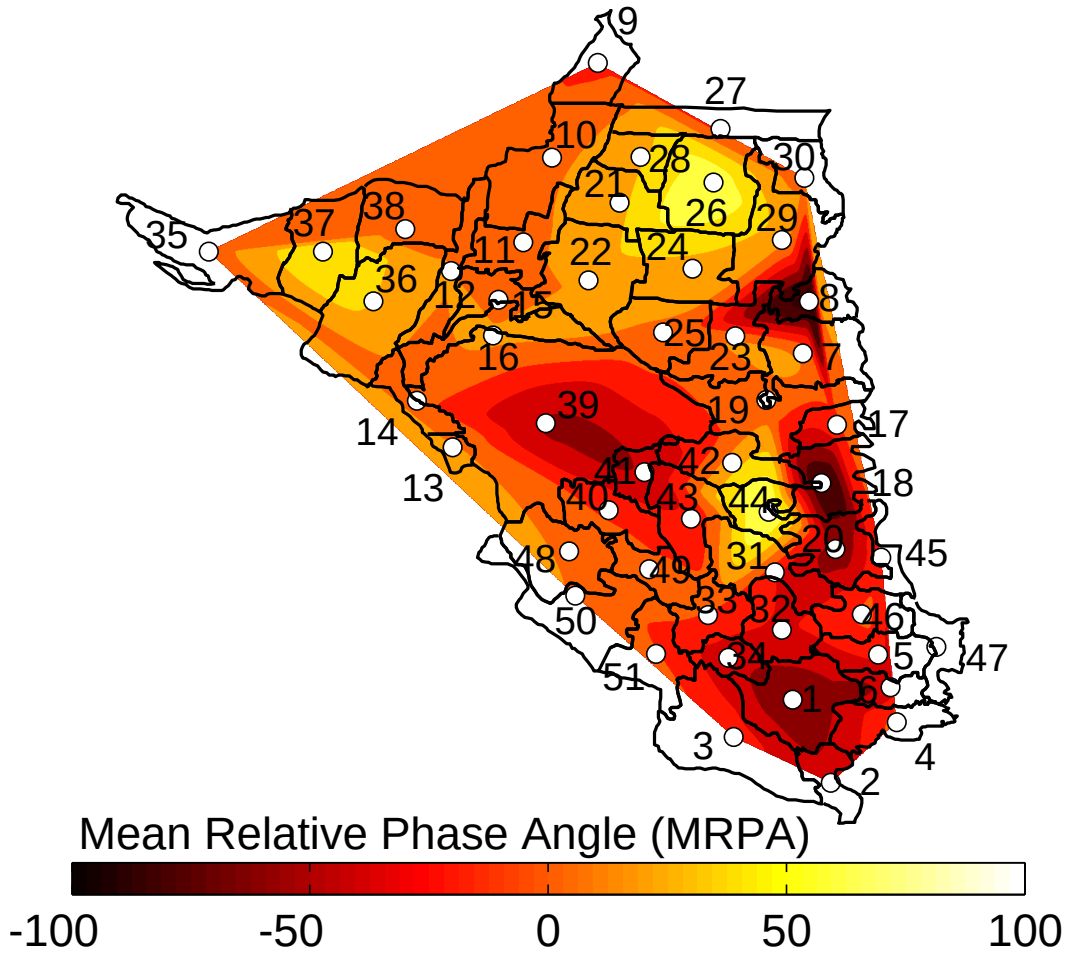

Figure 6.5: Epicenters of *Plasmodium falciparum* incidence. The spatial distribution of interpolated mean relative phase angles (MRPA) obtained from a series reconstructed from wavelet spectral components corresponding to the dominant period of 1 year indicates that the infections spread throughout the region from two different epicenters which are located in regions under the health centers HC-26 and HC-44 respectively.

region under health center HC-43 has been identified as the epicenter of *Plasmodium vivax* infection. Note that, this is a region which does not have very high population. Thus, this result contrasts with similar investigations carried out for diseases spread by direct transmission between infected individuals (such as measles) [182] where the epicenter coincides with the region having largest population. We surmise that despite the low population, HC-43 plays a key role in spreading malaria as it is a important crossroad where high-volume transportation routes going through the area intersect. In addition, it is also located next to a major irrigation canal that was under construction during the period the data was collected.

Fig. 6.5 shows that the *Plasmodium falciparum* infection appears to have two different epicenters located some distance apart from each other. One of these (HC-44) is adjacent to the epicenter HC-43 for *Plasmodium vivax*, so that the same reasons attributed to HC-43 being a source of infection may apply here. The other epicenter, HC-26, occurs in a region which has one of the highest populations in this area. It is also located close to the sub-divisional administrative headquarters, suggesting that people from areas outside the sub-division may be frequent visitors. The infection may thus enter the area under study from other areas through this region. The spatio-temporal propagating wave nature of the infections is further supported by the correlation between the phase angle differences and physical distance measured for different regions from the epicenters HC-43 and HC-26 (Figs. 6.6 and 6.7, respectively).

As already mentioned, the incidence time-series shows a dominant periodicity corresponding to 1 year. While there can be multiple environmental factors that are related to the spreading of malaria, here we have focused on seasonal variations in rainfall to account for this observed annual period in malaria incidence. Fig. 6.8 shows that the pattern of infections appear to follow the variation in rainfall with a short delay. We quantitatively establish this by calculating the time-delayed correlation between rainfall and incidence of *vivax* and *falciparum* (Fig. 6.9). The peak in correlation of *vivax* is seen for a delay

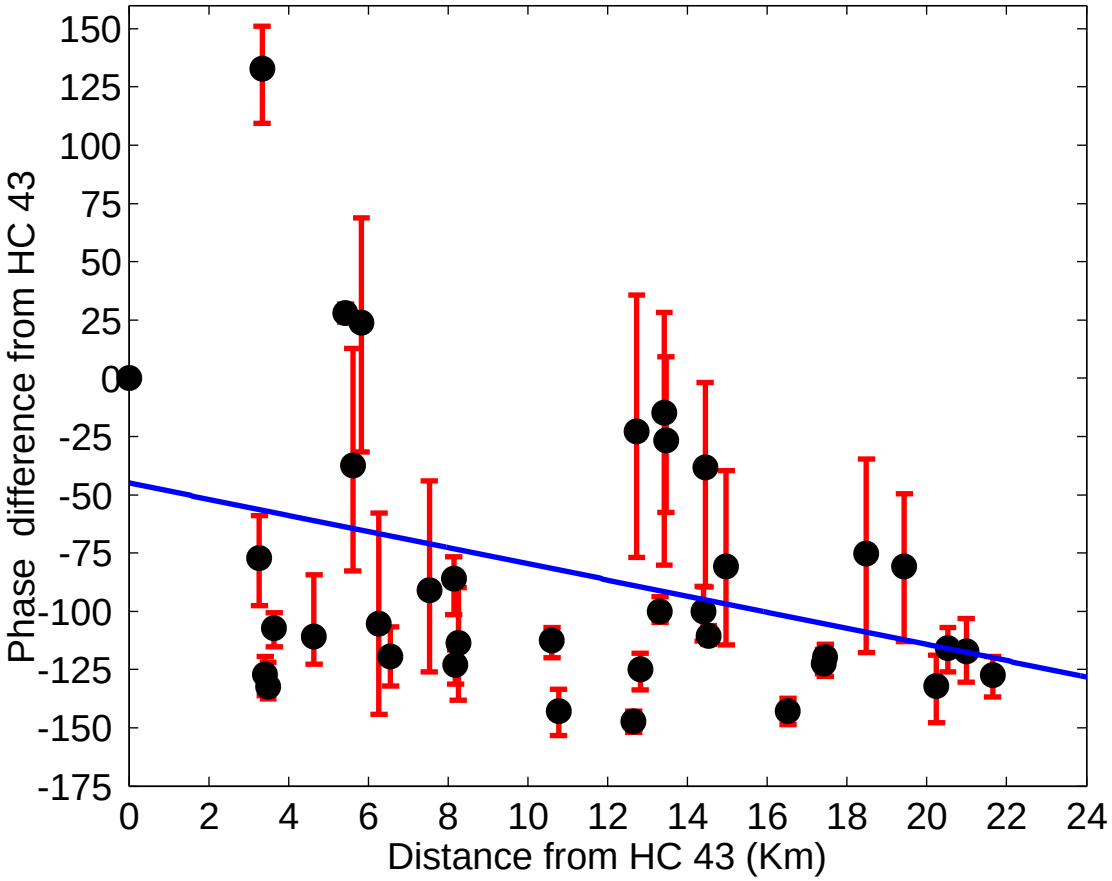

Figure 6.6: The pattern of spreading of *Plasmodium vivax* infection from the region under health center HC-43. Mean phase difference from HC-43 for the neighboring areas (served by 37 different health centers) shown as a function of the distance from the health center HC-43. Within 22 kms of the health center, there is a significant correlation between phase angle difference and distance as indicated by the correlation coefficient $r_{cor} = -0.35$ (95% bootstrap limits: -0.57 to -0.09 for 1000 bootstraps). The error bars are 99% bootstrapped confidence limits.

