Studies on the Structure and Dynamics of Protein Interaction Networks

Thesis Submitted to
University of Hyderabad
for the Degree of Doctor of Philosophy

Shubhada R. Hegde
Centre for DNA Fingerprinting and Diagnostics
Hyderabad

Registration Number: 07LBPH08

2011

Studies on the Structure and Dynamics of Protein Interaction Networks

Thesis Submitted in Partial Fulfillment for the Degree of
Doctor of Philosophy
to

The Department of Biochemistry
School of Life Sciences, University of Hyderabad

by

Shubhada R. Hegde

Centre for DNA Fingerprinting and Diagnostics
Nampally, Hyderabad - 500 001

2011

Registration Number: 07LBPH08

University of Hyderabad
School of Life Sciences
Department of Biochemistry
Hyderabad - 500 046. India

Declaration

The research work embodied in this thesis entitled "Studies on the Structure and Dynamics of Protein Interaction Networks" has been carried out by me at the Centre for DNA Fingerprinting and Diagnostics, Hyderabad, under the guidance of Dr. Shekhar C. Mande. I hereby declare that this work is original and has not been submitted in part or full for any other degree or diploma of any other university.

Shubhada R. Hegde

University of Hyderabad
School of Life Sciences
Department of Biochemistry
Hyderabad - 500 046. India

Certificate

This is to certify that this thesis entitled "Studies on the Structure and Dynamics of Protein Interaction Networks", submitted by Ms. Shubhada R. Hegde for the degree of Doctor of Philosophy to the University of Hyderabad is based on the work carried out by her at the Centre for DNA Fingerprinting and Diagnostics, Hyderabad. This work is original and has not been submitted for any diploma or degree of any other university or institution.

Dr Shekhar C. Mande
Thesis supervisor
CDFD, Hyderabad

Head, Dept. of Biochemistry
University of Hyderabad

Dean, School of Life Science
University of Hyderabad

Table of Contents

Acknowledgements

It is indeed a great pleasure to thank all the people who have guided, motivated and worked towards materializing this thesis. This work was carried out in the Structural Biology laboratory, CDFD, while I was registered as an external PhD candidate with the Department of Biochemistry, University of Hyderabad. I am thankful to the Director and the other scientific and non-scientific staff for providing wonderful environment for research in CDFD. On a similar note, Prof. Ramaiah and other members in the Biochemistry Department, University of Hyderabad were very kind in guiding me through seemingly complicated official procedures. Also, I heartily acknowledge the financial support by University Grants Commission during this period.

I have been fortunate to receive the guidance for my thesis from Dr. Shekhar C. Mande, with whom my association dates back to 2004. The scientific atmosphere in his lab during my summer training days then had motivated me to come back as a graduate student. Besides being an excellent mentor for the projects, he made sure I read sufficiently, developed scientific communication skills and attended right conferences. In his company, learning teamwork, socializing with people and having progressive attitude become natural. Sir, I will always be drawing inspiration from you and I hope I will live up to your training in future. I also would like to acknowledge my doctoral committee members, Dr. H A Nagarajaram and Dr. Abhijit Sardesai, who assessed my progress regularly and made sure I am on the right track.

Special thanking note must go to all my lab members. I have learnt basic programming skills and graph theory from Dr. P Manimaran. Dhananjay was always there to troubleshoot my silly OS associated problems. Sailu and Kshama were very kind in replying my every e-mail and helping me understand the project initially. Heartfelt thanks to Anamika for a wonderful company in Gandipet lab. Also, my seniors in the lab- Akif, Santosh, Madhav, Debashree and Pramod were of immense strength. I also thank past and present members of the lab- Aditi, Ashwani, Bala, Jayshree, Mamta, Navita, Neeraja, Payel, Roshna Susan and Swastik for their support and encouragement. I am immensely benefited from the formal lab meetings and informal scientific discussions with all of them. On a parallel note, I thank Sheeba, Madhuri and Hassan for providing excellent support for all the non-scientific affairs. It was great to work with project students in the lab, namely Rajesh, Chandrani, Khushbu, Neha, Jhinak and Sushma; their enthusiasm was really contagious.

SUN-CoE, CDFD had provided high-end computational facility for the projects and I must thank all the members of SUN-CoE and Bioinformatics for their timely assistance. Kavita was an excellent systems administrator who helped us with our every minute problem in Gandipet lab. Besides, she has been an amazing friend and a person to look up to. I will always miss our innumerable discussions on books, movies, music and life in general. Days spent with LCB members in Gandipet were truly memorable. I would like to thank with great pleasure Vijay, Jamshaid and Anupam for their help and guidance. I am also grateful to all the members of LBG for involving me in the discussions on a number of occasions.

I thank our collaborators Dr. Kanury Rao (ICGEB, New Delhi), Dr. Vsevolod Makeev (Russian Academy of Science, Moscow) and Dr. Sharmila Mande (TCS, Hyderabad), with whom I had several stimulating interactions. Outside CDFD, I had wonderful time working with Noor, Srikanth, Elya, Julia and Hannah. Noor has been a great friend and motivator all these years, and was very kind to host me at her house in Dehli during my visit.

I acknowledge CDFD for providing me the financial support to attend many national and international conferences. Mrs. Rani Simran Kaur and Bibusita were wonderful hosts when I visited New York for CSHL conference in 2011. I thank Department of Science and Technology, Govt. of India for funding my trip to Lindau Nobel Laureates Meeting-2011. DFG was very generous in supporting my subsequent visits to several other institutes in Germany.

From my initial days in CDFD, Ghazala and Nora have shared so many special moments. They stood by me during difficult phases and helped me grow as a person. Love you both so much☺. I have had some of the best times with my other batch mates Neelima, Sreejata and Rohan as well. Special thanks to Neels for accommodating me at her house innumerable times. It was great to be with Charita, Tej and Ranjani anytime. CDFD hostel has been my home for the last five years, and I am grateful to every hosteller for making it a special place to stay.

Thanks to all the social network sites and my innumerable friends on-line who kept me sane and hearty while I was stuck in the lab most of the time. I have a deep sense of gratitude to my Hindustani music Guru Shri Suresh Karhade, who took me along a world as interesting and complex as my research domain.

I am blessed with a wonderful family that supports unconditionally whatever I choose to do. My parents, brother and my entire extended family were my strength all throughout. I am quite indebted to Sadakaka, Jayakka and Shraddha, who made 'Hudukata' my other home. I can't thank Sumakka enough for being there for me always. My friends from the native place are my treasure since childhood days, and I take this opportunity to thank them all for their goodwill. It is because of the love, patience and belief of all of these people that I have this thesis in my hand.

Shubhada Hegde

Preface

Systems biology aims to understand complex biological phenomena in a collective manner. However, underlying complexity of biological systems makes it inevitable to find novel methods for their representation and analysis. Interaction of proteins in a cell can be illustrated in the form of a network with nodes as proteins and the edge between them indicating a physical or a functional association. Graph theoretical analysis of such a representation is an informative approach for comprehending biological functions, conditional responses and system perturbations. Using gene expression datasets, this thesis aims to delineate conditional responses of an organism through otherwise static representation of protein functional linkages. The thesis focuses on two prokaryotic organisms namely, *E. coli* and *M. tuberculosis,* and attempts to understand interactions of proteins and their biological consequences in a condition dependent manner. *E. coli* being widely studied prokaryotic organism is an attractive model for computational systems studies. On the other hand *M. tuberculosis*, a pathogen causing the disease tuberculosis, remains an enigma in terms of both its pathogenicity and latency. A promising approach in computational systems studies is to establish methodologies and principles in a well studied organism and translate such approaches to understand obscure functions in a less studied organism.

The thesis presented is themed as 'computational systems biology', and is divided into five chapters. **Chapter 1** introduces to complex systems and biological complexity, and details the need for systems-level analyses in biology. This chapter also reviews existing literature for biomolecular interactions, computational methods for the prediction of protein:protein interactions and their potential applications. The chapter culminates with future directions for analysing protein interactions in an effective manner.

One of the complications of analysing protein interactions is underestimating the inherent dynamics of the interactions which results from conditionally regulated gene expression. In **Chapter 2**, the work aimed at analysing conditional networks that are constructed by integrating static protein interaction maps and gene expression data is presented. Response of cells to changing environmental conditions is governed by the dynamics of intricate biomolecular interactions. It may be

reasonable to assume that proteins being the dominant macromolecules that carry out routine cellular functions, understanding the dynamics of protein:protein interactions might yield useful insights into the cellular responses. The large-scale protein interaction data sets are, however, unable to capture the changes in the profile of protein:protein interactions. In order to understand how these interactions change dynamically, conditional protein linkages for *E. coli* have been constructed by integrating functional linkages and gene expression information. As a case study, UV exposure in wild-type and SOS deficient *E. coli* at 20 minutes post irradiation has been analyzed. The conditional networks exhibit similar topological properties. Although the global topological properties of the networks are similar, many subtle local changes are observed, which are suggestive of the cellular responses to perturbations. Some such changes correspond to differences in the path lengths among the nodes of carbohydrate metabolism correlating with its loss in efficiency in the UV treated cells. Similarly, expression of hubs under unique conditions reflects the importance of these genes. Various centrality measures applied on the networks indicate increased importance for replication, repair and other stress proteins for the cells under UV treatment, as anticipated. The work therefore proposes a novel approach to study an organism at the systems level by integrating genome-wide functional linkages and the gene expression data.

In order to extrapolate protein interaction network dynamics across number of growth conditions, it is important to characterize the nature of coregulated gene expression, which is the basis for conditional cellular responses. In this direction, analysis of the large-scale gene expression datasets in *E. coli* is reported as **Chapter 3**. One of the challenges of systems biology is to comprehend biological processes by consolidating accumulated information. In this regard, gene expression data of *E. coli* across multiple growth conditions was analyzed to understand gene-gene relationships exhibited at expression levels. A large compendium of gene expression dataset, which covers 466 growth conditions, was used for the analysis. The quality of the data was tested for operonic gene pairs which are expected to exhibit correlation in their expression. Using gene expression data, the genes of *E. coli* were profiled into three classes: Widely expressed, Conditionally expressed and Less expressed. Using gene expression intensities across growth conditions, it was observed that widely expressed genes also show higher level of expression compared to conditionally expressed and less expressed genes. Interestingly, genes that are correlated in expression are

proximal in an interaction network and exhibit similarity in their mRNA half lives. In addition, two tightly regulated gene clusters in *E. coli* were identified using anticorrelated gene pairs. The implication of such an effort towards understanding the rules that govern coregulation of genes in enhancing our understanding of network dynamics appears promising.

Other significant focus of the thesis is to translate the knowledge of protein interactions and their dynamics to comparatively less understood pathogenic organism *M. tuberculosis.* **Chapter 4** of the thesis deals with predicting genome-wide protein functional linkages for *M. tuberculosis*. A prediction for genome-wide protein functional linkages was obtained based on gene expression correlations along with genome-context methods namely, phylogenetic profile, gene distance, operonic frequency, that were combined using a Support Vector Machine. With 88% prediction accuracy, 32546 protein functional linkages were predicted for 3571 proteins of *M. tuberculosis*. The network showed scale-free topology and small world property with about 95% of the interactions in the core cluster. Analysis of the centrality measures indicated a strong correlation between node centrality and gene essentiality in *M. tuberculosis*. Predicted network along with search options for individual protein interactions are made available on the URL http://www.cdfd.org.in/MtbPPI/.

About 90% of the people infected with *M. tuberculosis* carry latent bacteria which are believed to get activated upon immune suppression. One of the fundamental challenges in the control of tuberculosis is therefore to understand molecular mechanisms involved in the onset of latency and/or reactivation. The attempt to address this problem at the systems level by integration of functional interactions with large scale gene expression studies and predicted transcription regulatory network, and finally simulations with a Boolean model of the network is detailed in **Chapter 5** of the thesis. The set of predicted protein functional linkages along with gene expression data of the available models of latency was employed to identify proteins involved in mediating switch signals during dormancy. Notably, genes that are up and down regulated during dormancy are not only coordinately regulated under dormancy-like conditions but also under a variety of other experimental conditions. Their synchronized regulation indicates that they form a tightly regulated gene cluster and might form a latency-regulon. Conservation of these genes across bacterial species suggests a unique evolutionary history that might be associated with *M.*

tuberculosis dormancy. Finally, simulations with a Boolean model based on the regulatory network with logical relationships derived from gene expression data reveals key regulators switch to *M. tuberculosis* latency. Therefore, the analysis based on the interaction network and the conditional responses captured from diverse latency models reveal a potential model of *M. tuberculosis* latency.

Major contribution of the thesis is towards emphasizing the importance of network dynamics and associated biological responses. Along with proposing a novel approach for studying the changing profile of protein interactions, thesis also presents a systems overview of the nature of correlated and anticorrelated gene pairs. In addition, systems perspective of *M. tuberculosis* provides some of the interesting hypotheses related to the mechanism of latency switch. Functional characterization of some of the identified proteins in the context of *M. tuberculosis* latency might prove useful in enhancing our knowledge on tuberculosis disease mechanisms. Furthermore, genome-wide data on protein functional linkages and gene expression correlations, which are made freely available on the database, will be a useful resource for *M. tuberculosis* researchers in general.

Chapter 1

Introduction and Review of Literature

1.1 Complexity

Complexity exists at each level of organization, the spectrum of which varies from subatomic particles to societies to ecosystems. Despite this hierarchy, properties such as emergence, robustness, modularity and evolution are common to all the levels. Interestingly, these systems are neither truly deterministic nor are purely random, which makes the system to be characterized as 'existing on the edge of chaos' [Mazzocchi 2008]. Such properties impart interesting characteristics to the system, for example dynamism and adaptability to internal and external perturbations. In a broader sense, complex systems can be characterized those possessing number of individual components; the complex and dynamic interactions among which lead to multiple pathways, and hence renders evolvability. The representation of such systems requires, among other representations, nonlinear differential equations [Whitesides and Ismagilov 1999].

Biological systems are inherently complex at each level of organization in terms of their number of interacting components and also their behavioral outcome. For example, response of a cell to its environment involves sensing external factors, coordinated transcription regulation and feed-back mechanisms [Weng et al. 1999]. Similarly, nervous system offers an interesting example with its complex structural makeup and highly efficient information processing ability. In humans, billions of neurons function in coherence to perform cognitive tasks [Koch and Laurent 1999; Singer 2007]. Complex adaptive system such as ecosystem represents a dynamic interaction between hierarchy of organisms and the biosphere [Levin 1998; Hartvigsen et al. 1998].

As the system components interact non-linearly, new properties emerge at each scale of complexity [Flake 2000]. For instance, interaction between sensor kinase and its cognate response regulator in a two-component system brings about specific cellular responses upon sensing environmental conditions [Alex and Simon 1994]. Here, molecular interactions result in a complex cascade of reactions, the outcome of which is not easily explained by the sum total of individual proteins. A new set of laws, methods and perspectives are required to understand the implications of such events. Additionally, complex systems acquire robustness with parallelism and

redundancy suggesting their adaptive behavior to a variety of environmental conditions. This can be illustrated in a cellular pathway where multiple routes function to synthesize the end product. Notably, the scale of redundancy can also be seen at parts lists' level wherein individual components are usually replaceable in a system [Flake 2000]. Another interesting property of a complex system is its modular organization with compartmentalized functional units. Such segregation of functions makes the system highly efficient and robust. For instance, an organism can be viewed as a modular unit of multiple organ systems which are specialized for a function, and each system works in coherence with other systems to bring about complex organismal behavior. Hence, the features such as emergence, robustness and modularity characterize a complex system.

An important aspect of complex systems is that the properties that could be attributed to components are different from what would emerge out of their interactions. As the American physicist P. W. Anderson noted "The behavior of large and complex aggregates of elementary particles, it turns out, is not to be understood in terms of a simple extrapolation of the properties of a few particles.....Psychology is not applied biology, nor is biology applied chemistry" [Anderson 1972]. Individual components develop collective behavior by self-organization; the characterization of such systems involves dynamic variables that share nonlinear relation to produce complex behavior. Philosophy of reductionism emphasizes on the decomposition of complex phenomena into smaller parts, enabling simpler investigation. On the other hand, holism takes into account the properties that emerge at each hierarchical level of component interactions, and argues for new laws to describe each level of organization. For example, understanding disease requires not just about gene mutation but the context in which a specific protein functions. Similarly, disease treatment involves, in addition to knowing molecular functions, patient specific and other socio-cultural factors. Therefore, mere knowledge of a non-functional enzyme is not useful in systems level understanding to model a disease [Beresford 2010]. One has to therefore understand the structure of the system, interactions among its components and the underlying principles of decision making in order to predict the behavior of a complex system.

Biological complexity is beginning to be appreciated by visualizing them as complex network of interacting molecules. General systems theory, which deals with organization of complex systems, their control mechanisms, hierarchies and component interactions, is widely applied to study underlying complexity in biological systems [Stelling et al. 2004; Oberhardt et al. 2009]. The need for such holistic approach in drug discovery and development of vaccines is reviewed by Van Regenmortel (2004). In a similar context, properties such as self-organization, modularity and centrality have been studied to understand biological systems [Barabasi and Oltvai 2004].