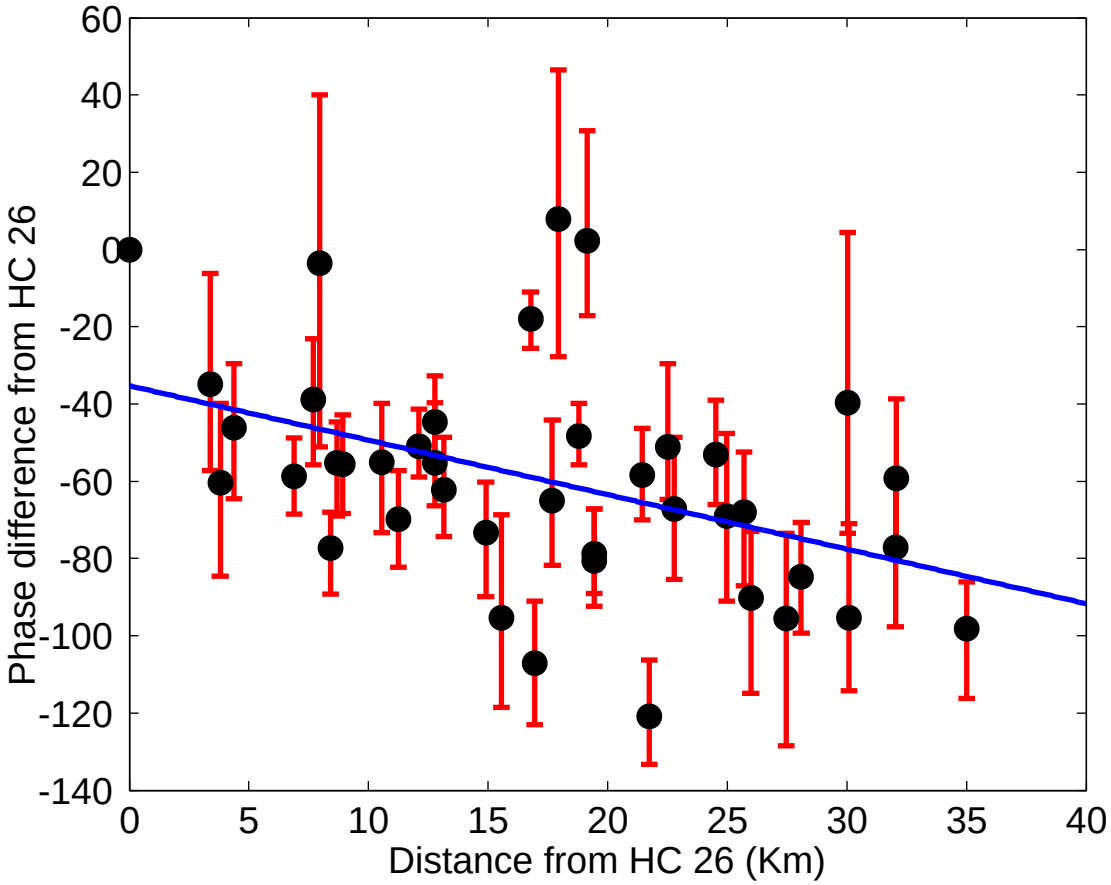

Figure 6.7: The pattern of spreading of *Plasmodium falciparum* infection from the region under health center HC-26. Mean phase difference from HC-26 for the neighboring areas (served by 41 different health centers) shown as a function of the distance from the health center HC-26. Within 35 kms of the health center, there is a significant correlation between phase angle difference and distance as indicated by the correlation coefficient $r_{cor} = -0.43$ (95% bootstrap limits: -0.65 to -0.21 for 1000 bootstraps). The error bars are 99% bootstrapped confidence limits.

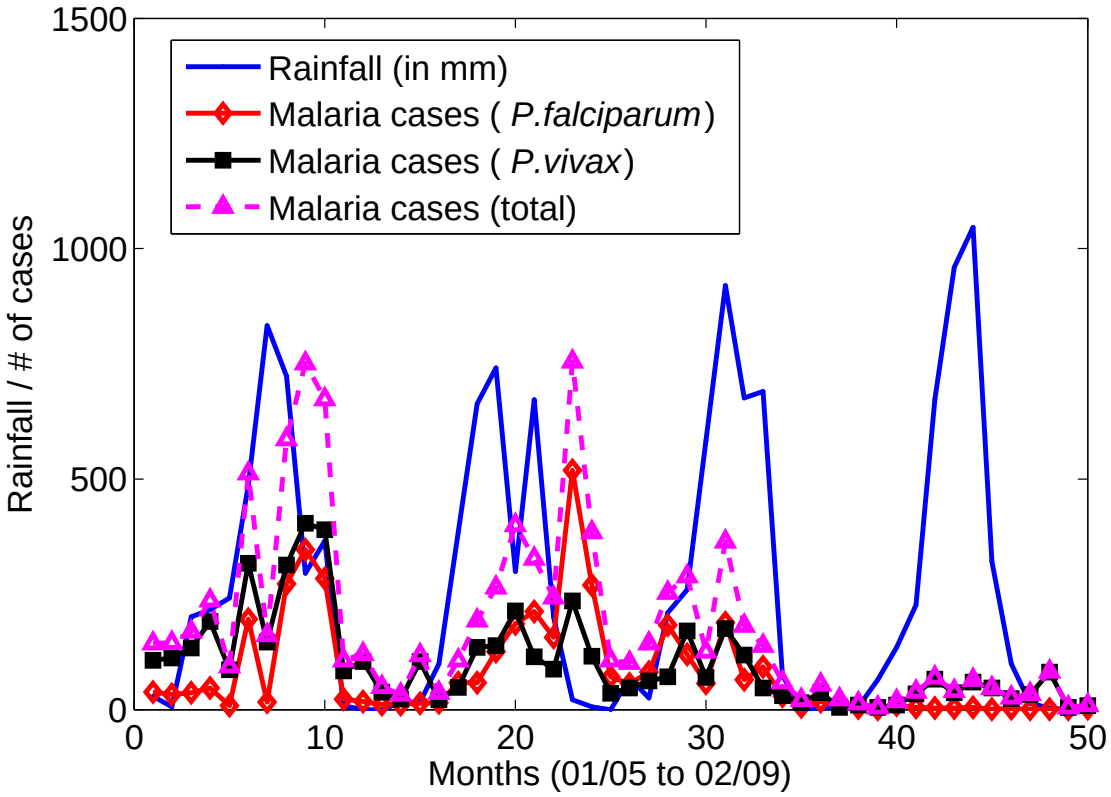

Figure 6.8: Monthly time-series of rainfall and malaria incidence in the sub-division during the period of January 2005 to February 2009. It can be seen that variation in the incidence of malaria types follows that of rainfall with a short delay in time.

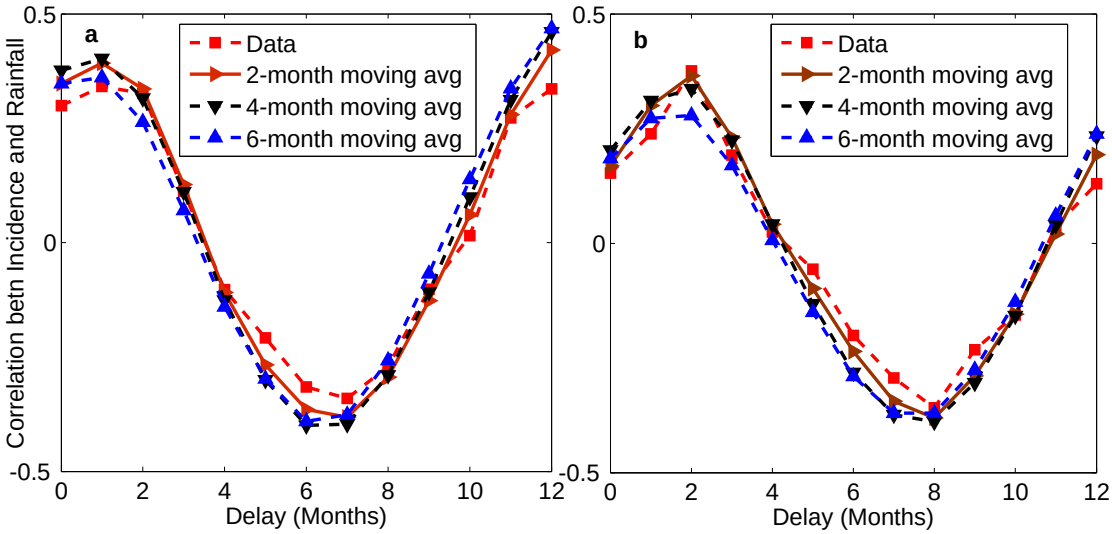

Figure 6.9: Correlation between rainfall and malaria incidence. (a) The time-delayed correlation calculated between the amount of rainfall and the number of *Plasmodium vivax cases* reported every month shows a peak for a delay period of 1-2 months. (b) The corresponding time-delayed correlation between the number of *Plasmodium falciparum* cases and amount of rainfall shows highest value for a delay period of 2-3 months.

of 1-2 months, while the peak for *falciparum* is observed for a delay of 2-3 months. We conjecture that the difference in the delay periods for *vivax* and *falciparum* has to do with the different geographical nature of their epicenters.

Malaria transmission model

In order to further establish the origin of the annual periodicity of incidence as resulting from the temporal variations in rainfall, we take recourse to quantitative modeling of malaria transmission in a spatially detailed scenario. We use a network model incorporating spatial heterogeneity and time varying emergence of mosquitoes, with the dynamics of individual nodes being similar to model proposed earlier by Smith *et al.* [184]. In our model the network has $N = 51$ nodes corresponding to the 51 health centers in the empirical data. The interaction between the nodes is governed by an adjacency matrix A_{ij} which is constructed from the matrix of actual physical distances between the health centers. Thus, a link is assumed to exist between two nodes i and j, i.e., $A_{ij} = 1$, if the physical distance between the corresponding pair of health centers is less than d_c km, and $A_{ij} = 0$, otherwise. In the results shown here, $d_c = 6.5$ km. However, our results are not sensitively dependent on the exact value of d_c.

To describe the dynamics of the epidemic at each node, we define X_i as the proportion of humans in node i who are infected (H_i being the human population density in that node) and Z_i as the density of infectious mosquitoes (M_i being the population density of mosquitoes in that node). The local dynamics is given by the following equations:

$$\frac{dX_i}{dt} = ab\frac{Z_i}{H_i}(1 - X_i) - rX_i + D_c\left(\sum_j^N A_{ij}X_j - k_iX_i\right),$$

$$\frac{dM_i}{dt} = \epsilon_t - gM_i,$$

$$\frac{dZ_i}{dt} = acX_i(M_i - Z_i) - gZ_i,$$

where k_i is the degree of node i, ϵ denotes the rate at which adult mosquitoes emerge from

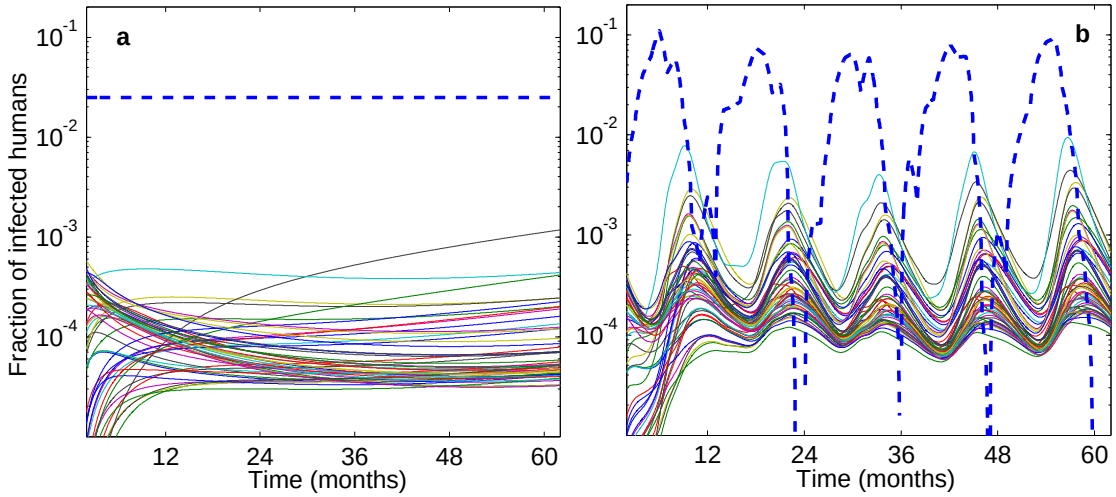

Figure 6.10: Time-evolution of the fraction of infected humans in a spatially extended malaria transmission model. (a) In the absence of any temporal variation in the parameters, the time series for different regions show monotonic variation. (b) When the rate of emergence of adult female mosquitoes from the larval habitat (ϵ_t) varies over time, correlated with the recorded rainfall in that area (having a mean-value $\langle\epsilon\rangle = 0.025$ identical to its constant value in the time-independent case), the time-series for different HCs show seasonal peaks with a period of 1 year. In both cases $r = 0.01$, $g = 0.1$, $a = 0.1$, $b = 0.3$, $c = 0.3$ and $D_c = 0.01$. Initial conditions used are $M_i = 2H_i$ and $X_{43} = 0.01$.

larval habitat, $1/r$ is the average duration of infection period in days, $1/g$ the mean lifetime of mosquito, a is the human feeding rate of a mosquito, b is the probability an uninfected human becomes infected through a single bite from an infectious mosquito, and c is the probability that a mosquito becomes infected from biting an infectious human host. Note that the emergence rate ϵ depends on rainfall [186], and should thus be considered as a time-varying parameter ϵ_t. In fact, we construct the time-series for ϵ_t by interpolating the empirical rainfall data. This is because rainfall generates new possible breeding sites as well as enhances the existing ones. Thus, malaria incidence generally increases with rainfall; however, excessive rainfall can reduce transmission due to flushing out of aquatic breeding habitats [186].

Fig. 6.10 shows that in the absence of any temporal variation in the parameters, none of the nodes show any periodic (or even, non-monotonic) variation in malaria incidence. However, on driving the emergence rate ϵ by the temporal variation of rainfall obtained empirically, we immediately observe seasonal peaks in the incidence corresponding to a

period of 1 year. This suggests that periodically varying environmental factors, especially rainfall, can impose their periodicity on the incidence of certain vector-borne diseases.

6.4 Discussion and Conclusion

In this chapter we have analyzed the time-series data of malaria infection in 51 different health centers located in a region of norther Bengal by using wavelet analysis. Through the use of relative phase angles obtained from wavelet reconstruction of the time-series we establish that infections spread in the form of travelling waves emerging from specific epicenters. The wave nature of spreading is supported by the correlation between phase angle difference and the physical distance between different health centers. The identified epicenters for *Plasmodium vivax* and *falciparum* appear to be distinct and may be explained on the basis of being located in the most favorable conditions for dense mosquito habitat which includes water stagnation and forest coverage, relatively high human populations or well-connected landscapes. We also observe that rainfall significantly affects the dynamics of malaria incidence with the pattern of infections following that of rainfall with a short time delay. The importance of environmental seasonal variations is further established by using a spatially detailed model of malaria transmission. In the absence of any temporal variations in the parameter values, the model exhibits monotonic temporal behavior in infections. However, if the emergence rate of adult mosquitos are driven by periodic variations in rainfall, we immediately observe seasonal peaks in the infection time-series which have a period of one year.