1.2 Systems Biology

Systems biology is an approach to study biological complexity by focusing on system structure, its dynamics, control methods and design principles [Kitano 2002]. An underlying conviction of systems biology is that the components of a system are interconnected, and therefore a thorough understanding requires analysis of the system in its entirety. Systems biological studies can be 'top-down', wherein comprehensive '-omics' data is used to describe molecular mechanisms, or 'bottom-up' which considers formulations derived from individual components for model building [Bruggeman and Westerhoff 2006]. In another view, systems level analysis begins with integrating large scale experiments and other available resources which are subsequently modelled *in silico.* The hypotheses thus derived are tested using computational simulations and/or experiments on the bench [Butcher et al. 2004]. Figure 1 illustrates generalized steps that are involved in a typical systems biological study.

Over a decade, systems perspective in biological studies has increased our insights into biochemical pathways, underlying molecular interactions and control mechanisms. In addition to providing fundamental understanding of biological phenomena, such studies find immense applications in areas as diverse as drug discovery and synthetic biology. Graph theoretical and simulation studies on biomolecular networks have become efficient means of understanding cell interactomes [Barabasi and Oltvai 2004; Oberhardt et al. 2009]. Systems engineering

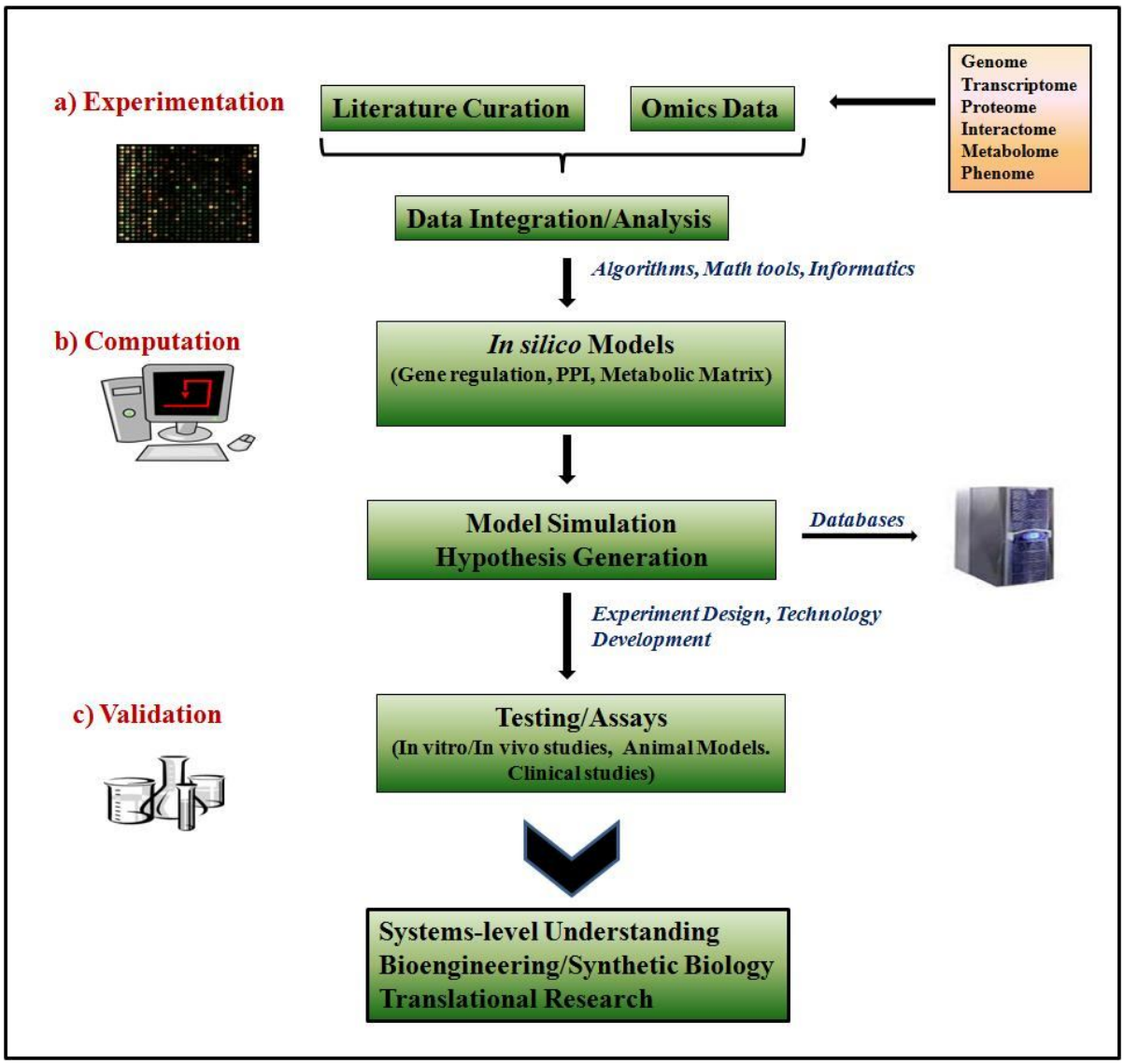

Figure 1: Systems biological research. Studies in systems biology are driven by large-scale experiments combined with existing literature (a). These are then reconstructed to derive models of gene regulation, protein-protein interactions or biochemical pathways using computational algorithms and mathematical tools. Biological hypotheses are subsequently generated using model simulations and analysis (b) and are validated (c). In addition to enhancing our understanding of biological principles, such studies form the basis for synthetic biology and translational research.

studies have revealed general properties of biological systems such as redundancy, robustness and modularity [Stelling et al. 2004]. Approaches such as integration of omics data for gene perturbation or drug administration, computational models of disease physiology and monitoring combinatorial effects have emerged as some of the

alternatives for effective drug discovery [Butcher et al. 2004]. In a similar line, pharmaco-metabonomics and meatagenomic studies are the promising approaches for molecular epidemiology as well as personalized healthcare [Nicholson 2006]. Moreover, Genome-scale metabolic models have been constructed for number of organisms and these find applications in areas such as metabolic engineering, studying pathway conservation across species, and visualizing metabolism as a network of interconnected pathways to improve gene annotations and select drug targets. All these aspects of metabolic reconstruction have been reviewed in Oberhardt et al. [2009]. Importantly, knowledge of system structure and its dynamics, which is the ultimate objective of systems biology, is the basis for synthetic biology and systems engineering [Smolke and Silver 2011].

The future of systems biology hence appears promising for improving our fundamental understanding of biological complexity and its manipulation for the betterment of our lives. For a computational systems biologist, the tasks in hand are to develop novel algorithms for the analysis of large scale data, incorporate them into successful models, and develop appropriate visualization tools. Data from genome-wide experiments come with substantial amount of noise, and to derive meaningful information from such sources is indeed challenging. To derive an organism-level perspective by consolidating information at different levels is the eventual goal of systems biology.

1.3 Graph Theory: Notations

One of the efficient ways to represent biological complexity is by visualizing them as a network of interacting components. The structure, dynamics and evolution of such networks (also termed graphs) are studied using the methods in graph theory. Following are the notations that are routinely used in graph theory [Mason and Verwoerd 2007].

1.3.1 Nodes and Edges

A finite graph G consists of a set of vertices (or nodes) $V(G)$, such that $V(G) = \{v1.....vn\}$, and an edge set $E(G)$ where each edge $(u,v) \in E(G)$. In a broader classification, graphs can be either directed or undirected. For a directed graph, edge (u,v) begins at u and ends at v [Figure 2(a)].

1.3.2 Degree

For an undirected graph G, the degree of a node u is the number of links of u in the graph. Therefore,

$$k = \deg(u) = |N(u)|$$

where, $N(u) = \{v \in V(G) : (u,v) \in E(G)\}$. For a directed graph, node u has two degree components: in-degree ($k_{in}(u)$) and out-degree ($k_{out}(u)$). The number of edges that terminate at node u is termed $k_{in}(u)$ and the number of edges that begin at u is termed $k_{out}(u)$ [Figure 2(b)]. The nodes with high degree are termed as hubs in the network.

1.3.3 Clustering Coefficient

Clustering coefficient denotes the density of interactions in the neighborhood of a node u in the graph G. For an undirected graph, clustering coefficient of a node u is given by:

$$C_u = \frac{2e}{k(k-1)}$$

where, k is the degree of the node u. The average clustering coefficient of the network is the measure of probability that there exists an edge between neighbors of a randomly chosen node in the network [Figure 2(c)].

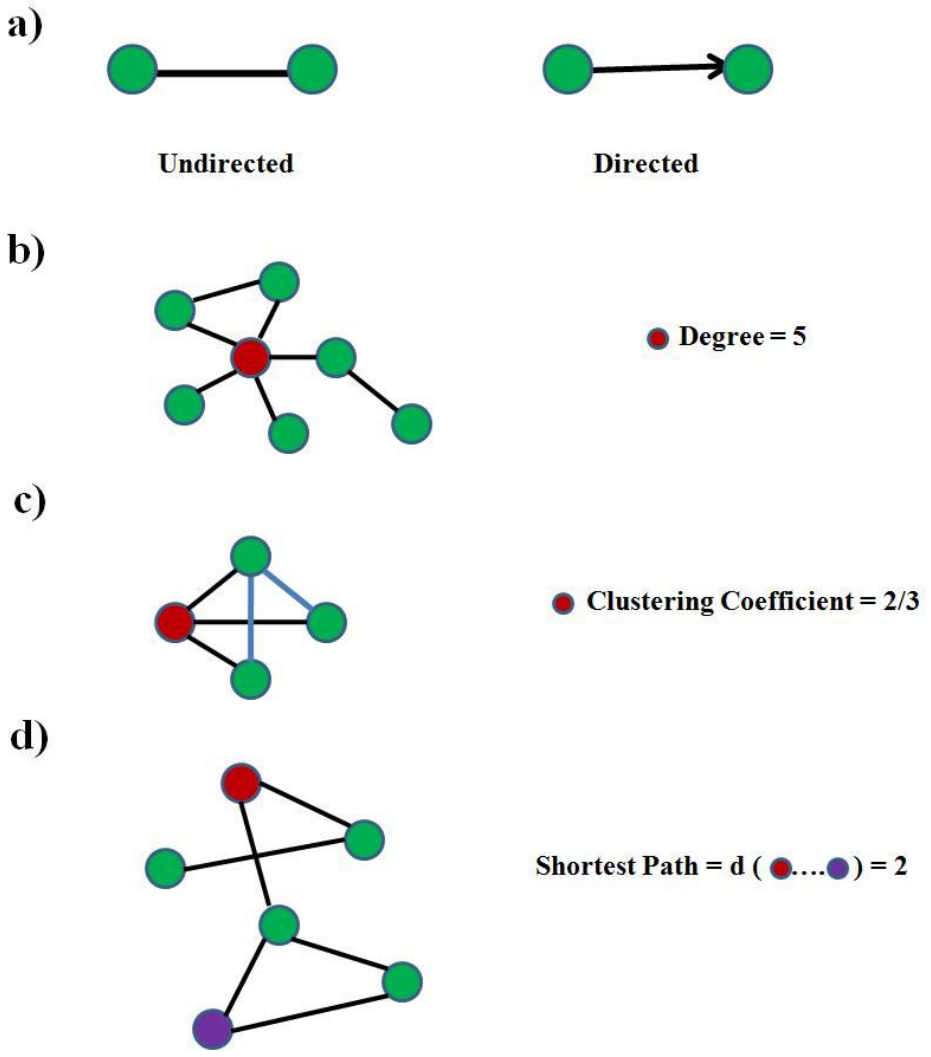

Figure 2: Graph theoretical notations a) Interactions are either directed or undirected b) In an undirected network, degree of the node is the number of interactions of that node. As illustrated, the degree of a node colored in red is 5 c) Clustering coefficient measures the local density of interactions. For a node in an undirected graph, it is calculated as the ratio of the number of interactions neighbors have and the possible number of interactions between neighboring nodes. Therefore in the figure above, clustering coefficient of the red colored node is 2/3 and d) Shortest path length is the minimum number of steps taken to reach from a source node to the target node. The shortest path length from the red node to the blue node is 2.

1.3.4 Pathlength (shortest path) and Network Diameter

Pathlength (or geodesic distance) $p_G(u,v)$, from nodes u and v is the minimum number of steps to be taken to reach from node u to v in graph G [Figure 2(d)]. The graph G is said to be connected if, for each pair of nodes (u,v) in graph G, there exists a path from u to v. Average pathlength of graph is the average of shortest paths of all possible pairs.

Network diameter of graph G is the maximum value of the set of pathlengths in G. Therefore,

$$d = \max_{u,v \in G} p_G(u,v)$$

where d is the network diameter.

1.3.5 Small world phenomenon

A small world network is the network where average pathlength is of the same order of $\log(n)$ where n is the total number of vertices in the network [Watts and Strogatz 1998].

1.3.6 Degree Distribution

Degree distribution $P(k)$ is the probability that a randomly selected node has k links. It is calculated as,

$$P(k) = \frac{n_k}{n}$$

Where n_k is the number of nodes in the network with degree k and n is the number of nodes number of nodes in the network.

1.3.7 Subgraphs, motifs and modules

A subset of nodes from a graph that are connected in a specific manner is termed a subgraph [Barabasi and Oltvai 2004]. Motifs are the subgraph structures that are overrepresented in the network compared to random networks [Milo et al. 2002].

Modules, on the other hand, are the group of nodes that are highly connected among them to achieve a specialized function [Hartwell et al. 1999].

1.4 Mathematical Models of Networks

Based on the topological parameters and characteristic features that have been observed for many real world networks, several mathematical models have been proposed to account for the origin and evolution of networks. Described below are some of the models to understand network structures.

1.4.1 Random Networks

The random network model by Erdös and Rényi [Figure 3(A)] begins with N nodes, and for each pair of nodes (u,v), an edge is placed with a probability p. Thus generated graph will have approximately $pN(N-1)/2$ randomly placed edges. Importantly, the degree distribution of such graphs follows Poisson distribution and the clustering coefficient is independent of the degree of a node [Erdös and Rényi 1960].

1.4.2 Scale-free Networks

The Barabási–Albert model of scale-free networks [Figure 3(B)] considers the evolution of network structure. The dynamics of network involves two components, a) *Growth:* at each time j, new node with degree m is added to the network, and b) *Preferential attachment:* Probability that the newly added node is connected to the existing node u is proportional to the degree of the node u. Scale-free networks show power law degree distribution $P(k) \sim k^{-\Upsilon}$ with $2<\gamma<3$, where the probability of finding a high degree node is lower. Similar to random networks, clustering coefficient of a node does not depend on its degree in scale-free networks [Barabási and Albert 1999].

1.4.3 Hierarchical Networks

Hierarchical networks are generated by combining existing clusters in an iterative manner to generate hierarchical structure [Figure 3(C)]. This incorporates modularity which is observed in many real world networks. The distribution of degree as well as clustering coefficient follows power law. Such structure denotes that nodes with

smaller degree are part of highly clustered regions, whereas a few high degree nodes connect such highly clustered regions in the neighborhood [Ravasz 2002].

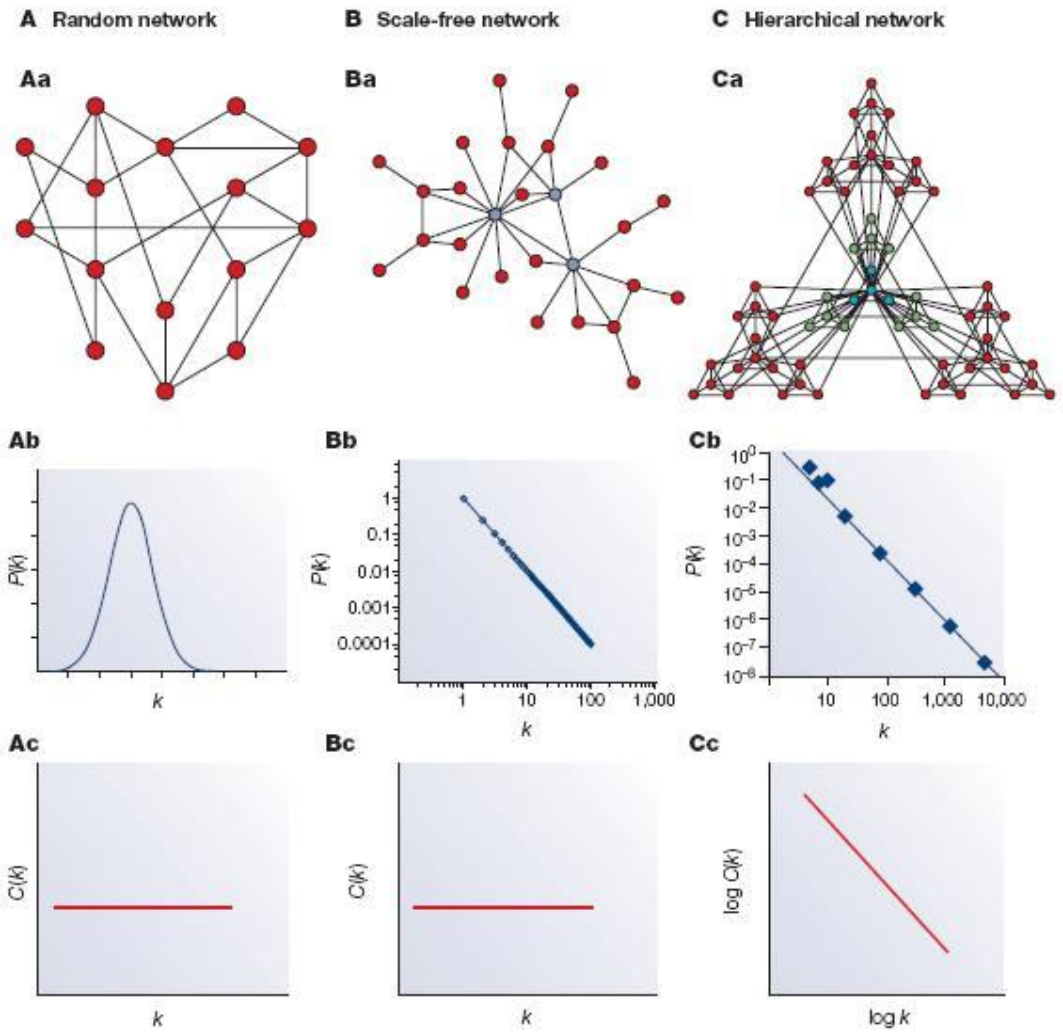

Figure 3: Network Models. a) Random network model: Growth of the network is by adding an edge to the node with a probability *p* (Aa). Network has Poisson degree distribution (Ab) and clustering coefficient is independent of node degree. b) Scale-free network model: Network evolution features are growth and preferential attachment (Ba). Degree distribution of scale-free network follows power law (Bb) and clustering coefficient is independent of node degree (Bc) and c) Hierarchical network model: It is generated by iterative addition of clusters (Ca). Degree distribution, as well as clustering coefficient distribution, follows power law (Cb and Cc). Reprinted by permission from Macmillan Publishers Ltd: [Nat Rev Genet.] (Barabási AL and Oltvai ZN. (2004) Network biology: understanding the cell's functional organization. *Nat Rev Genet.* 5: 101–113), copyright (2004).