Table 6.1: **Identities of the 51 health centers (HCs) in the rural sub-division from which malaria incidence data has been collected for analysis.**

Health center id	Health center	Health center id	Health center
1	Chapadanga	27	Meenglass
2	Basususba	28	SyleeTE
3	Babupara	29	NewMalPanchayat
4	SouthMatiali	30	ToonBari
5	ChakMaulani	31	KrantiSC
6	SouthChakMaulani	32	UttarSaripakuriP
7	WestTesimla	33	UttarKhalpara
8	HaiHaiPathar	34	SouthKhalpara
9	PatharJhoraTE	35	Ellenbury
10	Manabari	36	LeeshRiverTE
11	OdlabariP	37	Washabari
12	HindiSchool	38	ChandaCompany
13	Gajaldoba10	39	Sologhoria
14	Gajaldoba7	40	SouthHanskhali
15	SouthOdlabari	41	AnandapurTE
16	OdlabariTG	42	Baroghoria
17	Baradighi	43	Dhalabari
18	Kumlai	44	KodalKati
19	EastDamdim	45	Lataguri
20	NeoraNadiTG	46	UttarMatiali
21	RanicheraTE	47	JharMatiali
22	DamdimSC	48	Gachimari
23	BaintguriTE	49	JogeshChTE
24	KumlaiTE	50	WestDolaigaon
25	DamdimTE	51	Karaibari
26	RangamatiTE		

7

Conclusions

The work described in this thesis is a contribution towards developing a a systems-level description of biology by focussing on the different networks that are relevant in the microscopic and macroscopic dynamics of host-pathogen interactions. Although the components of the networks analyzed in the various chapters are quite different, ranging from kinase proteins to human individuals, focusing on the network description of these systems allow us to observe the functional importance of mesoscopic structural organization, in particular motifs and modules, of these networks. In the following subsections, the important results and conclusions reported in the thesis are summarized. We conclude with a brief discussion of possible future extensions of our results.

7.1 Summary of main results

Branched motifs enable long-range interactions in signaling networks through retrograde propagation

One of the most challenging problems in biology is to understand how robust yet sensitive coordination of response to stimuli is achieved in the intra-cellular signaling network.

The molecular components of this network are often arranged into modules containing branched motifs where the signal from an upstream signaling intermediate is channeled through two (or more) parallel pathways which can be counteractive. To generate appropriate cellular response, the activity in the parallel branches need to be suitably regulated, possibly through direct interaction between them such that activation of an intermediate in one pathway inhibits an intermediate in the other pathway. By contrast, we show in this chapter that even in the absence of such direct interactions, long-range coordination is possible between branches of signaling pathways through retrograde propagation of information. Thus a high level of homeostatic regulation of cellular responsiveness to stimuli can be achieved without a concomitant increase in the connection complexity of the signaling network. An important aspect of retrograde propagation in branched pathways that is distinct from previous work on retroactivity focusing exclusively on single chains is that varying the type of perturbation, e.g., between pharmaceutical agent mediated inhibition of phosphorylation or suppression of protein expression, can result in opposing responses in the other branches. This can have potential significance in designing drugs targeting key molecules which regulate multiple pathways implicated in systems-level diseases such as cancer and diabetes. Our results can potentially explain the evolutionary advantage of pathogens which target only selected intra-cellular signaling components as they do not need to devise extremely complicated interception strategies involving many types of molecules in order to survive within the host.

Mesoscopic organization of cancer gene network

The focus in cancer research has been gradually shifting away from the study of individual molecules and the effect of single gene mutations to an emerging consensus that it is a complex disease involving significant disruptions of the intra-cellular signalling network. One of the drawbacks of a network-based approach to analyzing cancer is the immensely large number of cellular agents whose interactions need to be investigated.

We have tried to solve this problem by taking a mesoscopic view of the cancer network, built up by considering the bipartite network of 146 tumor types and 927 cancer genes (and the corresponding proteins). Projecting this data onto a single network, we construct a largest connected component of 910 genes. Partitioning of this network yields 25 communities, with genes within each community having much stronger interaction with each other than with members belonging to other communities. We then project this onto the modular decomposition of the largest connected component of the human protein-protein interaction network having 9270 proteins grouped into 542 communities. Considering the distance of the cancer gene communities in the abstract protein-modular space allows us to build a relational dendrogram between different tumor types (as well as between different classes of cancer) which allows us to appreciate the relations between different types of tumors (and cancer categories). For example, our analysis shows that the hormonally related disease types of breast cancer and ovarian cancer indeed occur very close to each other in the dendrogram hierarchy. We also investigate the functional role of different cancer genes as revealed by their importance in the modular organization of the network by investigating the joint distribution of their participation coefficients and their within module degree z-scores. We have identified about 36 genes as "connector hubs" that occupy crucial positions in the cancer network and which can be potential targets for therapeutic efforts.

Epidemiological dynamics of the 2009 influenza A(H1N1)v outbreak in India

Influenza is a viral disease that has periodic breakouts almost every year in different parts of the world. In the last few centuries there have been several pandemics of this disease, the most infamous being the 1918-19 "Spanish flu" that killed about 50 million people worldwide. It is important to understand the dynamics of the initial stages of such pandemics in order to come up with possible control strategies. In this chapter we have

analyzed the incidence time-series data in the initial stage of the A(H1N1)v influenza pandemic in India during the period June 1- September 30, 2009. Using a variety of statistical fitting procedures, we obtain a robust estimate of the exponential growth rate $\langle\lambda\rangle \simeq 0.15$ in the number of infections. This corresponds to a basic reproduction number $R_0 \simeq 1.45$ for influenza A(H1N1)v in India, a value which lies towards the lower end of the range of values reported for different countries affected by the pandemic. Our study indicates that the seasonal and regional variations in the spreading rate of an epidemic need to be taken into account while devising strategies for controlling it.

Persistence of epidemics in networks with modular organization

Spread of infectious diseases (such as influenza, measles, etc) from person to person mainly depends on the structure of human contact networks. Empirical data shows evidence for community or modular organization in such networks. In a modular network, individuals tend to have many more and/or stronger links with members within their own community (or module) compared to members of other modules. Thus, investigating the dynamics of epidemic spreading on modular networks may provide us insights on how to efficiently control human pandemics. In this chapter we have explored the persistence of epidemic dynamics on modular networks where individuals after having recovered from an infection can again becomes susceptible with a certain probability. Our study of SIRS (Susceptible-Infectious-Recovered-Susceptible) dynamics in a modular network suggests that under certain circumstances an epidemic can become persistently recurrent in a population. We show how the probability of persistence of an epidemic depends on the network mesoscopic organization as well as on individual dynamical parameters such as the infection rate. In particular, we show that highly contagious diseases, which quickly die out in a population with homogeneous contact structure, can survive indefinitely (becoming endemic) when there is strong community organization in the population.

Spatiotemporal patterns of incidence for a vector-borne infectious disease

Malaria is a mosquito-borne infectious disease that results in more than a million people dying around the world each year. It is endemic in many tropical and subtropical countries, including India, and is a heavy burden on their public health systems. Identifying and understanding characteristic spatio-temporal patterns of malaria incidence can help in arriving at better control and containment strategies for epidemics. In this chapter we have investigated the spatio-temporal incidence of *Plasmodium falciparum* and *Plasmodium vivax* in a data-set describing infections occurring in villages grouped under 51 health centers in a district sub-division of northern Bengal. We use the method of wavelet phase analysis to identify certain health centers acting as epicenters of the diseases, from which the infections spread as traveling waves. In contrast to earlier studies of diseases spread by direct transmission between infected individuals, we observe that epicenters for a vector-borne infectious disease do not necessarily occur in zones having the highest population density. The identification of a dominant periodicity in the data corresponding to one year points to the important role played by seasonal variation in environmental factors affecting vector population growth. We confirm this by simulating a spatially detailed model of malaria transmission where a periodic environmental signal (viz., rainfall with a periodicity of one year) that affects the adult mosquito emergence rate is seen to result in a periodically varying incidence of infection having the same period.