1.5 Centrality Measures in Networks

In order to characterize importance of individual nodes in a network, it is to be noted that not all nodes are equally important in a network in terms of maintaining topological features or mediating communication within the network. Traditionally, the concept of centrality is used to assess criticality of nodes in the maintenance of topological parameters of complex networks. Centrality measures are typically used to find important individual in social network, to rank most relevant web pages in the World Wide Web or to identify essential gene in the protein interaction networks [Freeman 1979; Page et al. 1998; Jeong et al. 2003]. Discussed below are the three widely used centrality measures in graph theory [Latora and Marchiori 2007].

1.5.1 Degree Centrality

Degree centrality is based on the assumption that an important node makes greater number of connections in the network. For an undirected graph G with N number of nodes, the degree centrality of the node i is given by,

$$C^D{}_i = \frac{\sum_{j \in G} a_{ij}}{N-1}$$

1.5.2 Closeness Centrality

Whereas degree centrality is a measure of local importance of a node, closeness centrality considers global positioning of a node in the network. The underlying assumption in closeness centrality is that a node which is closer to all other nodes in the network is critical for a network. For an undirected graph G with N number of nodes, the closeness centrality of the node i is given by,

$$C^C{}_i = \frac{N-1}{\sum_{j \in G} d_{ij}}$$

where d_{ij} is the shortest path between the nodes i and j.

1.5.3 Betweenness centrality

A node may be important for mediating information transfer between two other nodes in the network. Betweenness centrality captures this aspect of criticality for a node in the network. Conceptually, it measures the number of shortest paths that pass through a given node in the network. For an undirected graph G with N number of nodes, the betweenness centrality of the node i is given by,

$$C_i^B = \frac{\sum_{j<k \in G} n_{jk}(i) / n_{jk}}{(N-1)(N-2)}$$

where n_{jk} is the number of shortest paths between the nodes j and k, and $n_{jk}(i)$ is the number of shortest paths between j and k that traverse through the node i.

1.6 Network Robustness

Interaction networks, which are the abstract representation of complex systems, are robust in terms of structural properties as well as function [Barabasi and Oltvai 2004]. Due to scale-free topology, probability that a network will disintegrate due to accidental node failure is very less [Albert et al. 2000]. However, attack on nodes which possess high degree will break the network into isolated clusters. Biological insight of such a property is shown by Jeong and co-workers in *S. cerevisiae* protein interaction network. About 10% of essential genes in *S. cerevisiae* are observed to have links less than 5. On the other hand, more than 60% of essential genes have over 15 interactions, indicating that essential genes are highly connected in the protein interaction network [Jeong et al. 2001]. Additionally, segregation of functional units as modules in the network also imparts robustness to the system, wherein, vulnerability of a single module does not usually hamper the structural and functional integrity of the rest of the modules [Barabasi and Oltvai 2004]. Overall, network robustness highlights the ability of the system to respond to external perturbations and changing environmental conditions.

1.7 Biomolecular Networks

Cellular processes are driven by interactions among various biomolecules, and the complexity of such interplay can be illustrated by an interaction map [Zhu et al. 2007]. Based on the type of molecules interacting and the nature of interactions, biomolecular interactions can be broadly divided into five classes [Figure 4(a-e)].

1.7.1 Protein Interaction Networks: Protein-protein interactions represent physical or functional association between two proteins [Eric and Fields 1995; Bork et al. 2004]. Large scale two-hybrid studies have been performed to study genome-wide protein interactions in organisms such as *S. cerevisiae, D. melanogaster* and *C. elegans* [Uetz et al. 2000; Giot et al. 2003; Li et al. 2004]. In addition, affinity purification technique followed by mass spectrometry has identified protein complexes in *E. coli* and *S. cerevisiae* [Butalnd et al. 2005; Krogan et al. 2006]. Protein functional linkages, on the other hand, can be predicted by genome-context methods and similar other computational approaches [Yellaboina et al. 2007]. Protein interaction networks find immense applications in understanding biological complexity. Graph theoretical analysis on protein interaction networks has enhanced our understanding of gene essentiality, modular organization of functional pathways and protein function [Jeong et al. 2001; Barabasi and Oltvai 2004]. Large-scale experimental and predicted protein interactions for a number of organisms can be accessed from databases such as STRING, BIND and DIP [von Mering et al. 2003; Bader et al. 2003; Xenarios et al. 2002].

1.7.2 Gene Regulatory networks: Gene regulatory interactions denote regulatory mechanisms between transcription factors and their target genes. Such interactions are directed, where the source node is the transcription factor and the regulated gene is the target node. Reconstruction of gene regulatory networks can be either literature curated or data driven, or both [Herrgard et al. 2004]. Genome-wide ChIP-chip experiments have been performed to identify regulatory interactions in *S. cerevisiae* and have been assembled into regulatory network [Lee et al. 2002]. Using gene expression data and regulatory sequence information, Wang et al (2002) have

identified transcription modules that are conditionally activated. In *E. coli*, literature curated gene regulatory network was used to identify commonly occurring motif structures [Shen-Orr et al. 2002]. Studies on the topology of regulatory network in *S. cerevisiae* revealed positioning of the transcription factor in the network hierarchy and its correlation to expression dynamics [Jothi et al. 2009]. Effects of gene duplication and gene transfer on the evolution of regulatory networks have been reviewed in [Teichmann and Babu 2004; Perez and Groisman 2009]. The database RegulonDB houses literature curated as well as computationally predicted gene regulatory interactions of *E. coli* [Gama-Castro et al. 2008].

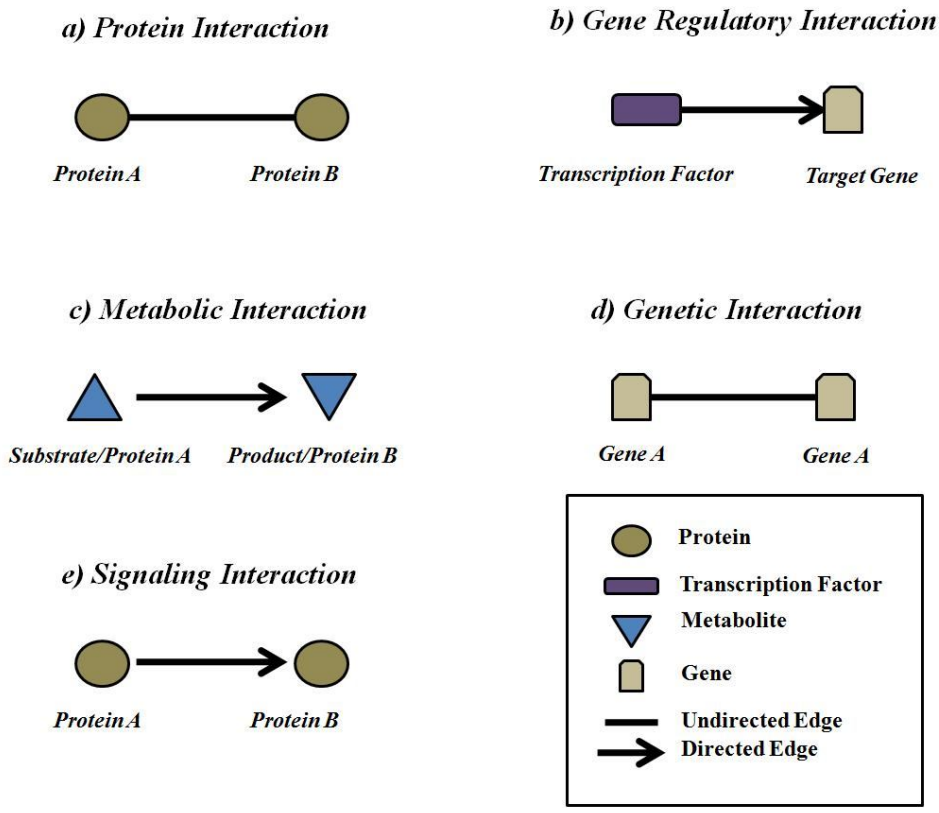

Figure 4: Biomolecular Networks. a) Protein interaction networks are undirected representing physical or functional link between two proteins b) Gene regulatory interactions are the directed link between a transcription factor as the source node and the regulated gene as its target c) Metabolic interactions represent conversion of one metabolite to another metabolite in a chemical reaction d) Genetic interactions are the synthetic interactions between two genes and e) Signaling interaction denotes direction of information flow in the cellular signaling cascade.

1.7.3 Metabolic networks: Metabolic networks represent biochemical interactions in cellular pathways. A directed edge denotes an enzymatic reaction where a substrate is transformed into a product [Jeong et al. 2000]. Genome-scale metabolic models have been constructed for *E. coli* and *S. cerevisiae* to study metabolism and phenotypic behavior [Reed et al. 2003; Duarte et al. 2004]. Metabolic networks also have been successfully used to study network dynamics. In *S. cerevisiae,* metabolic model was used for *in silico* gene deletion studies performed across multiple growth conditions [Duarte et al. 2004]. Using metabolic networks, constraint based analyses such as flux balance analysis (FBA) was used to account for gene dispensability in *S. cerevisiae* [Papp et al. 2004]. KEGG database provides metabolic pathway information for multiple organisms which are derived by extensive literature curation and genome annotation [Kanehisa and Goto 2000].

1.7.4 Genetic Interaction networks: Two genes are said to be interacting genetically if the simultaneous deletion of the genes leads to cell death, or the mutation in one gene either enhances or represses the phenotype of mutation in the other gene [Guarentel 1993]. In *S. cerevisiae*, systematic construction of double mutants generated large-scale genetic interaction network which revealed clusters of functionally related genes and the cross-wiring between different pathways [Tong et al. 2001; Costanzo et al. 2010]. Similar double deletion study followed by two dimensional hierarchical clustering of the data identified distinct functional modules in *E. coli* [Butland et al. 2008]. Furthermore, computational approaches have been suggested to construct genome-wide genetic interaction networks [Pandey et al. 2010]. In addition to revealing novel functional associations between genes/pathways, genetic interactions also assist in assigning functional classes to unannotated genes [Lippert et al. 2010].

1.7.5 Signaling networks: Signaling networks represent signal transduction events wherein nodes are proteins or small molecules and the transfer of signals is denoted by an edge between them [Pawson and Scott 1997]. Reconstruction of signaling networks, their complexity, and spatio-temporal dynamics of cellular signaling

mechanisms have been reviewed recently [Papin et al. 2005; Kholodenko 2006]. Graph theoretical analysis of the interaction network specific to RNAi targeted cellular signaling machinery revealed dynamics of signaling modules pertaining to cell cycle progression [Jailkhani et al. 2011]. The Alliance for Cellular Signaling (AfCS)–Nature 'Molecule Pages' is the database designed to collect information about molecules involved in cell signaling and the pathways [Li et al. 2002].

1.8 Protein Interaction Networks

Every cellular process is driven by interactions between protein molecules, the nature of which may vary widely in a living cell. Certain classes of proteins can be found as multisubunit entities, some of the classical examples being hemoglobin, ribonucleotide reductase and protein kinase A. Protein complexes, on the other hand, are transiently interacting cellular machineries to perform functions such as transcription, translation, folding and transport [Phizicky and Fields 1995]. Transient interactions among individual proteins are also critical for large number of processes including signal transduction, protein modifications and enzyme catalysis [Phizicky and Fields 1995]. Therefore, a cell can be viewed as a tightly regulated web of interacting proteins, and understanding such complexity requires studying them as a whole system. A popular means of comprehending such complex systems is by representing them as a network of interacting entities. Subsequent graph theoretical analysis and mathematical modeling of networks is a promising approach to unravel such complex molecular interactions [Barabasi and Oltvai 2004].

Based on the nature of association, protein interactions can be either physical or functional. Physical interaction implies physical contact with molecular docking which is observed in signal transduction events, enzyme catalysis of the substrate proteins etc. On the other hand, functional interaction signifies participation of proteins in the same pathway, or association of proteins in complex machinery in which proteins might not necessarily interact physically [Rivas and Fontanillo 2010]. Decades of research in biology has accumulated large amount of data on protein interactions which could be compiled to construct high-confident interaction network. Over the years, high-throughput experiments have been performed in various organisms to

generate large-scale protein interaction maps. Table 1 lists some of the databases for retrieving protein interactions, and visualization tools for interaction networks.

Protein Interaction Databases		
Database	**URL**	**Reference**
BIND	http://bind.ca/	Bader et al. 2003
STRING	http://string.embl.de/	von Mering et al. 2003
DIP	http://dip.doe-mbi.ucla.edu/	Xenarios et al. 2002
HPRD	http://www.hprd.org/	Peri et al. 2003
BioGRID	http://www.thebiogrid.org/	Stark et al. 2006
MINT	http://mint.bio.uniroma2.it/mint/	Chatr-aryamontri et al. 2007
MIPS	http://mips.helmholtz-muenchen.de/proj/ppi/	Pagel et al. 2005
IntAct	http://www.ebi.ac.uk/intact/	Aranda et al. 2010
MiMI	http://mimi.ncibi.org/MimiWeb/main-page.jsp	Gao et al. 2009
OPHID	http://ophid.utoronto.ca/	Brown and Jurisica 2005
Predictome	http://visant.bu.edu/	Mellor et al. 2002
Network Visualization Tools		
Tool	**URL**	**Reference**
Cytoscape	http://www.cytoscape.org/	Shannon et al. 2003
Visant	http://visant.bu.edu/	Hu et al. 2008
Pajek	http://vlado.fmf.uni-lj.si/pub/networks/pajek/	Batagelj and Mrvar 1998
Osprey	http://biodata.mshri.on.ca/osprey/servlet/Index	Breitkreutz et al. 2003

Table 1: List of protein interaction databases and network visualization tools

Comprehensive interaction maps have been generated for *S. cerevisiae* using yeast-two hybrid technique [Ito et al. 2000, 2001; Uetz et al. 2000] and affinity purification followed by mass spectrometry [Gavin et al. 2002; Ho et al. 2002]. Yeast-two hybrid technique is also used to generate interaction maps for other model organisms such as *C. elegans* [Li et al. 2004] and *D. melanogaster* [Giot et al. 2003]. Using TAP-mass spectrometry for *E. coli*, two studies have generated large maps of interacting protein complexes [Butland et al. 2005; Arifuzzaman et al. 2006]. For the human pathogen *H. pylori*, protein interaction network is built using large scale yeast-two hybrid method [Rain et al. 2001]. In addition, multiple datasets for human protein

interactions is made available by both affinity purification followed by mass spectrometry and yeast-two hybrid technique [Ewing et al. 2007; 2004 Rual et al. 2005; Stelzl et al. 2005]. However, such large-scale experimental studies are limited to only model organisms for technical reasons.

The recent availability of complete genome sequences for multiple organisms makes it feasible to predict genome-wide protein interaction networks for several organisms including pathogens [Chapter 5 in this thesis]. Furthermore, computational approaches are faster, less expensive and generally provide good coverage of protein interactions. Several genome-context based algorithms have been proposed to predict protein interactions from genome-sequences [Galperin and Koonin 2000; Valencia and Pazos 2002]. Three such methods have been used to predict functional linkages for *M. tuberculosis* [Strong et al. 2003]. Using genome-context methods combined using support vector machine, genome-wide protein functional linkages were predicted for *E. coli* [Yellaboina et al. 2007]. Databases such as BIND and STRING house protein interactions derived from both experiments and computational predictions (Table 1).

1.9 Computational Methods for Protein Interaction Prediction

Following are the widely applied computations approaches to predict protein-protein interactions in completely sequenced genomes.

1.9.1 Phylogenetic Profile

The basis for Phylogenetic profile method is that two interacting proteins are conserved evolutionarily across genomes (Figure 5). This method was proposed by Pellegrini et al (1999) wherein such profile was created for the proteins of *E. coli* across the then sequenced 16 other genomes. Enault et al (2003) have suggested a method to improvise phylogenetic profile construction by replacing bit scores with normalized scores of the BlastP value. Combination of phylogenetic profiles with gene proximity information was applied in *E. coli* to identify protein functional linkages [Zheng et al. 2002]. With the availability of large number of genome sequences, this method has been proven to be highly useful for predicting protein functional linkages [Strong et al. 2003; Yellaboina et al. 2007].

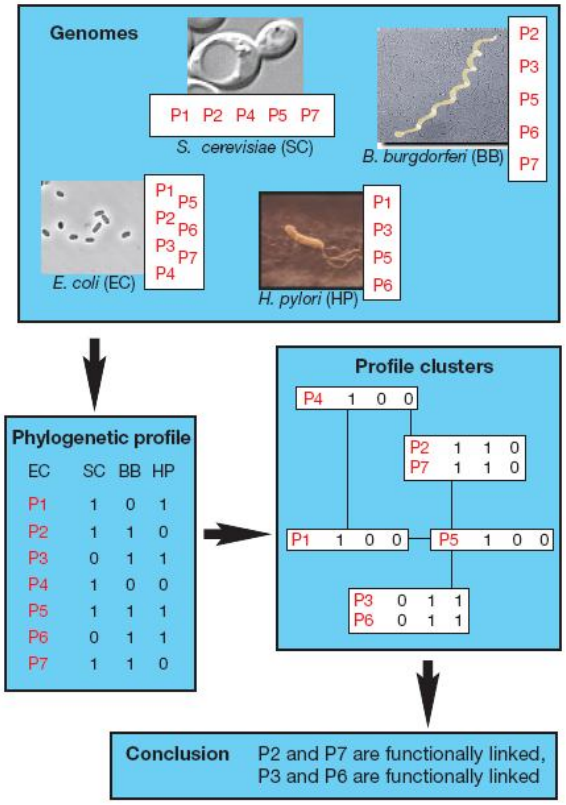

Figure 5. Construction of a Phylogenetic profile. Proteins are clustered based on their absence or presence across multiple genomes, and the pairs that are co-conserved are likely to interact functionally. In the figure above, protein P2 and P7, and P3 and P6 are functionally linked as they are clustered together according to their phylogeny. Reprinted by permission from Macmillan Publishers Ltd: [Nature] (Eisenberg D, Marcotte EM, XenariosI, and Yeates TO (2000) Protein function in the post-genomic era. *Nature* 405: 823–826), copyright (2000).

1.9.2 Domain Fusion/Rosetta Stone

Domain fusion (also called Rosetta stone) prediction method is based on the observation that two individual proteins which occur as a fused protein in another organism are likely to be functionally linked (Figure 6). Marcotte et al have studied domain fusion events to predict putative protein interactions in *E. coli* and *S. cerevisiae* [Marcotte et al. 1999]. The same approach was used by Enright et al to detect about 88 fusion events in three query genomes [Enright et al. 1999]. Domain fusion, however, will not be applicable in detecting those interactions which have

evolved through mechanisms other than gene fusion, or the cases where fused protein has disappeared during evolution. In addition, promiscuous domains such as SH3 might interfere with the accuracy of interaction prediction based on domain fusion [Marcotte et al. 1999]. Domain fusion is used as one of the features to predict protein functional linkages in the pathogenic bacterium *M. tuberculosis* [Strong et al. 2003].

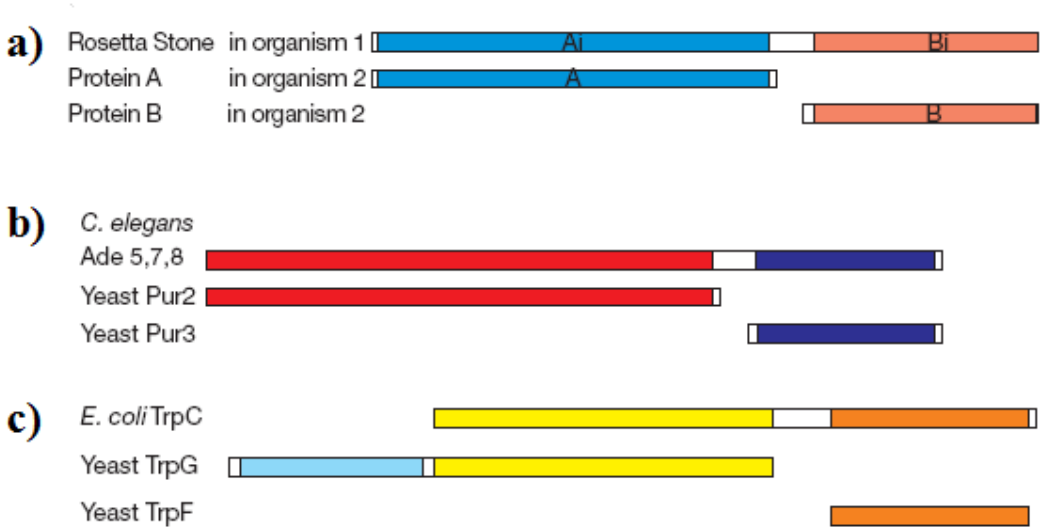

Figure 6. Domain fusion (or Rosetta Stone) method for functional linkage prediction. If two independent proteins in one organism occur as fused domains in another organism, then they are likely to share a functional relationship (a). In the figure above, two yeast proteins Pur2 and Pur3 occur as fused protein Ade 5,7,8 in *C. elegans* (b). Similarly, *E. coli* protein TrpC is composed of domains which are found in yeast proteins TrpG and TrpF (c). Reprinted by permission from Macmillan Publishers Ltd: [Nature] (Eisenberg D, Marcotte EM, XenariosI, and Yeates TO (2000) Protein function in the post-genomic era. *Nature* 405: 823–826), copyright (2000).

1.9.3 Gene expression correlation

Similarity in the expression pattern is one of the non-homology based methods to predict protein functional association [Marcotte 2000]. The assumption for prediction is that two interacting proteins share functional relation, and therefore should exhibit similar expression pattern across growth conditions (Figure 7). For *S. cerevisiae,* correlated mRNA expression was used as one of the methods to derive functional links between proteins [Marcotte et al. 1999]. Also, relation between protein interaction and

correlated expression has been studied in four diverse species [Bhardwaj and Lu 2005]. Based on these studies, correlations derived from large-scale gene expression datasets for *M. tuberculosis* has been used as one of the predictive features for protein interaction prediction [Chapter 4 in this thesis].

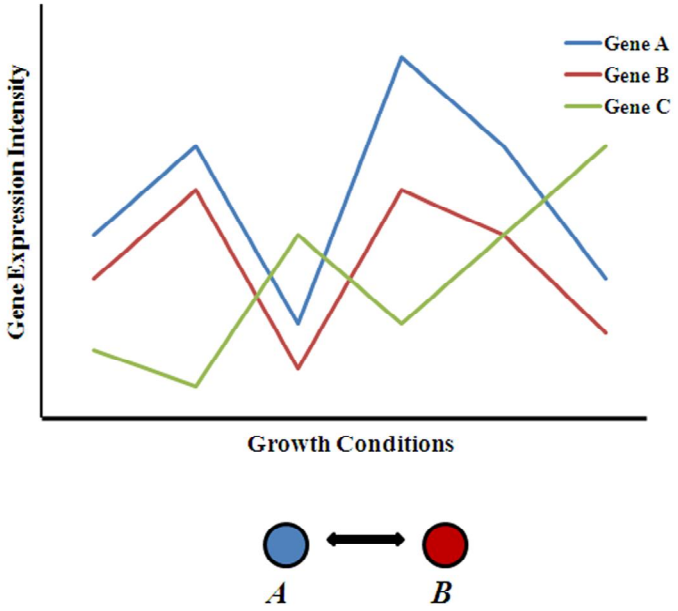

Figure 7. Gene Expression correlations. Interacting proteins have similar expression pattern at their mRNA level across number of growth conditions. In the illustration above, proteins A and B are likely to interact as they exhibit similar expression pattern across growth conditions.

1.9.4 Conserved Gene Neighborhood

Conserved gene neighborhood is a method that predicts protein interactions based on proximity between gene pairs across genomes (Figure 8). Dandekar and colleagues have showed higher probability for conserved gene clusters or pairs to interact physically [Dandekar et al. 1998]. Taking *trp* operon as an example, authors note the conservation of operon architecture for many gene pairs [Dandekar et al. 1998]. Based on gene directionality and distance, gene clusters have been identified in bacterial chromosomes, the conservation of which indicates a close functional relationship [Overbeek et al. 1999]. Similarly, Korbel and colleagues report chromosomal proximity as a measure of co-regulation, which in turn implies functional association between proteins [Korbel et al. 2004]. Since functionally related genes occur as operonic pairs

in a genome, and this structure is observed to be conserved across genomes [Dandekar et al. 1998], prediction based on genome-context methods for *E. coli* included operonic frequency as one of the features [Yellaboina et al. 2007]. This study also employed gene distance measure that is calculated as distance between translation start sites between two genes [Yellaboina et al. 2007].

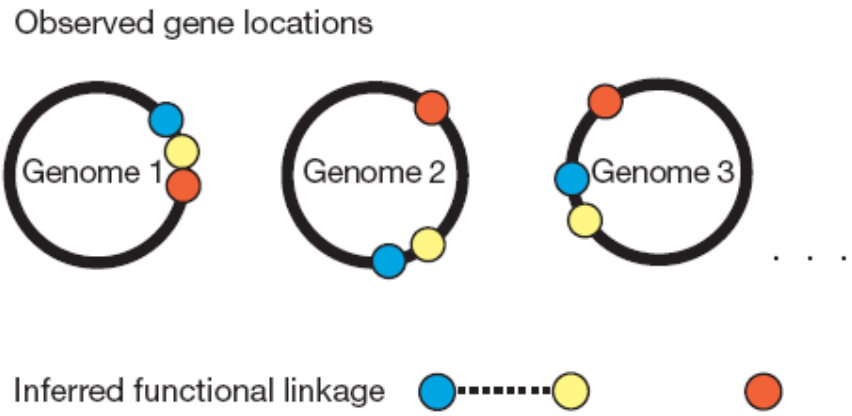

Figure 8. Gene neighborhood method for functional linkage prediction. Protein pairs show higher likelihood to share functional linkage if the genes encoding them are closely located in multiple genomes. The closeness can be measured in terms of gene distance, operonic structure or as gene clusters. In the cartoon above, proteins colored in blue and yellow share a functional relationship as their genes are located close on multiple genomes. Reprinted by permission from Macmillan Publishers Ltd: [Nature] (Eisenberg D, Marcotte EM, XenariosI, and Yeates TO (2000) Protein function in the post-genomic era. *Nature* 405: 823–826), copyright (2000).

Since each computational method captures different aspect of protein interactions, it is ideal to combine multiple predictive features. The modest approach in this regard is to take consensus from different computational methods for interaction prediction. Strong and colleagues have considered those interactions in *M. tuberculosis* for further analysis which were predicted by at least two of the four methods used for prediction [Strong et al. 2003]. For *S. cerevisiae,* large set of protein interactions accumulated from experiments and computational predictions were classified into different confidence category based on the source of interaction [Marcotte et al. 1999]. In *E. coli,* machine learning algorithm, Support Vector Machine (SVM), was used for genome-wide functional linkage prediction [Yellaboina et al.

2007]. A similar machine learning based prediction is made for the proteins of *M. tuberculosis* using genome-context methods and gene expression data as predictive features [Chapter 4 in this thesis]. Several databases including protein interaction database STRING provide confidence score for each interaction based on number of methods reporting a particular interaction and the predictive capacity of each method [von Mering et al. 2003].

1.10 Network Evolution and Dynamics

Most of the network studies carried out so far do not take into consideration the dynamics associated with its structure. One aspect of network dynamics is its evolution over a large time-scale, which indeed is a slow change. Evolution of interaction networks is attributed to genetic changes such as gene duplication and gene loss, which affect both the nodes (i.e proteins) and the links in the networks. In addition, point mutations, insertion or deletion events also contribute to changes in the protein structure, which in turn will affect the associations between protein pairs [Yamada and Bork 2009]. Interestingly for gene regulatory networks, gene duplication has been shown to be a major event in the network evolution [Teichmann and Babu 2004]. It is observed that hubs have slow evolutionary rate suggesting that they are highly conserved [Prachumwat and Li 2006]. The duplication or loss of hubs, or their changes at structural level often have undesirable effects as they affect large number of connections that hubs make. Furthermore, environment also poses substantial amount of constrains for evolution [Yamada and Bork 2009].

Another important aspect of network dynamics is associated with the changes at spacio-temporal scale in the cell. Intuitively, not all the protein interactions that are experimentally shown or computationally predicted take place together all the time in a cell. Subset of genes is expressed based on environmental conditions, and therefore interactions at protein level are inherently dynamic. Protein interactions also vary based on the cellular localization of a protein. Embryogenesis is an example where spatially co-ordinated gene expression leads to progression of systematic developmental stages. In a multicellular organism, each cell type has its own set of gene regulatory circuits, and hence the network of interacting proteins varies

considerably across these tissue types. Interestingly, Han and co workers classify hubs in the *S. cerevisiae* protein interaction networks as 'date' and 'party' hubs. Date hubs connect to different interacting partners dynamically whereas party hubs interact with all the partners simultaneously [Han et al. 2004]. In a murine protein interaction network, Lu and colleagues observe that proteins with higher connectivity show less expression changes in a condition mimicking asthma [Lu et al. 2007]. Komurov and White identify static and dynamic modules in the protein interaction network of *S. cerevisiae* which differ in function as well as network topology [Komurov and White 2007]. In a similar line, dynamics of genome-wide protein functional linkages specific to wild type and UV treated *E. coli* cells were studied using graph theoretical measures to understand cellular responses upon perturbation [Hegde et al. 2008]. Using gene expression data, active subnetworks of *M. tuberculosis* gene regulatory network (termed 'responsive origons') during hypoxia and stationary phase were identified [Balazsi et al. 2008]. Interestingly, different hierarchical layers of *S. cerevisiae* transcription factors have varied dynamic properties ensuring effective cellular responses to changing conditions [Jothi et al. 2009]. Therefore, dynamics of protein interactions is the integrated aspects of interaction evolution and spatio-temporal expression.

Biomolecular networks, including protein interactions, are the abstract representations of complex cellular systems. Even with such simplified version of molecular interactions, networks pose complexity in terms of growth, diversity of nodes and the type of interactions [Strogatz 2001]. The foremost challenge is to construct high confidence network at genome-scale for a given organism. Notably, there is poor overlap between different methods used for deriving interaction maps, the reason for which could be either a high rate of false positives or false negatives for each method [Futschik et al. 2007]. Nonetheless, interaction maps have contributed immensely to study the design of biological systems. However, understanding of such systems become complete only when interaction dynamics at spatio-temporal scale is taken into consideration. In addition, networks describing interactions of different biomolecules have to be integrated to arrive at a true map of cellular interactions. A notable suggestion is to study networks in the context of post-translational regulations

by factors such as miRNAs [Zhu et al. 2007]. Analysis of such a network, though at a higher scale of complexity, would reveal orchestration of a variety of cellular interactions. A new direction in network biology is therefore to integrate all versions of biomolecular interactions and study them in the perspective of evolution and spatio-temporal dynamics.

1.11 References

1. Albert R, Jeong H and Barabasi AL (2000) Error and attack tolerance of complex networks. *Nature* **406:** 378-382.

2. Alex LA and Simon MI (1994) Protein histidine kinases and signal transduction in prokaryotes and eukaryotes. *Trends Genet* **4:** 133-138.

3. Anderson PW (1972) More is different. *Science* **177:** 393-396.

4. Aranda B, Achuthan P, Alam-Faruque Y, Armean I, Bridge A et al. (2010) The IntAct molecular interaction database in 2010. *Nucleic Acids Res* **38:** D525-D531.

5. Bader GD, Betel D and Hogue CW (2003) BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res* **31:** 248-250.

6. Barabasi AL and Albert R (1999) Emergence of scaling in random networks. *Science* **286:** 509-512.

7. Barabási AL and Oltvai ZN (2004) Network biology: understanding the cell's functional organization. *Nat Rev Genet* **5:** 101-113.

8. Batagelj V and Mrvar A (1998) Pajek - Program for Large Network Analysis. *Connections* **21:** 47–57.

9. Beresford MJ (2010) Medical reductionism: lessons from the great philosophers. *QJM* **103:** 721-724.

10. Bork P, Jensen LJ, von Mering C, Ramani AK, Lee I et al. (2004) Protein interaction networks from yeast to human. *Curr Opin Struct Biol* **14:** 292-299.