7.2 Outlook

In this thesis, we have addressed several problems that can contribute to a general understanding of how structure of networks affect diffusion of signals or contagia, in the context of host-pathogen interactions. A natural extension of the work presented here would be to develop a comprehensive theory of how appearance of specific mesoscopic organizational

structures affect communication over the system. For instance, we have shown here that branching in the intra-cellular signaling cascade, that results in pathways diverging from a common upstream kinase, gives rise to unexpected long-range communication in the network through retrograde propagation. A natural extension of this research would be to ask about the role of convergent pathways in the transmission of information in such networks. One could intuitively surmise that this may result in congestion at the node where multiple pathways meet and this may identify vulnerable points in the signaling network that pathogens can exploit.

Another possible extension relates to the investigation of the mesoscopic structural organization in the network of cancer-related genes. One could conceive of adding more factors to those already considered. For instance, it is possible to consider including data related to DNA microarray analysis of samples from cancer patients. Alternatively, a network can be constructed by connecting tumor types or cancer categories with the specific pharmaceutical drugs or clinical treatment used for such diseases. From this bipartite network, gene network projections can be constructed whose mesoscopic organization can then be compared to the one analyzed in this thesis.

Proceeding to the scale of epidemic networks, we can extend the work reported in this thesis by investigating the effect of increasing speed of transportation between population centers on the initial spreading dynamics, characterized by the basic reproduction number, of an epidemic. This issue is important in view of the debate on how appropriate the "well-mixed population" assumption is at different spatial scales. For example, one could argue that representing the spreading dynamics of a large region (e.g., an entire country) by a single basic reproduction number may be misleading. To see how well a single number represents the epidemic dynamics of large regions with multiple urban centers one could set up a spatially detailed model where the spreading rate of the disease at each location, in the absence of any communication between neighboring areas, is different. Once people are allowed to move from one region to another, it is possible that the disease will spread

at a rate that is common to all the regions, provided that the dispersion of values for the local spreading rates is not too high. However, one can still ask whether this common rate will be close to the mean value of the different local spreading rates, or whether it will approach the maximum of all these rates. An equally important question is what is the minimum rate of communication that is required in order for the entire system to attain a common spreading rate. These studies would have practical implications on the validity of calculating basic reproduction numbers for epidemics in different situations.

Bibliography

[1] International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. Nature, 431: 931-945

[2] Kitano H (2002) Computational systems biology. Nature, 420: 206-210.

[3] Kitano H (2002) Systems biology: A brief overview. Science, 295: 1662-1664.

[4] Wiener N (1948) Cybernetics, or Control and Communication in the Animal and the Machine (MIT Press, Cambridge, MA).

[5] Ashby W R (1952) Design for a Brain (Chapman and Hall, London).

[6] McCulloch W S (1965) Embodiments of Mind (MIT Press, Cambridge, MA).

[7] von Bertalanffy L (1968) General System Theory: Foundations, Development, Applications (George Braziller, New York).

[8] Sinha S, Jesan T and Chatterjee N (2009) Systems biology: From the cell to the brain. in Current Trends in Science: Platinum Jubilee Special, edited by Mukunda N (Indian Academy of Sciences, Bangalore) 199-205.

[9] Strogatz S H (2001) Exploring complex networks. Nature, 410: 268-276.

[10] Newman M E J (2010) Networks: An Introduction (Oxford University Press, Oxford).

[11] Wasserman S and Faust K (1994) Social Network Analysis: Methods and Applications (Cambridge University Press, Cambridge).

[12] Albert R and Barabasi A L (2002) Statistical mechanics of complex networks. Rev. Mod. Phys., 74: 47-97.

[13] Newman M E J (2003) The structure and function of complex networks. SIAM Review, 45:167-256.

[14] Watts D J and Strogatz S H (1998) Collective dynamics of "small-world" networks. Nature, 393: 440-442.

[15] Barabasi A L and Albert R (1999) Emergence of scaling in random networks. Science, 286: 509-512.

[16] Barabasi A L and Oltvai Z N (2004) Network biology : Understanding the cell âĂŹs functional organization. Nature Reviews Genetics, 5: 101-113.

[17] Kauffman S A (1969) Metabolic stability and epigenesist in randomly constructed genetic nets. Journal of Theoretical Biology, 22: 437-467.

[18] Samal A and Jain S (2008) The regulatory network of E. coli metabolism as a Boolean dynamical system exhibits both homeostasis and flexibility of response. BMC Systems Biology, 2: 21.

[19] Dassow V G, Meir E, Munro E M and Odell G M (2000) The segment polarity network is a robust developmental module. Nature, 406: 188-192.

[20] Tyson J J, Chen K C and Novak B (2001) Network dynamics and cell physiology. Nature Reviews Mol. Cell Biology, 2: 908-916.

[21] Jeong H, Tombor B, Albert R, Oltvai Z N and Barabasi A L (2000) The large-scale organization of metabolic networks. Nature, 407: 651-654.

[22] Wagner A and Fell D A (2001) The small world inside large metabolic networks. Proc. Roy. Soc. Lond. B, 268: 1803-1810.

[23] Guimera R and Amaral L A N (2005) Functional cartography of complex metabolic networks. Nature, 433: 895-900.

[24] Tyson J J, Chen K C and Novak B (2003) Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell. Curr. Opin. Cell Biology, 15: 221-231.

[25] Sinha S (2009) From network structure to dynamics and back again: Relating dynamical stability and connection topology in biological complex systems. In Dynamics On and Of Complex Networks (Editors: N Ganguly, A Deutsch and A Mukherjee, Boston, Birkhauser), [also at arxiv:0804.0977].

[26] del Sol A, Fujihashi H and O'Meara P (2005) Topology of small-world networks of protein-protein complex structures. Bioinformatics, 21: 1311-1315.

[27] Bagler G and Sinha S (2007) Assortative mixing in protein contact networks and protein folding kinetics. Bioinformatics, 23: 1760-1767.

[28] Hartwell L H, Hopfield J J, Leibler S and Murray A W (1999) From molecular to modular cell biology. Nature 402: C47-52.

[29] Pavlopoulos G A, Secrier M, Moschopoulos C N, Soldatos T G, Kossida S, Aerts J, Schneider R and Bagos P G (2011) Using graph theory to analyze biological networks, BioData Mining, 4: 10.

[30] Jeong H, Mason S P, Barabasi A L and Oltvai Z N (2001) Lethality and centrality in protein networks. Nature, 411: 41.

[31] Jeong H, Oltvai Z N and Barabasi A L (2003) Prediction of protein essentiality based on genomic data. ComPlexUs, 1: 19-28.

[32] Goh K I, Cusick M E, Childs D V B, Vidal M, and Barabasi A L (2007) The human disease network. Proc. Natl. Acad. Sci. USA, 104: 8685-8690.

[33] Kann M G (2007) Protein interactions and disease: computational approaches to uncover the etiology of diseases. Brief Bioinform, 8: 333-346.

[34] Harris M A, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C *et al* (2004) The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res., 32: D258-D261.

[35] Futreal A P, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N and Stratton M R (2006) A census of human cancer genes. Nature Reviews Cancer, 4: 177-183.

[36] Jonsson P F and Bates P A, (2006) Global topological features of cancer proteins in the human interactome. Bioinformatics, 22: 2291-2297.

[37] Ideker T, and Sharan R (2008) Protein networks in disease. Genome Res., 18: 644-652.

[38] Pellegrini M, Haynor D and Johnson J M (2004) Protein interaction networks. Expert Rev. Proteomics, 1: 239-249.

[39] Bader S, Kuhner S and Gavin A C (2008) Interaction networks for systems biology. FEBS Letters, 582: 1220-1224.

[40] Shen-Orr S, Milo R, Mangan S and Alon U (2002) Network motifs in the transcriptional regulation network of Escherichia coli. Nat Genet, 31: 64-68.

[41] Karlebach G and Shamir R (2008) Modelling and analysis of gene regulatory networks. Nature Reviews Molecular Cell Biology, 9: 770-780.

[42] Folger O, Jerby L, Frezza C, Gottlieb E, Ruppin E and Shlomi T (2011) Predicting selective drug targets in cancer through metabolic networks. Molecular Systems Biology, 7: 501.