11. Breitkreutz BJ, Stark C and Tyers M (2003) Osprey: a network visualization system. *Genome Biol* **4:** R22.

12. Brown KR and Jurisica I (2005) Online predicted human interaction database. *Bioinformatics* **21:** 2076-2082.

13. Bruggeman FJ and Westerhoff HV (2007) The nature of systems biology. *Trends Microbiol* **15:** 45-50.

14. Butcher EC, Berg EL and Kunkel EJ (2004) Systems biology in drug discovery. *Nat Biotechnol* **22:** 1253-1259.

15. Butland G, Babu M, Díaz-Mejía JJ, Bohdana F, Phanse S et al. (2008) eSGA: *E. coli* synthetic genetic array analysis. *Nat Methods* **5:** 789-795.

16. Butland G., Peregrin-Alvarez J.M., Li J., Yang W., Yang X et al. (2005) Interaction

network containing conserved and essential protein complexes in *Escherichia coli*. *Nature* **433:** 531–537.

17. Chatr-aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV et al. (2007) MINT: the Molecular INTeraction database. *Nucleic Acids Res*. **38:** D572–D574.

18. Costanzo M, Baryshnikova A, Bellay J, Kim Y, Spear ED et al. (2010) The genetic landscape of a cell. *Science* **327:** 425-431.

19. De Las Rivas J and Fontanillo C (2010) Protein-protein interactions essentials: key concepts to building and analyzing interactome networks. *PLoS Comput Biol* **6:** e1000807.

20. Duarte NC, Herrgård MJ and Palsson BØ (2004) Reconstruction and validation of *Saccharomyces cerevisiae* iND750, a fully compartmentalized genome-scale metabolic model. *Genome Res* **14:** 1298-1309.

21. Erdos P and Renyi A (1960) On the evolution of random graphs. *Publ Math Inst Hung Acad Sci* **5:** 17-61.

22. Flake GW (2000) The Computational Beauty of Nature: Computer Explorations of Fractals, Chaos, Complex systems, and Adaptaion. The MIT Press, Massachusetts.

23. Freeman LC (1979) Centrality in social networks: Conceptual clarification. *Social Networks* **1:** 215–239.

24. Futschik ME, Chaurasia G and Herzel H (2007) Comparison of human protein-protein interaction maps. *Bioinformatics* **23:** 605-611.

25. Gama-Castro S, Jiménez-Jacinto V, Peralta-Gil M, Santos-Zavaleta A, Peñaloza-Spinola MI et al. (2008) RegulonDB (version 6.0): gene regulation model of *Escherichia coli* K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Res* **36:** D120-D124.

26. Gao J, Ade AS, Tarcea VG, Weymouth TE, Mirel BR et al. (2009) Integrating and annotating the interactome using the MiMI plugin for cytoscape. *Bioinformatics* **25:** 137-138.

27. Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B et al. (2003) A protein interaction map of *Drosophila melanogaster. Science* **302:** 1727-1736.

28. Guarente L (1993) Synthetic enhancement in gene interaction: a genetic tool come of age. *Trends Genet* **9:** 362-366.

29. Hartvigsen G, Kinzig A and Peterson G (1998) Complex Adaptive Systems: Use and Analysis of Complex Adaptive Systems in Ecosystem Science: Overview of Special Section. *Ecosystems* **1:** 427–430.

30. Hartwell LH, Hopfield JJ, Leibler S and Murray AW (1999) From molecular to modular cell biology. *Nature* 402: C47-C52.

31. Hegde SR, Manimaran P and Mande SC (2008) Dynamic changes in protein functional linkage networks revealed by integration with gene expression data. *PLoS Comput Biol* **4:** e1000237.

32. Herrgård MJ, Covert MW and Palsson BØ (2004) Reconstruction of microbial transcriptional regulatory networks. *Curr Opin Biotechnol* **15:** 70-77.

33. Hu Z, Snitkin ES and DeLisi C (2008) VisANT: an integrative framework for networks in systems biology. *Brief Bioinform* **9:** 317–325.

34. Jailkhani N, Ravichandran S, Hegde SR, Siddiqui Z, Mande SC et al. (2011) Delineation of key regulatory elements identifies points of vulnerability in the mitogen-activated signaling network. *Genome Res* [Epub ahead of print].

35. Jeong H, Mason SP, Barabási AL and Oltvai ZN (2001) Lethality and centrality in protein networks. *Nature* **411:** 41-42.

36. Jeong H, Oltvai ZN, and Barabasi AL (2003) Prediction of protein essentiality based on genomic data. *ComPlexUs* **1:** 19–28.

37. Jeong H, Tombor B, Albert R, Oltvai ZN and Barabási AL (2000) The large-scale organization of metabolic networks. *Nature* **407:** 651-654.

38. Jothi R, Balaji S, Wuster A, Grochow JA, Gsponer J et al. (2009) Genomic analysis reveals a tight link between transcription factor dynamics and regulatory network architecture. *Mol Syst Biol* **5:** 294.

39. Kanehisa M and Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28:** 27-30.

40. Kholodenko BN (2006) Cell-signalling dynamics in time and space. *Nat Rev Mol Cell Biol* **7:** 165-76.

41. Kitano H (2002) Systems biology: a brief overview. *Science* **295:** 1662-1664.

42. Koch C and Laurent G (1999) Complexity and the nervous system. *Science* **284:** 96-98.

43. Krogan NJ, Cagney G, Yu H, Zhong G, Guo X et al. (2006) Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae Nature* **440:** 637-643.

44. Latora V and Marchiori M (2007) A measure of centrality based on network efficiency. *New J Phys* **9:** 188.

45. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z et al. (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae. Science* **298:** 799-804.

46. Levin SA (1998) Ecosystems and the Biosphere as Complex Adaptive Systems. *Ecosystems* **1:** 431-436.

47. Li J, Ning Y, Hedley W, Saunders B, Chen Y et al. (2002) The Molecule Pages database. *Nature* **420:** 716-717.

48. Li S, Armstrong CM, Bertin N, Ge H, Milstein S et al. (2004) A map of the interactome network of the metazoan *C. elegans*. *Science* 303: 540-543.

49. Lippert C, Ghahramani Z and Borgwardt KM (2010) Gene function prediction from synthetic lethality networks via ranking on demand. *Bioinformatics* **26:** 912-918.

50. Mason O and Verwoerd M (2007) Graph theory and networks in Biology. *IET Syst Biol* **1:** 89-119.

51. Mazzocchi F (2008) Complexity in biology. Exceeding the limits of reductionism and determinism using complexity theory. *EMBO Rep* **9:** 10-14.

52. Mellor JC, Yanai I, Clodfelter KH, Mintseris J and DeLisi C (2002) Predictome: a database of putative functional links between proteins. *Nucleic Acids Res* **30:** 306–309.

53. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D et al. Network motifs: simple building blocks of complex networks. *Science* **298:** 824-827.

54. Nicholson JK (2006) Global systems biology, personalized medicine and molecular epidemiology. *Mol Syst Biol* 2: 52.

55. Oberhardt MA, Palsson BØ and Papin JA (2009) Applications of genome-scale metabolic reconstructions. *Mol Syst Biol*. 5: 320.

56. Page L, Brin S, Motwani R, and Winograd T (1998) The PageRank Citation Ranking: Bringing Order to the Web, Stanford Digital Library Technologies Project, Tech. Rep., 1998. http//citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.31.1768.

57. Pagel P, Kovac S, Oesterheld M, Brauner B, Dunger-Kaltenbach I et al. (2005) The MIPS mammalian protein-protein interaction database. *Bioinformatics* **21:** 832-834.

58. Pandey G, Zhang B, Chang AN, Myers CL, Zhu J et al. (2010) An integrative multi-network and multi-classifier approach to predict genetic interactions. *PLoS Comput Biol* **6:** e1000928.

59. Papin JA, Hunter T, Palsson BO and Subramaniam S (2005) Reconstruction of cellular signalling networks and analysis of their properties. *Nat Rev Mol Cell Biol* **6:** 99-111.

60. Papp B, Pál C and Hurst LD (2004) Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast. *Nature* **429:** 661-664.

61. Pawson T and Scott JD (1997) Signaling through scaffold, anchoring, and adaptor proteins. *Science* **278:** 2075-2080.

62. Perez JC and Groisman EA (2009) Evolution of transcriptional regulatory circuits in bacteria. *Cell* **138:** 233-244.

63. Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK et al. (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res* **13:** 2363–2371.

64. Phizicky EM and Fields S (1995) Protein-protein interactions: methods for detection and analysis. *Microbiol Rev* **59:** 94-123.

65. Prachumwat A and Li WH (2006) Protein function, connectivity, and duplicability in yeast. *Mol Biol Evol* **23:** 30-39.

66. Ravasz E, Somera AL, Mongru DA, Oltvai ZN and Barabási AL (2002) Hierarchical organization of modularity in metabolic networks. *Science* **297:** 1551-1555.

67. Reed JL, Vo TD, Schilling CH and Palsson BO (2003) An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR). *Genome Biol* **4:** R54.

68. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT et al. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13:** 2498–2504.

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

70. Singer W (2007) Understanding the brain. How can our intuition fail so fundamentally when it comes to studying the organ to which it owes its existence?

EMBO Rep **8:** S16-19.

71. Smolke CD and Silver PA (2011) Informing biological design by integration of systems and synthetic biology. *Cell* **144:** 855-859.

72. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A et al. (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* **34:** D535–D539.

73. Stelling J, Sauer U, Szallasi Z, Doyle FJ 3rd and Doyle J (2004) Robustness of cellular functions. *Cell* **118:** 675-685.

74. Strogatz SH (2001) Exploring complex networks. *Nature* **410:** 268-276.

75. Teichmann SA and Babu MM (2004) Gene regulatory network growth by duplication. *Nat Genet* **36:** 492-496.

76. Tong AH, Lesage G, Bader GD, Ding H, Xu H et al. (2004) Global mapping of the yeast genetic interaction network. *Science* **303:** 808-813.

77. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS et al. (2000) A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* **403:** 623-627.

78. Van Regenmortel MH (2004) Reductionism and complexity in molecular biology. Scientists now have the tools to unravel biological and overcome the limitations of reductionism. *EMBO Rep* **5:** 1016-1020.

79. von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P et al. (2003) STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res* **31:** 258-261.

80. Wang W, Cherry JM, Botstein D and Li H (2002) A systematic approach to reconstructing transcription networks in *Saccharomyces cerevisiae*. *Proc Natl Acad Sci U S A* **99:** 16893-16898.

81. Watts D and Strogatz S (1998) Collective dynamics of small-world networks. *Nature* **393:** 440–442.

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

83. Whitesides GM and Ismagilov RF (1999) Complexity in chemistry. *Science* **284:** 89-92.

84. Xenarios I, Salwínski L, Duan XJ, Higney P, Kim SM et al. (2002) DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res* **30:** 303-305.

85. Yellaboina S, Goyal K and Mande SC (2007) Inferring genome-wide functional linkages in *E. coli* by combining improved genome context methods: comparison with high-throughput experimental data. *Genome Res* 17: 527-535.

86. Zhu X, Gerstein M and Snyder M (2007) Getting connected: analysis and principles of biological networks. *Genes Dev* **21:** 1010-1024.

Chapter 2

Dynamic Changes in Protein Functional Linkage Networks Revealed by Integration with Gene Expression Data

---

**[*This Chapter is Published as:*
Hegde SR, Manimaran P and Mande SC (2008) Dynamic changes in protein functional linkage networks revealed by integration with gene expression data. *PLoS Comput Biol* 4: e1000237]**

2.1 Introduction

Gene expression pattern in all organisms is a property of the environmental conditions in which they grow. Expression of a large number of genes is turned on or off conditionally and temporally allowing the organisms to adapt to different growth or changing environmental conditions. While some genes are constitutively expressed under many different conditions, presumably being essential for the organism to carry out basic cellular processes for growth and sustenance, many genes are expressed only under defined conditions. DNA microarray offers a powerful tool to study such gene expression profiling. Studying the gene expression pattern under different conditions therefore offers an attractive approach to study the response of an organism to changing environmental conditions.

The traditional analysis of microarray data involves measuring differential expression between two samples after background elimination and data normalization. An unsupervised classification method such as clustering or principal component analysis is popularly used to identify genes that have a similar regulation pattern [Murphy 2002; Slonim 2002]. Although measuring relative gene expression levels is the preferred method of analysis, the individual signal intensities, which contain valuable information on absolute gene expression, are often not considered. Studying the pattern of absolute expression of genes, rather than relative expression between two conditions, might provide an alternate useful approach for comparative analysis.

A few attempts have been made to analyze differences in gene expression arising out of different conditions of growth. The gene expression profiling in *E. coli* has revealed varied mRNA transcripts in the cells growing in minimal and rich media as well as in their exponential and transitional phases of growth [Wei et al. 2001]. In a similar line, the gene expression dynamics and its relevance to *E. coli* physiology is shown by the protein expression profiles and their correlation with the gene expression profiles [Champion et al. 2003; Corbin 2003]. Absolute gene expression analysis in fission yeast has shown that the basic cellular functions are carried out by the conserved genes and that the organism specific genes are expressed conditionally for the specialized processes [Mata and Bahler 2003]. These studies have provided a

wealth of data on the molecular and genetic basis of response of organisms to changing environmental conditions.

While the analysis of gene expression data provides useful insights into the adaptation process, it is believed that the response of organisms is dictated by the dynamics of biomolecular interactions profile. One of the inspirations to carry out the present study was to understand the changing landscape of protein-protein interactions under different environmental conditions. The protein-protein interaction studies carried out experimentally usually represent only a fraction of all the possible interactions among different cellular proteins [von Mering et al. 2002]. Moreover, the protein interaction networks available, for example in *E. coli* [Butland et al. 2005; Arifuzzaman et al. 2006], represent only the static protein interaction networks and are able to capture the interactions involving only those genes whose products are expressed under the unique experimental conditions. On the other hand, understanding of the dynamics of protein interaction networks demands profiling genome-wide protein-protein interactions under many different experimental conditions. Such experiments are currently prohibitive in time and resources. Therefore, profiling interactions under changing environmental conditions presents enormous challenge to the experimental biologists.

There have been a few attempts to combine protein: protein interaction networks and gene expression data [Horvath and Dong 2008]. The dynamics of yeast interactome studied using mRNA expression revealed two kinds of hubs namely, “date hubs” and “party hubs” [Han et al. 2004]. The former are believed to bind different proteins at different time or location, and the latter are believed to bind to their partners simultaneously. Also, the hub proteins were shown to have lower levels of differential expression compared to the non-hub proteins [Lu et al. 2007]. Further, it was observed that static and dynamic proteins cluster into different modules in yeast protein interaction network [Komurov and White 2007]. In another study the PPI networks were studied in the context of genes expressed during aging [Xue et al. 2007]. These sub-networks and the modules therein were examined for understanding the aging process. Thus, although a few attempts have been made in integrating the gene expression data with protein interaction networks, the analysis of differential

gene expression in the context of corresponding networks remains an underexplored area. This study is an attempt in this direction

2.2 Results

Gene expression information of *E. coli* has been used to identify the genes that are expressed in the prevailing conditions and reduced networks have been constructed from the predicted genome-wide parent functional linkage network. Figure S1(A) in Appendix I schematically shows the approach, which is used for the analysis of *E. coli* expression data. The sub-networks thus constructed are hypothesized to represent a real functional interaction picture of the cell. Further, various graph theoretical measures have been applied to extract the relevant biological information from these sub-networks. This is proposed to be a novel methodology in which a raw microarray data can be analyzed by incorporating molecular interaction information with gene expression.

2.2.1 Construction of the conditional protein-protein interaction networks

Predicted functional interaction network for *E. coli,* which comprises 78,048 interactions among 3,682 proteins, was used as the parent network [Yellaboina et al. 2007]. This functional linkages network was obtained by training a Support Vector Machine on high confidence interactions in the EcoCyc database and assuming that cytoplasmic and periplasmic protein do not interact with each other. The predicted data set has fewer interactions in common with the experimentally derived networks [Butland et al. 2005; Arifuzzaman et al. 2006]. However, the overlap increases significantly if the indirect interactions are taken into consideration.

As a case study, gene expression data from UV exposure in wild type and SOS deficient *E. coli* at 20 minutes post irradiation has been chosen for the analysis [Courcelle et al. 2001]. The four conditions studied are Untreated Wild Type (UWT), UV treated Wild Type (TWT), untreated *lexA* mutant (UML) and UV treated *lexA* mutant (TML). Graph theoretical measures were applied to the four sub-networks that were derived by this methodology, and then compared amongst each other (Figure S1(B) in Appendix I).

The *E. coli* genome has the capability to encode more than 4000 genes. Out of these, approximately 40-60% are likely to be expressed under any defined condition [Champion et al. 2003]. Using the methodology described in Materials and Methods, processing of the raw microarray data revealed expression of around 2000 genes (Table S1 in Appendix I), which is in close agreement with the earlier studies on absolute gene expression [Corbin 2003]. Networks under each of the four conditions (henceforth referred to as conditional networks) were constructed by mapping the expressed proteins on the parent network. The conditional networks possess around 30,000 interactions among the expressed proteins.

2.2.2 Global properties of the conditional networks

It is anticipated that the effect of turning off or on of the genes expressed under the four conditions will be reflected in the conditional networks. While, this is likely to lead to many local perturbations in the network, the global properties of the four networks are not likely to change significantly. Various topological properties of the conditional networks under the perturbations such as mutation (*lexA*) or UV treatment along with the network corresponding to wild-type were therefore studied.

The four conditional networks exhibit similar network parameters (Table 1). The core cluster comprises >95% of the nodes in the network. The overlap of the interactions and the nodes in UWT-TWT and UML-TML is shown in Figure S2 in Appendix I. The degree distributions in all the four conditional networks show power law behavior with the degree exponent of 1.1. Therefore, the networks are scale-free, indicating that these are similar to other real world networks. The scale-free property also is suggestive of their resistance to random node failure [Albert et al. 2000]. The network diameters for the studied graphs imply the small-world property where the number of steps required for reaching from one node to the other is not more than 9. Other properties such as average clustering coefficient and mean eccentricity are similar in all the graphs (Table 1). The fractal dimensions calculated using cluster growing method indicates that the networks are self similar with all the four conditional networks possessing similar fractal dimensions. Similarly, network efficiency is also comparable in the conditional networks in spite of the imposed

perturbations. Thus, there is no significant difference in the global network properties of the four conditional networks.

Property	Parent Network	UWT	TWT	UML	TML
Nodes	3,682	1,899	1,865	1,957	1,947
Edges	78,048	34,893	34,680	31,900	33,513
Percentage core Nodes	96.9	97.4	97.9	96.1	95.5
Average Degree	42.4	36.7	37.2	32.6	34.4
Degree Exponent	1.2	1.1	1.1	1.1	1.1
Diameter	11	8	8	8	9
Mean Eccentricity	7.99	5.66	5.78	5.89	6.07
Average Clustering coefficient	0.23	0.21	0.21	0.22	0.22
Network efficiency	0.36	0.37	0.38	0.36	0.37

Table 1: Global properties of the sub-networks. Global network parameters for the parent network (15) and the conditional networks. UWT – UV Untreated Wild Type, TWT – UV Treated Wild Type, UML – UV Untreated *lexA* mutant, TML – UV Treated *lexA* mutant.

2.2.3 Unique nodes of the conditional networks

Each of the four conditional networks possesses unique nodes corresponding to the genes that are expressed differentially. Interestingly, the uniquely expressed genes include a few hubs and transcription factors. The lists of proteins that are identified to be uniquely expressed are listed in the Table S2 in Appendix I. As anticipated, some of the proteins involved in DNA repair, recombination and cell structure determination are observed only in the UV treated wild type network in comparison with the untreated wild type network. In addition to the UV damage response genes, other environmental stress related genes are also uniquely expressed

in the UV treated wild type cells. Thus, the comparative analysis of PPI networks appears to identify a few crucial features that might be physiologically relevant.

Mapping of the unique nodes of each of the four comparison sets to different metabolic pathways for *E. coli*, as enlisted in KEGG database [Kanehisa and Goto 2000], revealed that replication and repair proteins, as expected, are more in number in the UV treated networks compared to their untreated counterparts (Figure S3 in Appendix I). Interestingly, many genes coded on carbohydrate metabolism operons are seen only in the untreated wild type network. The presence of these gene clusters in the untreated wild type cells and their absence in UV irradiated cells appears to suggest the repression of sugar metabolism in UV treated *E. coli* cells. Earlier study has shown that a few carbohydrate metabolism operons indeed exhibit reduced expression under UV exposure [Courcelle et al. 2001]. Similarly, it is observed that the genes belonging to membrane transport more in number as uniquely expressed in the *lexA* mutant cells when compared with the respective wild type cells. Thus, the four-way comparisons of expressed genes, as anticipated, highlight the importance of DNA repair and replication processes under UV exposure.

One of the interesting genes that is observed to be expressed only in the UV treated cells is the *hda* gene, protein product of which is involved in DnaA inactivation. It has been demonstrated earlier that cells suppress replication upon DNA damage. As Hda is known to repress hyper initiation of DNA replication by inactivating DnaA [Banack et al. 2005], the criticality for Hda in the UV treated networks appears to be significant. Interestingly, this protein has a high degree in the UV treated networks (Table S3 in Appendix I). Since high degree nodes are believed to be important in maintaining robustness of graphs [Albert et al. 2000], the high degree of Hda and its unique expression in UV treated cells signifies its importance when the cells are treated with UV radiation. Furthermore, this gene is observed to be expressed under UV exposure, both in the wild type as well as in the *lexA* mutant, suggesting that its expression is independent of the well characterized SOS response.

Another interesting example is the unique expression of genes involved in the iron uptake system in the untreated wild type cells. The proteins EntA, EntB and EntF function in the pathway of enterobactin synthesis and the proteins FepA and

FepB form a part of the channel to transport Fe-enterobactin complex inside the cell. When cells are UV treated, reactive oxygen species (ROSs) are synthesized via photo-Fenton reaction which leads to oxidative damage of structural proteins, enzymes, DNA and lipids. Thus, it is likely that cells repress iron uptake to protect cellular macromolecules from damage. The absence of these iron uptake proteins from UV treated wild type network supports this idea.

The analysis of uniquely expressed nodes under one condition, but not in another condition, indicates some of the possible effects of UV radiation on *E. coli*. The importance of repression of carbohydrate metabolism and iron uptake upon exposure to UV is apparent in the networks. Similarly, an important hub, Hda, is also apparently expressed only upon UV exposure. Thus, the uniquely expressed nodes in the networks indicate of how *E. coli* might respond to UV, thereby suggesting that such an analysis might be useful in other similar studies.

2.2.4 Analysis of the path length differences

An interesting aspect in systems analysis is to study the effect of selective removal of nodes on modifications in the shortest path lengths in the conditional networks. The shortest path lengths in a network signify the efficiency of communication between the nodes, and any alteration in these paths might suggest significance of these nodes under the two conditions. Importantly, the overall diameter of the four conditional networks is identical, indicating that diameter as a global property of the network is not subject to change. Moreover, all the networks have small world property; almost all nodes can be reached from every other in a small number of steps. This is not surprising, considering the biological robustness that is reflected in these networks. Thus, analysis of shortest path lengths might yield interesting insights into the relative importance of communication networks in the four sub-networks.

In order to analyze local changes in pathlength differences, the reduced pathlength matrices were constructed for the common nodes in network pairs under study. The pathlength difference of more than or equal to 3 for each node pair in two reduced networks were considered significant. As expected, for most of the node pairs,

there is no change in the pathlength as there are multiple paths to reach from one node to another node even in the event of a collapse of a particular path. Interestingly, considerable variation in the path for some node pairs was observed, manifesting their reduced connectivity in terms of efficient information exchange, two examples of which are discussed in detail below.

The shortest pathlength from AmyA, a cytoplasmic α-amylase to many of the glycogen metabolism enzymes is observed to be increased in UV treated wild type network (Figure 1). In the untreated wild type subgraph, AmyA is connected to glycogen metabolism enzymes through MalZ, which functions as a maltodextrin glucosidase. The increase in the path length is due to the absence of MalZ node in the cells treated with UV and thereby resulting in isolation of AmyA with respect to proteins belonging to starch and sucrose metabolism in *E. coli*. Thus, the importance of repression of carbohydrate metabolism upon UV treatment is highlighted not only by the repression of a few carbohydrate metabolism operons [Courcelle et al. 2001], but also by the reduced efficiency of communication between different glycogen metabolism genes.

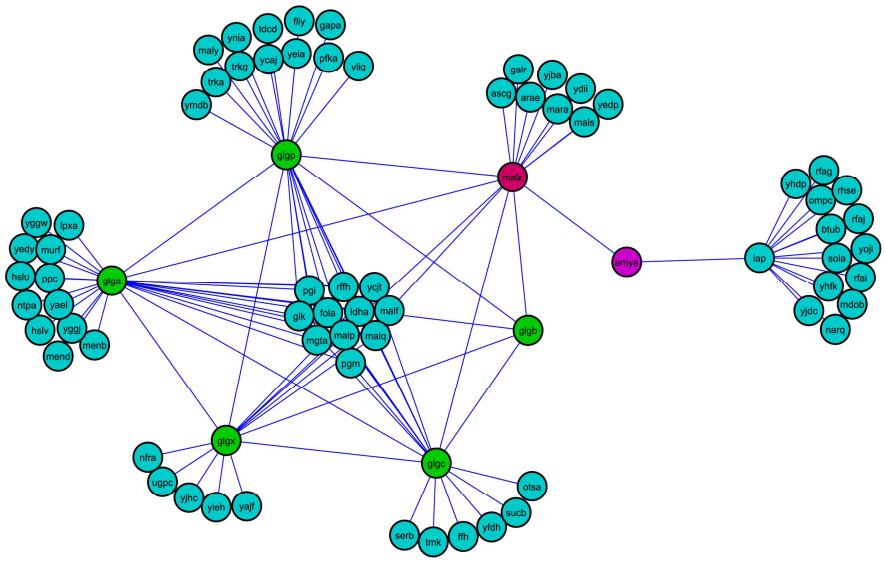

Figure 1: Path length Analysis for the starch and sugar metabolism genes. In the untreated wild type network, the subgraph for starch and sucrose metabolism pathway proteins is well connected. The absence of MalZ in treated wild type network increases the path from AmyA to some of the glycogen metabolism proteins significantly.

Another interesting example pertains to the phosphotransferase system in *E. coli*. The sub-network corresponding to a part of phosphotransferase (PTS) system in *E. coli* in the untreated mutant network is very well connected, whereas the mutant UV treated network lacks CmtB, a component of mannitol PTS permease (Figure S4 in Appendix I). This results in the increased shortest paths from YggD, which is a hypothetical transcriptional regulator of *cmt* operon, to other proteins which are part of the phosphotransferase system in *E. coli*. Since PTS regulates the uptake and metabolism of several sugars, a logical speculation is that the increased paths between YggD and other PTS components might be a strong indication for the temporal repression of sugar metabolism in cells treated with UV radiation.

2.2.5 Expression of the hubs

It has been reported that the highly connected nodes of the network (hubs) are three times more likely to be essential than the poorly connected nodes [Jeong et al. 2001]. Moreover, dynamics of yeast interactome has revealed two kinds of hubs, ones which are present under a variety of conditions, and the ones which appear only under certain specific conditions [Han et al. 2004]. Thus, it is important to analyze the presence or absence of hubs under the four conditions. The hubs of the parent network have been defined as the nodes having degree more than 60 and thereby identified 736 hubs in the network. Interestingly, around 70-75% of the hubs of the parental network are expressed in the conditional networks. As a null hypothesis, when test networks are constructed by choosing random nodes of the parent network, only 50% hubs are found (Figure 2). Thus, essentiality of the hubs appears to manifest itself by the expression of a large number of hubs under all the four different conditions. One such example of the Hda hub was described above. Similarly, the percentage representation of experimentally determined essential genes in *E. coli* [Baba et al. 2006] has been tested in each of the conditional networks compared to random networks. As expected, conditional networks harbour significantly more number of essential genes compared to randomly generated networks (Figure 2).

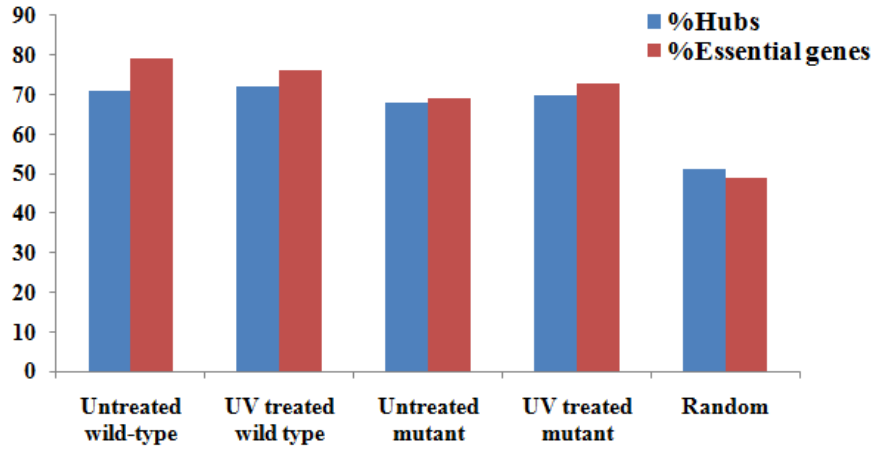

Figure 2: Expression of hubs and essential genes. Conditional networks are observed to represent significantly more number of hubs and essential genes compared to randomly generated networks.

2.2.6 Centrality measures

It is likely that the importance of a functional role of a gene might differ according to the prevailing condition of growth. The relative importance of a node in graph theory can be assessed by calculating various centrality measures. In this direction, different centrality measures of graph theory were analyzed with respect to their relevance to the four sub-networks.

Degree centrality is based on how well the node is connected in a graph. Degree centrality thus states that a node tends to be essential in a network if it is highly connected and its removal has severe impact on the overall topology and connectedness of the network [Albert et al. 2000; Jeong et al. 2001]. Similarly, if a node is positioned in such a way that it can communicate with other nodes quickly then the node is considered to be important in terms of closeness centrality. Betweenness centrality, on the other hand measures the number of shortest paths that traverse through a node. Both closeness and betweenness centrality have also been reported to be good measures to assess gene essentiality [Hahn and Kern 2005; Joy et al. 2005; Yu et al. 2007]. Based on these observations, three centrality values for the nodes in the conditional networks were calculated, and then for each of the nodes the pair wise difference between the different conditions were computed.

To address conditional or relative criticality of a node, the difference in the centrality measures for the common nodes in the comparison set was calculated. The

centrality measure difference is approximately normally distributed, thus about 99.7% values are expected to lie within 3 standard deviations of the mean value. For most of the nodes, there is no change in the centrality value as expected. Further, only those proteins whose centrality measure difference is more than 3 times the standard deviation of the distribution were chosen. When untreated wild type and UV treated wild type networks are compared, the proteins belonging to carbohydrate metabolism and energy metabolism such as BglX, Dld, GatB, GlgA, CydA, CydB and YneH have greater centrality measure in the untreated wild type network. The replication and repair proteins, namely RecN, RecO, Tag, HepA and HolC on the other hand have greater centrality values in the UV treated wild type network. Likewise, DnaA, DnaE, Mfd, RecJ and SbcB functioning in the replication and repair machinery possess significantly higher centrality values in UV treated mutant networks compared to their untreated counterparts. There is no considerable change in the centrality measure for the proteins of the pathways such as polyketide biosynthesis, cell motility and xenobiotics biodegradation. A detailed list of proteins with significant difference in centrality along with their functions in UWT- TWT and UML- TML comparison set is given as Table S4 in Appendix I. In order to check if the standard deviation cutoff (3.0) has any effect on the overall conclusions, the data was reanalyzed on the centrality using a cutoff of 2.0. However, the overall conclusions on the importance of carbohydrate metabolism in untreated cells, and those of DNA replication and repair in the UV-treated cells, remain the same. Therefore, it is concluded that the cutoff value of standard deviation, if modified, does not appear to alter the overall biological conclusions.

Further, to study the essentiality of the nodes depending on the UV treatment or the *lexA* mutation, degree centrality analysis of top 30% nodes in each of the conditional networks was undertaken. Using these criterion more than 550 nodes can be classified as high degree nodes under each condition (Table 2). When the common high degree nodes of wild-type networks (untreated as well as UV treated) and the common high degree nodes of *lexA* mutant (untreated as well as UV treated) networks are compared, interestingly 104 high degree nodes are unique to the wild type networks but are absent in either one or both mutant networks. Similarly 100 high

degree nodes are unique to the mutant networks but are absent in either one or both wild type networks. Therefore, these are considered as the nodes that are essential in a mutation independent and mutation dependent manner respectively. A similar comparison for the high degree nodes in the context of UV treatment revealed 42 nodes as essential depending on UV treatment and 57 nodes as essential when there is no UV treatment (Table 2). Further, these proteins are mapped onto different metabolic pathways in the KEGG database and are classified as *lexA* mutation independent, *lexA* mutation dependent, UV treatment dependent and UV treatment independent (Figure 3 and Table S5 in Appendix I). Some specific examples of this analysis are discussed below.

Conditional Network	High degree nodes	Networks Compared	Common Nodes	Criticality	Number of Proteins
UWT	570	UWT-TWT	527	Mutation Independent	104
TWT	560	UML-TML	523	Mutation Dependent	100
UML	587	UWT-UML	465	UV Independent	42
TML	584	TWT-TML	480	UV Dependent	57

Table 2: Degree Centrality of the Conditional Networks. The analysis of the top 30% nodes in terms of degree centrality alone in each conditional network revealed the nodes that are proposed to be essential depending on the UV treatment or the mutation. UWT – UV Untreated Wild Type, TWT – UV Treated Wild Type, UML – UV Untreated *lexA* mutant, TML – UV Treated *lexA* mutant.

Many repair proteins such as DinG, DnaN, MutM, MutS, RuvC, Rep and RecF that are likely to be indispensable for the UV treated networks were identified as important in terms of degree centrality. The criticality of some of the proteins that belong to lipid metabolism and cofactors and vitamins metabolism seems to be UV treatment dependent. One of the proteins that appears to be important from this anslysis in UV treated cells is UspA, the universal stress protein. Earlier study has

shown the role of UspA in resistance to DNA damaging agents and that its regulation is *lexA* independent [Diez et al. 2000]. The mutants lacking *uspA* were shown to be sensitive to UV irradiation. It is interesting to observe the importance of this protein through network centrality studies, which was otherwise not obvious from the classical microarray analysis. Similarly, the analysis also suggests the importance of the protein, ApaH. ApaH functions as a diadenosine tetraphosphatase and its substrate AppppA has been reported to regulate cell division [Nishimura et al. 1997], affect cell motility and catabolite repression [Farr et al. 1989] and shown to bind to several heat shock and oxidative stress proteins [Johnstone and Farr 1991]. Therefore, the observation that ApaH is critical in the studied networks in an UV dependent manner seems to be relevant in this regard.