[43] McCloskey D, Palsson B O and Feist A M (2013) Basic and applied uses of genome scale metabolic network reconstructions of Escherichia coli. Molecular Systems Biology, 9: 661

[44] Ma H and Zeng A P (2003) Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. Bioinformatics, 19: 270-277.

[45] Ravasz E, Somera A L, Mongru D A, Oltvai Z N and Barabasi A L (2002) Hierarchical organization of modularity in metabolic networks. Science, 297: 1551-1555.

[46] Bray D (2003) Molecular networks: The top - down view. Science, 301: 1864 -1865.

[47] Junker B H and Schreiber F, editors (2008) Analysis of Biological Networks (John Wiley, Hoboken, NJ).

[48] Weng G, Bhalla US and Iyengar R (1999) Complexity in biological signaling systems. Science, 284: 92-96.

[49] Kitano K H (2004) Biological robustness. Nature Reviews in Genetics, 5: 826-837.

[50] Bhalla U S and Iyengar R (1999) Emergent properties of networks of biological signaling pathways. Science, 283: 381-387.

[51] Bagowski C P and Ferrell J E (2001) Bistability in the JNK cascade. Curr. Biol., 11: 1176-1182.

[52] Bhalla U S, Ram P T and Iyengar R (2002) MAP kinase phosphatase as a locus of flexibility in a mitogen-activated protein kinase signaling network. Science, 297: 1018-1023.

[53] Ferrell J E (2002) Self-perpetuating states in signal transduction: positive feedback, double-negative feedback and bistability. Curr. Opin. Cell Biol., 14: 140-148.

[54] White J G, Southgate E, Thomson J N and Brenner S (1986) The structure of the nervous system of the nematode C. elegans. Phil. Trans. Roy. Soc. Lond. B, 314: 1-340.

[55] Chatterjee N and Sinha S (2007) Understanding the mind of a worm: Hierarchical network structure underlying nervous system function in C. elegans. Prog. Brain Res., 168: 145-153.

[56] Hethcote H W (2000) Mathematics of infectious diseases. SIAM Review, 42: 599-653.

[57] Anderson R M and May R M (1991) Infectious Diseases of Humans: Dynamics and Control (Oxford University Press, Oxford).

[58] Moore C and Newman M E J (2000) Epidemics and percolation in small-world networks. Phys. Rev. E, 61: 5678-5682.

[59] Kuperman M and Abramson G (2001) Small-world effect in an epidemiological model. Phys. Rev. Lett, 86: 2909-2912.

[60] Saramaki J and Kaski K (2005) Modelling development of epidemics with dynamic small-world networks. J. Theor. Biol., 234: 413-421.

[61] Deem M W (2007) Mathematical adventures in biology. Phys. Today, 60 (1): 42-47.

[62] Newman M E J (2002) Spread of epidemic disease on networks. Phys. Rev. E, 66: 016128.

[63] Pan R K and Sinha S (2009) Modularity produces small-world networks with dynamical time-scale separation. Europhys. Lett., 85: 68006.

[64] Sinha S (2005) Complexity vs. stability in small-world networks. Physica A, 346: 147-153.

[65] Dunne J A, Williams R J and Martinez N D (2002) Food-web structure and network theory: The role of connectance and size. Proc. Natl. Acad. Sci. USA, 99: 12917-12922.

[66] Krause A E, Frank K A, Mason D M, Ulanowicz R U and Taylor W W (2003) Compartments revealed in food-web structure. Nature, 426: 282-284.

[67] Amaral L A N, Scala A, Barthelemy M and Stanley H E (2000) Classes of small-world networks. Proc. Natl. Acad. Sci. USA, 97: 11149-11152.

[68] Latora L and Marchiori M (2001) Efficient behavior of small-world networks. Phys. Rev. Lett., 87: 198701.

[69] Freeman L C (1977) A set of measures of centrality based on betweenness. Sociometry, 40: 35-41.

[70] Newman M E J (2001) Scientific collaboration networks. I. Network construction and fundamental results. Phys. Rev. E, 64: 016131.

[71] Girvan M and Newman M E J (2002) Community structure in social and biological networks, Proc. Natl. Acad. Sci. USA, 99: 7821-7826.

[72] Boccaletti S, Latora V, Moreno Y, Chavez M and Hwang D U (2006) Complex networks: Structure and dynamics. Physics Reports, 424: 175-308.

[73] Arenas A, Diaz-Guilera A, Kurths J, Moreno Y and Zhou C (2008) Synchronization in complex networks. Physics Reports, 469: 93-153.

[74] Dorogovtsev S N, Goltsev A V and Mendes J F F (2008) Critical phenomena in complex networks. Rev. Mod. Phys. 80: 1275.

[75] Sinha S and Poria S (2011) Multiple dynamical time-scales in networks with hierarchically nested modular organization. Pramana, 77: 833-842.

[76] Pan R K and Sinha S (2008) Modular networks with hierarchical organization: The dynamical implications of complex structure. Pramana, 71: 331-340.

[77] Granell C, Gomez S and Arenas A (2011) Mesoscopic analysis of networks: Applications to exploratory analysis and data clustering. Chaos, 21: 016102.

[78] Newman M E J (2006) Modularity and community structure in networks. Proc. Natl. Acad. Sci. USA, 103: 8577-8582.

[79] Newman M E J (2012) Communities, modules and large-scale structure in networks. Nature Physics, 8: 25-31.

[80] Newman M E J and Girvan M (2004) Finding and evaluating community structure in networks. Phys. Rev. E, 69: 026113.

[81] Lancichinetti A and Fortunato S (2011) Limits of modularity maximization in community detection. Phys. Rev. E 84: 066122.

[82] Hintze A and Adami C (2008) Evolution of complex modular biological networks. PLoS Comput Biol, 4(2): e23. doi:10.1371/journal.pcbi.0040023.

[83] Seger R and Krebs E G (1995) The MAPK signaling cascade. FASEB Journal 9: 726-735.

[84] Orton R J, Sturm O E, Vyshemirsky V, Calder M, Gilbert D R and Kolch W (2005) Computational modelling of the receptor-tyrosine-kinase-activated MAPK pathway. Biochemical Journal, 392: 249-261.

[85] Wagner E F and Nebreda A R (2009) Signal integration by JNK and p38 MAPK pathways in cancer development. Nat Rev Cancer, 9: 537-549.

[86] Kim E K, Choi E J (2010) Pathological roles of MAPK signaling pathways in human diseases. Biochim Biophys Acta, 1802: 396-405.

[87] Fritsche-Guenther R, Witzel F, Sieber A, Herr R, Schmidt N, Braun S, Brummer T, Sers C and Blüthgen N (2011) Strong negative feedback from Erk to Raf confers robustness to MAPK signalling. Mol Sys Bio., 7: 489.

[88] Fortunato S (2010) Community detection in graphs. Physics Reports, 486: 75-174

[89] Newman M E J (2004) Fast algorithm for detecting community structure in networks. Phys. Rev. E, 69: 066133.

[90] Clauset A, Newman M E J and Moore C (2004) Finding community structure in very large networks. Physical Review E, 70: 066111.

[91] Lancichinetti A and Fortunato S (2009) Community detection algorithms: a comparative analysis. Phys. Rev. E, 80: 056117.

[92] Blondel V D, Guillaume J L, Lambiotte R and Lefebvre E (2008) Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment: P10008.

[93] Palla G, Derenyi I, Farkas I and Vicsek T (2005) Uncovering the overlapping community structure of complex networks in nature and society. Nature, 435: 814-818.

[94] Rosvall M and Bergstrom C T (2007) An information-theoretic framework for resolving community structure in complex networks. Proc. Natl. Acad. Sci. USA, 104: 7327-7331.

[95] Rosvall M and Bergstrom C T (2008) Maps of random walks on complex networks reveal community structure. Proc. Natl. Acad. Sci. USA, 105: 1118-1123.

[96] Chen J, Lu J A, Zhan C J and Chen G (2011) Laplacian Spectra and Synchronization Processes on Complex Networks. In Handbook of Optimization in Complex Networks : Theory and Applications, edited by Thai M T and Pardalos P M (Springer, Berlin).

[97] Meunier D, Lambiotte R and Bullmore E T (2010) Modular and hierarchically modular organization of brain networks. Frontiers in Neuroscience, 4: 200.

[98] Widmann C, Gibson S, Jarpe M B and Johnson G L (1999) Mitogen-activated protein kinase: Conservation of a three-kinase module from yeast to human. Physiol Rev, 79: 143-180.

[99] Kolch W (2000) Meaningful relationships: the regulation of the Ras/Raf/MEK/ERK pathway by protein interactions. Biochemical Journal, 351: 289-305.

[100] Lewin B, Cassimeris L, Lingappa V R and Plopper G, editors (2007) Cells (Jones and Bartlett, Sudbury, MA).

[101] Fanger GR, Widmann C, Porter A C, Sather S, Johnson G L and Vaillancourt R R (1998) 14-3-3 proteins interact with specific MEK kinase. J Biol Chem, 273: 3476-3483.

[102] Huse M (2009) The T-cell receptor signaling network. J Cell Science, 122: 1269-1273.

[103] Dal Porto J M, Gauld S B, Merrell K T, Mills D, Pugh-Bernard A E and Cambier J (2004) B cell antigen receptor signaling 101. Mol Immunol., 41: 599-613.

[104] Downward J (2001) The ins and outs of signalling. Nature, 411: 759-762.

[105] Komarova N L, Zou X, Nie Q, Bardwell L (2005) A theoretical framework for specificity in cell signaling. Mol. Sys. Biol., 1:0023.

[106] Qiao L, Nachbar R B, Kevrekidis I G and Shvartsman S Y (2007) Bistability and oscillations in the Huang-Ferrell model of MAPK signaling. PLoS Comput Biol., 3: e184.

[107] Ventura A C, Sepulchre J A, Merajver S D (2008) A hidden feedback in signaling cascades is revealed. PLoS Comput. Biol., 4: e1000041.

[108] Jesan T, Sarma U, Halder S, Saha B and Sinha S (2013) Branched motifs enable long-range interactions in signaling networks through retrograde propagation. PLoS ONE, 8: e64409.

[109] Huang CY and Ferrell JE (1996) Ultrasensitivity in the mitogen-activated protein kinase cascade. Proc Natl Acad Sci., USA, 93: 10078-10083.

[110] Millat T, Bullinger E, Rohwer J and Wolkenhauer O (2007) Approximations and their consequences for dynamic modelling of signal transduction pathways. Math Biosci., 207: 40-57.

[111] Alon U, Surette M G, Barkai N and Leibler S (1999) Robustness in bacterial chemotaxis. Nature 397: 168-171.

[112] Lange-Carter C A, Pleiman CM, Gardner AM, Blumer K J and Johnson G L (1993) A divergence in the MAP kinase regulatory network defined by MEK kinase and Raf. Science, 260: 315-319.

[113] Yan M, Dai T, Deak J C, Kyriakis J M, Zon L I, Woodgett J R and Templeton D J (1995) Activation of stress-activated protein kinase by MEKK1 phosphorylation of its activator SEK1. Nature, 372: 798-800.

[114] Lu X, Nemoto S and Lin A (1997) Identification of c-Jun NH2-terminal protein kinase (JNK)-activating kinase 2 as an activator of JNK but not p38. J Biol Chem., 272: 24751-24754.

[115] Deacon K and Blank J L (1997) Characterization of the mitogen-activated protein kinase kinase 4 (MKK4)/c-Jun NH2-terminal kinase 1 and MKK3/p38 pathways regulated by MEK kinases 2 and 3. MEK kinase 3 activates MKK3 but does not cause activation of p38 kinase in vivo. J Biol Chem., 30: 14489-14496.

[116] Deacon K, Blank J L. (1999) MEK kinase 3 directly activates MKK6 and MKK7, specific activators of the p38 and c-Jun NH2-terminal kinases. J Biol Chem., 274: 16604-16610.

[117] Abell A N, Granger D A and Johnson G L (2007) MEKK4 stimulation of p38 and JNK activity is negatively regulated by GSK3β. J Biol Chem., 282: 30476-30484.

[118] Ichijo H, Nishida E, Irie K, ten Dijke P, Saitoh M, Moriguchi T, Takagi M, Matsumoto K, Miyazono K and Gotoh Y (1997) Induction of apoptosis by ASK1, a mammalian MAPKKK that activates SAPK/JNK and p38 signalling pathways. Science, 275: 90-94.

[119] Zama T, Aoki R, Kamimoto T, Inoue K, Ikeda Y and Hagiwara M (2002) Scaffold role of a mitogen activated protein kinase phosphatase, SKRP1, for the JNK signaling pathway. J Biol Chem., 277: 23919-23926.

[120] Sutherland C L, Heath A W, Pelech S L, Young P R and Gold M R (1996) Differential activation of the ERK, JNK, and p38 mitogen-activated protein kinases by CD40 and the B cell antigen receptor. J Immunol., 157: 3381-3390.

[121] Blüthgen N, Bruggeman F J, Legewie S, Herzel H, Westerhoff H V and Kholodenko, B N (2006) Effects of sequestration on signal transduction cascades. FEBS Journal, 273: 895-906.

[122] Markevich N I, Hoek J B and Kholodenko B N (2004) Signaling switches and bistability arising from multisite phosphorylation in protein kinase cascades. J Cell Biol., 164: 353-359.

[123] Cloutier M and Wang E (2011) Dynamic modeling and analysis of cancer cellular network motifs. Integr Biol 3: 724-732.

[124] Soyer O S, Salathe M and Bonhoeffer M (2006) Signal transduction networks: Topology, response and biochemical processes. J Theor Biol., 238: 416-425.

[125] Shin S Y, Rath O, Choo S M, Fee F, McFerran B, Kolch W and Cho KH (2009) Positive- and negative-feedback regulations coordinate the dynamic behavior of the Ras-Raf-MEK-ERK signal transduction pathway. J Cell Sci., 122: 425-435.

[126] Olson J M, Hallahan A R (2004) p38 MAP kinase: a convergence point in cancer therapy. Trends Mol Med., 10: 125-129.

[127] Lawrence M C, Jivan A, Shao C, Duan L, Goad D *et al* (2008) The roles of MAPKs in disease. Cell Res., 18: 436-442.

[128] Berger S I and Iyengar R (2009) Network analyses in systems pharmacology. Bioinformatics, 25: 2466-2472.

[129] Hagemann D, Troppmair J and Rapp U R (1999) Cot protooncoprotein activates the dual specificity kinases MEK-1 and SEK-1 and induces differentiation of PC12 cells. Oncogene, 18: 1391-1400.

[130] Gantke T, Sriskantharajah S, Sadowski M and Ley S C (2012) IkB kinase regulation of the TPL-2/ERK MAPK pathway. Immunol Rev., 246: 168-182.

[131] Hui L, Bakiri L, Mairhorfer A, Schweifer N, Haslinger C, Kenner L, Komnenovic V, Scheuch H, Beug H and Wagner E F (2007) p38alpha suppresses normal and cancer cell proliferation by antagonizing the JNK-c-Jun pathway. Nature Genetics, 39: 741-749.

[132] Bennett B L, Satoh Y and Lewis A J (2003) JNK: a new therapeutic target for diabetes. Curr Opin Pharmacol., 3: 420-425.

[133] de Alvaro C, Teruel T, Hernandez R and Lorenzo M (2004) Tumor necrosis factor alpha produces insulin resistance in skeletal muscle by activation of inhibitor kappaB kinase in a p38 MAPK-dependent manner. J Biol Chem., 279: 17070-17078.

[134] Chowdhury T A (2010) Diabetes and cancer. QJM, 103: 905-915.

[135] Rung J, Cauchi S, Albrechtsen A, Shen L and Rocheleau G *et al* (2009) Genetic variant near IRS1 is associated with type 2 diabetes, insulin resistance and hyperinsulinemia. Nature Genetics, 41: 1110-1115.

[136] Knox C, Law V, Jewison T, Liu P, Ly S *et al* (2011) DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. Nucleic Acids Res., 39: D1035-1041.