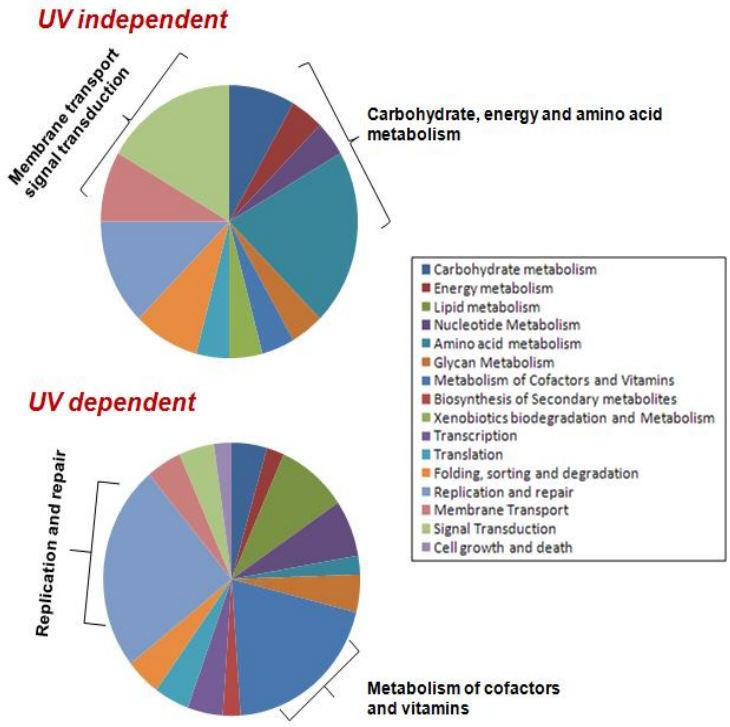

Figure 3: Classification of hubs according to the KEGG metabolic pathways. The high degree nodes of each of the four conditions are classified as critical in a UV treatment dependent or independent manner and *lexA* mutation dependent or independent manner and they are mapped onto different metabolic pathways of *E. coli* as enlisted in KEGG database.

2.3 Discussion

The analysis carried out is based on the predicted genome-wide functional linkages [Yellaboina et al. 2007]. In order to examine if the properties observed for the conditional sub-networks are consistent with those obtained from experimentally validated protein:protein interactions, a similar analysis for the two available experimental protein:protein interaction networks was carried out [Butland et al. 2005; Arifuzzaman et al. 2006]. The core protein interactions in the network reported by Butland *et al.* covers 1,255 proteins and 5,395 interactions among them [Butland et al. 2005]. The average degree for the core network is approximately 8.5 and the clustering coefficient is ~0.085. Similarly, the core protein interactions in the network reported by Arifuzzaman *et al*. covers 2,927 proteins and 11,105 interactions [Arifuzzaman et al. 2006]. The average degree for the core network is approximately 8.5 and the clustering coefficient is ~0.065. The low average degree and clustering coefficient for the experimental networks compared to the predicted functional linkages might be due to the inability of experimental methods to saturate the genome-wide interaction networks.

The obtained conditional networks derived from the experimental interactions show topological robustness similar to their parent networks. Interestingly, similar to the conclusions that have drawn based on the analyses derived from functional linkages network, the UV-dependent criticality of many of the replication and repair proteins through network centrality analysis as well as the analysis of unique nodes of the networks was observed. In addition, the expression of ~65% hubs in conditional networks derived from Arifuzzaman *et al.* data [Arifuzzaman et al. 2006] and ~85% hubs in the ones obtained from Butland *et al*. data [Butland et al. 2005] were observed. These numbers being significantly higher than those by choosing nodes in the networks randomly, suggest that one may obtain biologically important insights through such an analysis.

Some of the cutoffs applied in the current study might appear to be superficially arbitrary. For example, a gene was considered to be expressed if the net signal intensity corresponding to its spot was more than or equal to the median signal intensity of the spots within the sector. Although this cutoff might seem arbitrary, the

rationale for using median was based on the observation that gene expression is a stochastic event and hence the expression of a gene as well as copy number of the expressed protein differs from cell to cell even in an isogenic cell population [Elowitz et al. 2002; Cai et al. 2006]. Despite the inherent stochasticity, the response of a colony of bacteria to external stimuli is based on simultaneous expression of a set of genes. It is reported that the noise in gene expression is inversely proportional to the mean expression level [Bar-Even et al. 2006] and also that essential genes have lower noise in their expression [Fraser et al. 2004]. Therefore, with median as the cutoff, the noisy expression can be eliminated to identify those as genes as expressed which have: (i) high expression levels and (ii) respond to the growth condition. In order to minimize noise in such an identification process, each sector of the chip is considered independently to overcome the differences in the environments within the chip that contribute to expression variances. Thus, with the cutoff of 1.0, there is a reasonable of number of genes being expressed, i.e. ~40-60% [Champion et al. 2003], and also expression of approximately 75% of hub proteins.

Furthermore, the effect of different cutoffs on the overall conclusions of the analysis was tested. With the cutoff of 0.9 and 1.1, earlier conclusions such as increased importance of replication and repair proteins, and cofactor metabolism proteins in the UV treated cells, repression of carbohydrate metabolism upon UV treatment and importance of unique nodes of the conditional networks, remain identical. With further modification of these cutoff values to 1.2, approximately 1600 genes are being expressed which might be considered fewer than anticipated [Champion et al. 2003]. Similarly with modification of the cutoff value to 0.8, many more proteins are considered to be expressed, which might lead to noise in the expression analysis. Thus, although the cutoff value of 1.0 appears arbitrary, it leads to reasonable hypothesis on the response of *E. coli* to UV.

Thus, the comparative analysis does indeed reveal physiologically important changes in the four networks. Some of these changes would not have been apparent by measuring gene expression alone, or by the standard analysis of microarray data. This is partly due to the fact that the levels of expression of many genes do not change under different conditions, but nonetheless the profile of interactions surrounding

them changes significantly, thereby altering their significance in the broader picture of the cell. In this manner, studying the dynamics of protein:protein interactions appears to hold promise for the systems level understanding of an organism.

The analysis proposed in this study can also be potentially applied to disease interaction networks. For example, understanding how the interactions within a pathogen or a host change during the disease process, and the implications of these changes might yield useful insights into the disease. Further, this information can be used to deriving novel therapies against the diseases.

2.4 Methods

2.4.1 Microarray data processing and PPI network construction

The raw microarray data for *E. coli* were downloaded from the Stanford Microarray Database (SMD, http://smd.stanford.edu/) [Demeter et al. 2007]. The SMD lists sector information of the chip, and it is likely that environments within the chip differ considerably. Each spot was therefore assigned to one of sixteen possible sectors in the chip using sector information given in the raw data. A gene was considered to be expressed if the net signal intensity (i.e. background corrected) corresponding to its spot was more than or equal to the median signal intensity of the spots within the sector.

The conditional protein interaction network was built for the expressed genes by mapping them onto an existing predicted functional interaction network for *E. coli* [Yellaboina et al. 2007]. All orphan nodes were removed from the obtained network and the core interaction network was used for further analysis.

2.4.2 Global properties of the network

Network properties such as average degree, degree exponent, diameter, average clustering coefficient were calculated according to [Dorogovtsev and Mendes 2003]. The mean eccentricity was calculated according to [Dankelmann et al. 2004]. Fractal dimension was measured using cluster growing method [Song et al. 2005]. Network efficiency is the property that quantifies how well the nodes of the network exchange information, and this parameter was calculated according to [Latora and Marchiori 2004].

2.4.3 Path length Analysis

The shortest paths for all pairs of nodes in the network were calculated by Dijkstra's algorithm [Dijkstra 1959]. The difference in the path for a node pair in two different networks was analyzed.

2.4.4 Centrality Measures

Network centrality measures like degree centrality, closeness centrality and betweenness centrality were calculated [Latora and Marchiori 2004]. The definitions of these centrality measures are as follows:

Degree centrality of a node *i* in the network G is

$$C_i^D = \frac{\sum_{j \in G} A_{i,j}}{N-1}$$

Where, $A_{i,j}$ is the element in the adjacency matrix A for the nodes *i* and *j*, and N is the total number of nodes in G.

Closeness centrality of a node *i* in the network G is

$$C_i^C = \frac{N-1}{\sum_{j \in G} d_{i,j}}$$

Where $d_{i,j}$ is the shortest path between *i* and *j*.

Betweenness centrality of a node *i* in the network is

$$C_i^B = \frac{\sum_{j<k \in G} n_{jk}(i) / n_{jk}}{(N-1)(N-2)}$$

Where, $n_{jk}(i)$ is the number of shortest paths between *j* and *k* that traverse through *i* and n_{jk} is the total number of shortest paths between *j* and *k*.

2.4.5 Sub-network visualization

Sub-networks were visualized and analyzed using Cytoscape 2.4.1 [Shannon et al. 2003] and NAViGaTOR (http://ophid.utoronto.ca/navigator/).

2.5 References

1. Albert R, Jeong H and Barabasi AL (2000) Error and attack tolerance of complex networks. *Nature* **406:** 378-382.

2. Arifuzzaman M, Maeda M, Itoh A, Nishikata K, Takita C et al. (2006) Large-scale identification of protein-protein intractions of *Escherichia coli* K-12. *Genome Res* **16:** 686-691.

3. Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y et al. (2006) Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol.* **2:** 2006.0008.

4. Banack T, Clauson N, Ogbaa N, Villar J, Oliver D et al. (2005) Overexpression of the Hda DnaA-Related Protein in *Eshcerichia coli* inhibits Multiplication, Affects membrane permeability, and induces SOS response. *J Bacteriol* **187:** 8507-8510.

5. Bar-Even A, Paulsson J, Maheshri N, Carmi M, O'Shea E et al. (2006) Noise in protein expression scales with natural protein abundance. *Nat Genet* **38:** 636-43.

6. Butland G, Peregrin-Alvarez JM, Li J, Yang W, Yang X et al. (2005) Interaction network containing conserved and essential protein complexes in *Escherichia coli*. *Nature* **433:** 531-537.

7. Cai L, Friedman N and Xie XS (2006) Stochastic protein expression in individual cells at the single molecule level. *Nature* **440:** 358-362.

8. Champion MM, Campbell CS, Siegele DA, Russell DH and Hu JC (2003) Proteome analysis of *Escherichia coli* K-12 by two-dimensional native-state chromatography and MALDI-MS. *Mol Microbiol* **47:** 383-396.

9. Corbin RW, Paliy O, Yang F, Shabanowitz J, Platt M et al. (2003) Toward a protein profile of *Escherichia coli*: Comparison to its transcription profile. *Proc Natl Acad Sci U S A* **100:** 9232-9237.

10. Courcelle J, Khodursky A, Peter B, Brown PO and Hanawalt PC (2001) Comparative Gene Expression Profiles Following UV Exposure in Wild-type and SOS-deficient *Escherichia coli. Genetics* **158:** 41-64.

11. Dankelmann P, Goddard W and Swart CS (2004) The Average Eccentricity of a Graph and its Subgraphs. *Utilitas Math* **65:** 41-52.

12. Demeter J, Beauheim C, Gollub J, Hernandez-Boussard T, Jin H et al (2007) The Stanford Microarray Database: implementation of new analysis tools and open source release of software. *Nucleic Acids Res* **35:** D766-770.

13. Diez A, Gustavsson N and Nystrom T (2000) The universal stress protein A of *Escherichia coli* is required for resistance to DNA damaging agents and is regulated by a RecA/FtsK-dependent regulatory pathway. *Mol Microbiol* **36:** 1494-1503.

14. Dijkstra EW (1959) A note on two problems in connexion with graphs. *Numer Math* **1:** 263-271.

15. Dorogovtsev SN and Mendes JF (2003) Evolution of Networks: From Biological Nets to Internet and WWW, Oxford University Press, Oxford.

16. Elowitz MB, Levine AJ, Siggia ED and Swain PS (2002) Stochastic Gene Expression in a Single Cell. *Science* **297:** 1183-1186.

17. Farr SB, Arnosti DN, Chamberlin MJ and Ames BN (1989) An apaH mutation causes AppppA to accumulate and affects motility and catabolite repression in *Escherichia coli*. *Proc Natl Acad Sci U S A* **86:** 5010-5014.

18. Fraser HB, Hirsh AE, Giaever G, Kumm J and Eisen MB (2004) Noise Minimization in Eukaryotic Gene Expression. *PloS Biol* **2:** 834-838

19. Hahn MW and Kern AD (2005) Comparative Genomics of Centrality and Essentiality in Three Eukaryotic Protein-Interaction Networks. *Mol Biol Evol* **22:** 803-806.

20. Han JD, Bertin N, Hao T, Goldberg DS, Berriz GF et al. (2004) Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature* **430:** 88-93.

21. Horvath S and Dong J (2008) Geometric interpretation of gene co-expression network analysis. *PLoS Comput Biol* **4:** 1- 27.

22. Jeong H, Mason SP, Barabasi AL and Oltvai ZN (2001) Lethality and centrality in protein networks. *Nature* **411:** 41-42.

23. Johnstone DB and Farr SB (1991) AppppA binds to several proteins in *Escherichia coli*, including the heat shock and oxidative stress proteins DnaK, GroEL, E89, C45 and C40. *EMBO* **10:** 3897-3904.

24. Joy MP, Brock A, Ingber DE and Huang S (2005) High-Betweenness proteins in the Yeast Protein Interaction network. *J Biomed Biotechnol* **2:** 96-103.

25. Kanehisa M and Goto S (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* **28:** 27-30.

26. Komurov K and White M (2007) Revealing static and dynamic modular architecture of the eukaryotic protein interaction network. *Mol Syst Biol* **3:** 110.

27. Latora V and Marchiori M (2004) How the science of complex networks can help developing strategies against terrorism. *Chaos, Solitons and Fractals* **20:** 69-75.

28. Lu X, Jain VV, Finn PW and Perkins DL (2007) Hubs in biological networks exhibit low changes in expression in experimental asthma. *Mol Syst Biol* **3:** 98.

29. Mata J and Bahler J (2003) Correlation Between Gene expression and Gene conservation in Fission yeast. *Genome Res* **13:** 2686-2690.

30. Murphy D (2002) Gene expression studies using microarrays: Principles, problems and prospects. *Advan Physiol Educ* **26:** 256-270.

31. Nishimura A, Moriya S, Ukai H, Nagai K, Wachi M et al. (1997) Diadenosine 5',5'''-P1,P4-tetraphosphate (Ap4A) controls the timing of cell division in *Escherichia coli*. *Genes Cells* **2:** 401-413.

32. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT et al. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13:** 2498-2504.

33. Slonim DK (2002) From patterns to pathways: gene expression data analysis comes of age. *Nat Genet* **32:** 502-508.

34. Song C, Havlin S and Makse HA (2005) Self-similarity of complex networks. *Nature* **433:** 392-395.

35. von Mering C, Krause R, Snel B, Cornell M, Oliver SG et al. (2002) Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **417:** 399-403.

36. Wei Y, Lee JM, Richmond C, Blattner FR, Rafalski JA et al. (2001) High-Density Microarray mediated gene expression profiling of *Escherichia coli*. *J Bacteriol* **183:** 545-556.

37. Xue H, Xian B, Dong D, Xia K, Zhu S et al. (2007) A modular network model of aging. *Mol Syst Biol* **3:** 147.

38. Yellaboina S, Goyal K and Mande SC (2007) Inferring genome-wide functional linkages in *E. coli* by combining improved genome context methods: Comparison with high-throughput experimental data. *Genome Res* **17:** 527-535.

39. Yu H, Kim PM, Sprecher E, Trifonov V and Gerstein M (2007) The Importance of Bottlenecks in Protein Networks: Correlation with Gene Essentiality and Expression Dynamics. *PLoS Comput Biol* **3:** 713-720.

Chapter 3

Large Scale Analysis of the Gene Expression datasets of *Escherichia coli*

3.1 Introduction

Escherichia coli serves as a simple and widely studied prokaryotic model organism. The genome of *E. coli* K12, which is about 4.6 million base pairs, is capable of coding for 4288 proteins [Blattner et al. 1997]. Much of our understanding about prokaryotic biology including protein functions, cellular pathways and genetic modifications is derived from the studies in *E. coli*. In this regard, substantial amount of information is available on individual molecules as well as their coordinated functions. Therefore, *E. coli* is a potential model for systems level analysis.