[137] http://www.who.int/mediacentre/factsheets/fs297/en/

[138] For a recent popular account of the medical campaign to cure cancer see Mukherjee S (2010) The Emperor Of All Maladies: A Biography of Cancer (Scribner, New York).

[139] Hornberg J J, Bruggeman F J, Westerhoff H V and Lankelma J (2006) Cancer: A Systems Biology disease. Biosystems, 83: 81-90.

[140] Gong X, Wu R, Zhang Y, Zhao W, Cheng L, Gu Y, Zhang L, Wang J, Zhu J and Guo Z (2010) Extracting consistent knowledge from highly inconsistent cancer gene data sources. BMC Bioinformatics, 11: 76.

[141] Hamosh A, Scott A F, Amberger J S, Bocchini C A and McKusick V A (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Research, 33: D514-517.

[142] http://www.hprd.org/

[143] Prasad T S K *et al* (2009) Human Protein Reference Database - 2009 Update. Nucleic Acids Research, 37: D767-72.

[144] Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D and Alon U (2002) Network motifs: Simple building blocks of complex networks. Science, 298: 824-827.

[145] Bhattacharya K, Mukherjee G, Saramaki J, Kaski K and Manna S S (2008) The International Trade Network: weighted network analysis and modelling. J. Stat. Mech.: P02002. doi:10.1088/1742-5468/2008/02/P02002.

[146] Guimera R, Pardo M S and Amaral L A N (2007) Classes of complex networks defined by role-to-role connectivity profiles. Nature Physics, 3: 63-69.

[147] http://www.tp.umu.se/~rosvall/code.html

[148] Schaefer C F, Anthony K, Krupa S, Buchoff J, Day M, Hannay T nd Buetow K H (2009) PID: The Pathway Interaction Database. Nucleic Acids Res., 37: D674-9.

[149] http://pid.nci.nih.gov/

[150] http://seer.cancer.gov/

[151] Kaplan M M and Webster R G (1977) The epidemiology of influenza. Scientific American, 237(6): 88-106.

[152] Potter C W (2001) A history of influenza. J. Applied Microbiology, 91: 572-579.

[153] Johnson N P A S and Mueller J (2002) Updating the accounts: Global mortality of the 1918-1920 "Spanish" influenza pandemic. Bull. Hist. Med., 76: 105-115.

[154] 2009 H1N1 Flu: International Situation Update, 22 January 2010. Available from:

http://www.cdc.gov/h1n1flu/updates/international

[155] Jameel, S (2010) The 2009 influenza pandemic. Current Science, 98: 306-311.

[156] World Health Organization, Pandemic (H1N1) 2009 - update 96, 16 April 2010. Available from:

http://www.who.int/csr/don/2010_04_16/en/index.html

[157] Ministry of Health and Family Welfare, Government of India, Situation Update on H1N1, 11 April 2010. Available from:

http://mohfw-h1n1.nic.in/documents/PDF/SituationalUpdatesArchives/april2010/Situational%20Updates%20on%2011.04.2010.pdf

[158] Boëlle P Y, Bernillon P and Desenclos, J C (2009) A preliminary estimation of the reproduction ratio for new influenza A(H1N1) from the outbreak in Mexico, March-April 2009. Euro Surveill., 14: 19. pii: 19205.

[159] Cruz-Pacheco G, Duran L, Esteva L, Minzoni, A. A., Lopez-Cervantes M, Panayotaros P, Ortega A, Ruiz V I (2009) Modelling of the Influenza A(H1N1)v outbreak in Mexico City, April-May 2009, with control sanitary measures. Euro Surveill., 14: 26. pii:19254.

[160] Fraser C, Donnelly C A, Cauchemez S, Hanage W P, Van Kerkhove M D, Hollingsworth T D, Griffin J, Baggaley R F, Jenkins H E, Lyons E J, Jombart T, Hinsley W R, Grassly N C, Balloux F, Ghani A C, Ferguson N M, Rambaut A, Pybus OG, Lopez-Gatell H, Alpuche-Aranda, C M, Chapela I B, Zavala E'P, Guevara D M, Checchi F, Garcia E, Hugonnet S and Roth C (2009) WHO Rapid Pandemic Assessment Collaboration, Pandemic potential of a strain of Influenza A(H1N1): Early findings. Science, 324: 1557-1561.

[161] Yang Y, Sugimoto J D, Halloran, M E, Basta N E, Chao D L, Matrajt L, Potter G, Kenah E and Longini I M, (2009) The transmissibility and control of pandemic influenza A(H1N1) virus. Science, 326: 729-733.

[162] McBryde E, Bergeri I, van Gemert C, Rotty J, Headley E, Simpson K, Lester R, Hellard M and Fielding J, (2009) Early transmission characteristics of influenza A(H1N1)v in Australia: Victorian state, 16 May - 3 June 2009. Euro Surveill., 14: 42. pii: 19363.

[163] http://mohfw-h1n1.nic.in/

[164] http://www.world-gazetteer.com/

(Retrieved on February 18, 2010)

[165] http://www.mathworks.com/

[166] Morens D M, Folkers G K and Fauci A S (2004) The challenge of emerging and re-emerging infectious diseases. Nature, 430: 242 -249

[167] Wallinga J, Edmunds W J and Kretzschmar M (1999) Perspective: human contact patterns and the spread of airborne infectious diseases. Trends in Microbiology, 7: 372 .

[168] Liu Z and Hu B (2005) Epidemic spreading in community networks. EPL, 72: 315.

[169] Huang L, Park K and Lai Y C (2006) Information propagation on modular networks. Phys. Rev. E, 73: 035103.

[170] Huang W and Li C, (2007) Epidemic spreading in scale-free networks with community structure. J. Stat. Mech.: P01014.

[171] Zhao H and Gao Z Y (2007) Modular effects on epidemic dynamics in small-world networks. EPL, 79: 38002.

[172] Masuda N (2009) Immunization of networks with community structure. New J. Phys., 11: 123018.

[173] Salathé M and Jones J H (2010) Dynamics and control of diseases in networks with community structure. PLoS Comput Biol., 6: e1000736.

[174] Griffin R H and Nunn C L (2012) Community structure and the spread of infectious disease in primate social networks. Evol. Ecol., 26: 779.

[175] Earn D J D, Rohani P and Grenfell B T (1998) Persistence, chaos and synchrony in ecology and epidemiology. Proc. R. Soc. B, 265: 7

[176] Berglund EC, Nystedt B, Andersson SGE (2009) Computational resources in infectious disease: Limitations and challenges. PLoS Comput Biol 5(10): e1000481. doi:10.1371/journal.pcbi.1000481

[177] Murray CJ, Rosenfeld LC, Lim SS, Andrews KG, Foreman KJ, Haring D, Fullman N, Naghavi M, Lozano R and Lopez AD (2012) Global malaria mortality between 1980 and 2010: a systematic analysis. Lancet, 379: 413-431

[178] Sachs J and Malaney P (2002) The economic and social burden of malaria. Nature, 415: 680- 685.

[179] World Health Organization (WHO) Malaria fact sheet.

http://www.who.int/mediacentre/ factsheets/fs094/en/print.htm

[180] Torrence C and Compo G P (1998) A practical guide to wavelet analysis. Bull. Am. Meteorol. Soc, 79: 61-78.

[181] Wavelet software was provided by C Torrence and G Compo, and is available at:

http://paos.colorado.edu/research/wavelets/

[182] Grenfell B T, Bjornstad O N and Kappey J (2001) Travelling waves and spatial hierarchies in measles epidemics. Nature, 414: 716-723.

[183] Johnson D M, Bjornstad O N and Liebhold A M (2004) Landscape geometry and travelling waves in the larch budmoth. Ecology Letters, 7: 967-974.

[184] Smith D L, Dushoff J and McKenzie F E (2004) The risk of a mosquito-borne infection in a heterogeneous environment. PLoS Biology, 2: 1957-1964.

[185] Cazelles B, Chavez M, Constantin de Magny G, Guegan J F and Hales S (2007) Time-dependent spectral analysis of epidemiological time-series with wavelets. J. R. Soc. Interface, 4: 625-636. doi:10.1098/rsif.2007.0212.

[186] Parham P E and Michael E (2010) Modeling the effects of weather and climate change on malaria transmission. Environ. Health Perspect., 118: 620-626.