With genome-sequences in hand, it is relevant to address how gene expression is regulated in response to a particular growth condition. To qualitatively decipher the circuitry of apparently complex gene regulation, it is requisite to estimate the expression of genes in an organism. Microarray is one such technique that quantitatively describes gene expression [Lockhart et al. 1996]. It allows for the global measurement of mRNA transcripts in a cell. With technological developments, introduction of novel algorithms for data analysis and the availability of tools and software, microarray technique has found immense application in biological research [Stoughton 2005; Miller and Tang 2009]. Organized public databases thus became inevitable to accommodate increasing amount of expression datasets in number of organisms [Brazma 2000]. In this regard, databases such as NCBI-Geo [Barrett et al. 2007], MMMD [Faith et al. 2007] and ArrayExpress [Brazma et al. 2003] function as repositories for individual experiments carried out across laboratories. These databases facilitate a user to access data in large scale and perform genome-wide studies.

In order to investigate the characteristics of gene expression and gene-gene association, RMA normalized affymetrix expression data was analyzed which was obtained from Many Microbe Microarray database for *E. coli* [Faith et al. 2008]. Furthermore, systems studies in *E. coli* by integrating expression correlations calculated using gene expression data with other high-throughput studies such as mRNA half lives, gene deletion experiments, protein functional linkages and biochemical pathways were performed. Biological insights about coordinated

regulation in expression and associated properties of coregulated gene pairs were studied. Results of these analyses are presented in this chapter.

3.2 Results

3.2.1 Profiling of *E. coli* genes

It is observed earlier that not all genes are expressed in a given condition in an organism [Champion et al. 2003]. Genes coding for proteins that perform basic cellular functions are invariably expressed in all the conditions, and are therefore termed essential genes. In addition, condition specific cellular processes are turned on based on the expression of conditionally essential genes. The other class of genes which are not expressed in most of the conditions is termed non-essential, and they impart redundancy to the system. While experimental profiling of the genes has been carried out previously [Baba et al. 2006; Posfai et al. 2006], the availability of large-scale gene expression data allows one to perform such studies in a faster and less expensive manner.

In order to categorize the genes of *E. coli* based on expression, publicly available gene expression data was used. Uniformly normalized microarray data for 4297 genes in 466 unique growth conditions was considered for the analysis. A gene was scored for its presence depending on the expression intensity in each condition (Methods). The profile thus obtained was tested for the known essential and non-essential genes in *E. coli*. Intuitively, essential genes show expression in about 88% of the conditions and non-essential genes are expressed in about 29% of the conditions. A similar test was done for the hub proteins which are defined as the highly connected proteins in an interaction network. Previous studies indicate a correlation between essentiality and higher connectivity for proteins in the interaction network [Jeong et al. 2001]. Interestingly, in the analysis it was observed that hubs are expressed in 78% conditions. Therefore, based on the number of conditions a gene gets expressed, *E. coli* genes were profiled into three categories: genes that are expressed in majority of the conditions, genes that are expressed only under a few conditions and genes that are dispensable for cell survival. These three classes are named as Widely expressed, Conditionally expressed and Less expressed respectively (Table S1 in Appendix II).

Further interest was to find out whether genes that are expressed in most of the conditions are also expressed at higher levels. When average expression intensities of the genes from these three classes were tested, it was observed that widely expressed genes indeed have higher level of expression intensity followed by conditionally expressed and less expressed genes (Figure 1(a)). To confirm this observation, an independent dataset was used for the analysis wherein gene expression is measured in both LB media and M9 media [Bernstein et al. 2002]. It is clear from Figure 1(b) that widely expressed genes are also expressed at higher levels compared to conditionally and less expressed genes.

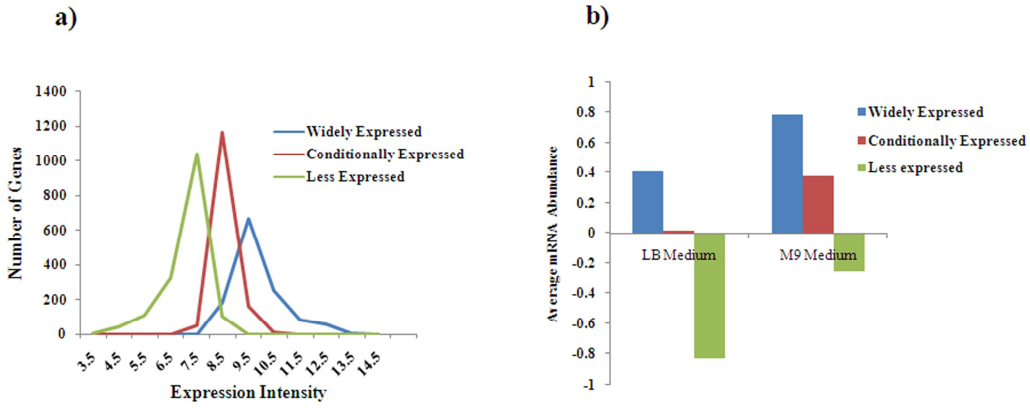

Figure 1: Expression levels for different classes of genes. Widely expressed genes are expressed at higher intensities compared to conditionally expressed and less expressed classes. a) MMMD expression data (Faith et al. 2008) b) Bernstein et al (2002) data.

When phyletic retention was studied for these classes, widely expressed genes were more conserved across genomes compared to conditionally expressed and less expressed classes (Figure 2(a)). In addition, widely expressed class was enriched for the orthologs of *Mycoplasma genitalium* (P-value < 5.09e-0122) suggesting that it consists of proteins from the conserved pathways. About 50% of the hubs are found to be present in the widely expressed class, reinforcing their essential functions (Figure 2(a)). Interestingly, 'Conditionally expressed' class showed higher expression variance suggesting its growth-dependent expression (Figure 2(b)).

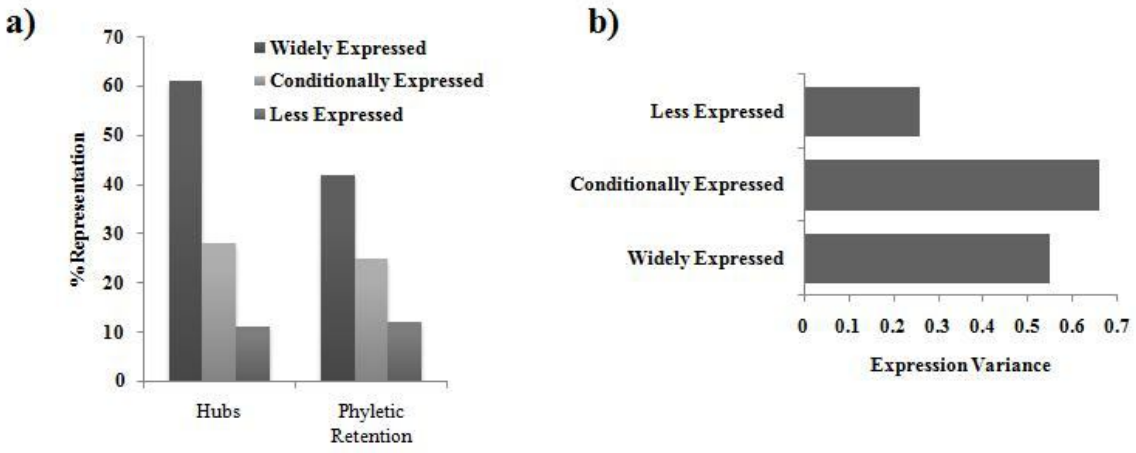

Figure 2: Properties of gene classes. a) Widely expressed gene class is enriched for hubs and is conserved across genomes. b) Conditionally expressed genes show higher expression variance compared to widely expressed and less expressed genes.

In order to test whether the stability of the transcripts from these classes differ, mRNA half life measurement data in both LB as well as M9 media were considered [Bernsetin et al. 2002]. Notably, genes of the less Expressed class code for more stable transcripts compared to the other two classes (Figure 3). It appears that though the essential genes are transcribed in large amount, their transcripts are degraded faster, suggesting a faster cellular response in transcription and their tighter regulation.

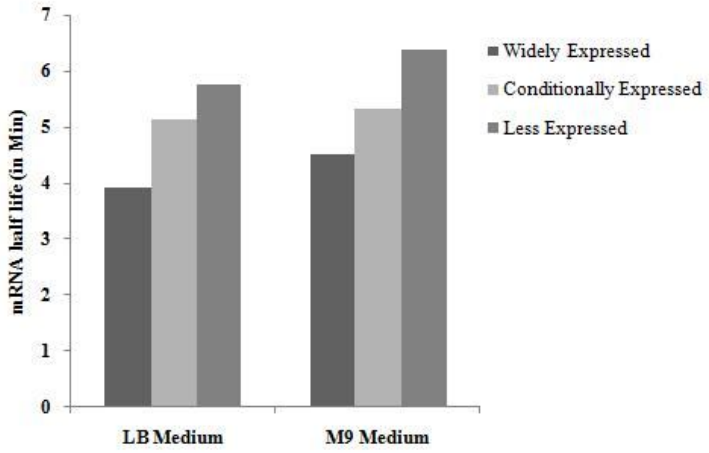

Figure 3: Gene expression and transcript stability. mRNAs coded by less expressed genes are relatively stable compared to the ones coded by widely expressed genes.

In order to understand the role of proteins from these three classes, centrality measures in the protein functional linkages were calculated [Yellaboina et al. 2007]. The proteins coded by widely expressed genes possess high degree as well as high betweenness centrality followed by conditionally expressed and less expressed classes (Figure 4). This implies that widely expressed genes form the backbone of a functional interaction network and play a critical role in information transfer. The genes from conditionally expressed class might temporally connect to this core of interacting proteins. Less expressed genes, on the other hand, have fewer connections and do not seem to play any significant role in communication within the network.

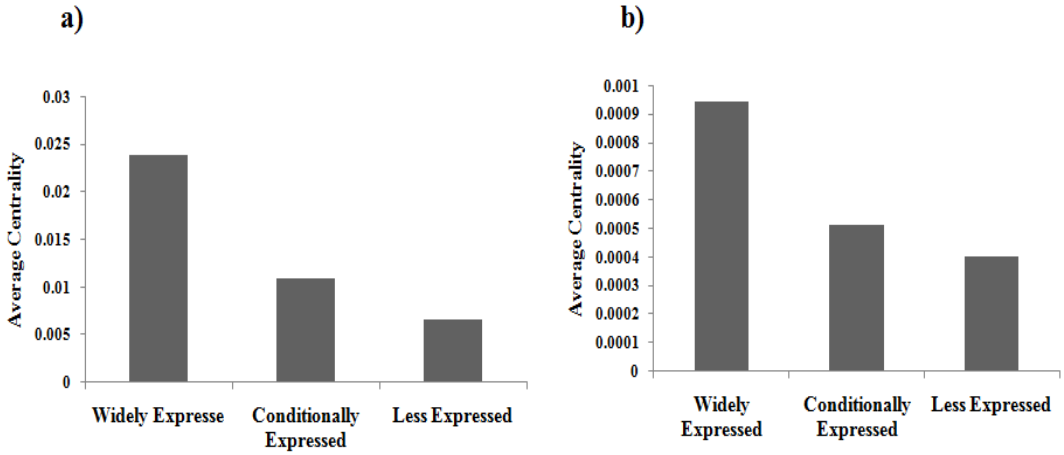

Figure 4: Network centrality and gene expression. Genes from the widely expressed class have higher centrality values. a) degree centrality and b) betweenness centrality.

Furthermore, metabolic pathway representation of genes from these three classes was examined. Pathways such as amino acid metabolism (P-value < $2.2e^{-16}$), nucleotide metabolism (P-value < 0.0011), transcription (P-value < 0.0005) and translation (P-value < $2.2e^{-16}$) are enriched for widely expressed genes. On the other hand, genes from conditionally expressed class are present in higher proportion in cell motility (P-value < $2.2e^{-16}$) and polyketide metabolism pathways (P-value < 0.04). Pathway enrichment therefore suggests essential cellular functions performed by widely expressed gens.

3.2.2 Gene Expression Correlations

To study co-regulation of gene pairs, genome-wide Pearson correlation coefficient was calculated based on gene expression data. Interestingly, most of the gene pairs are not correlated in their expression (Figure 5(a)). In order to assess the quality of the data, the distribution of operonic expression correlations was compared with randomly generated gene pairs. As shown in Figure 5(b), operonic gene pairs are significantly correlated in their expression compared to random gene pairs.

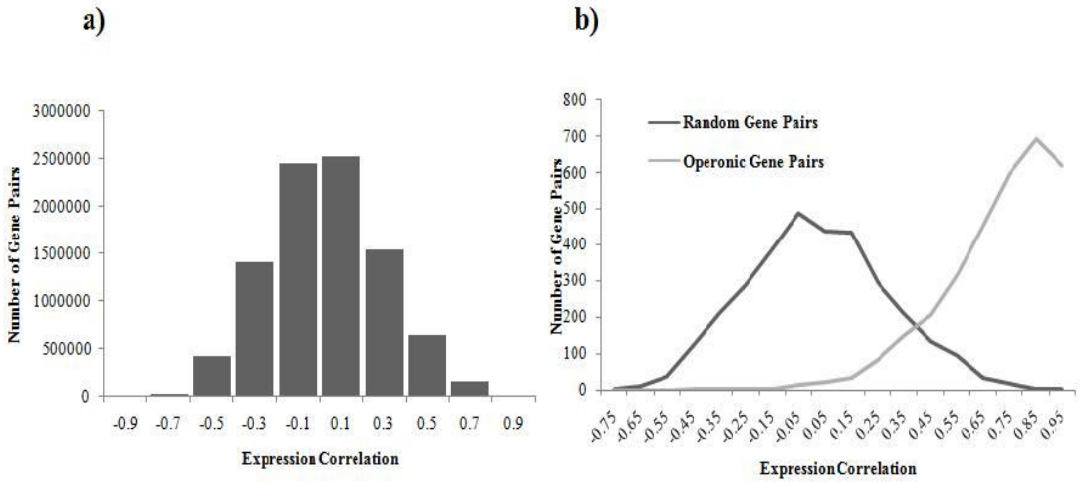

Figure 5: Expression correlations. a) Distribution of genome-wide gene expression correlations suggests that large fraction of gene pairs are not correlated in expression. b) Operonic gene pairs are highly correlated in expression. Equal number of randomly generated gene pairs' correlation is plotted as control.

A similar test was done for the protein functional linkages predicted based on combined genome-context methods [Yellaboina et al. 2007]. It is evident from Figure S1 in Appendix II that functionally linked proteins have higher expression correlation. This property is indeed used in predicting protein functional linkages in diverse organisms [Bhardwaj and Lu 2005; Hegde et al. unpublished]. This relation emphasizes the fact that a similar expression profile ensures timely interaction between proteins.

Even though large fraction of the genome-wide gene pairs do not correlate in expression, there are 174611 and 33014 gene pairs which have expression correlation better than 0.6 (highly correlated) and that less than -0.6 (less correlated) respectively. A strict regulation exhibited by correlated gene pairs, which may not fall in the same

operon, might indicate higher order of gene regulation such as regulons and modulons. However, the biological significance of the gene pairs that anticorrelate in expression is not clear.

Next, it was interesting to test whether correlation or inverse correlation of these gene pairs has any biological significance. Two proteins are likely to be proximal in the landscape of protein interaction network if they perform similar function [Zhou et al. 2002]. In a network, the distance between proteins is measured typically in terms of the length of shortest paths between them. In order to assess the functional relationship between coregulated gene pairs, shortest paths between correlated and inversely correlated gene pairs in the functional linkages network [Yellaboina et al. 2007]. The distribution of shortest paths indicates that the correlated gene pairs are positioned closer on the network compared to inversely correlated gene pairs (P-value < $2.2e^{-16}$, Figure 6), suggesting a coordinated regulation of functionally linked gene pairs.

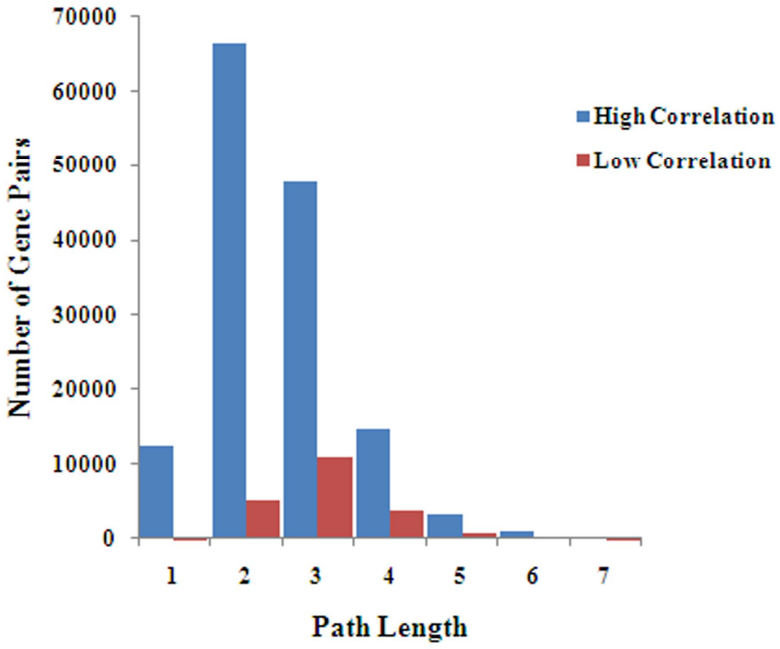

Figure 6: Expression correlation and Network pathlength. Highly correlated gene pairs are located closer in the interaction map compared to less correlated gene pairs, implying their functional association.

It was interesting to test whether correlated gene pairs also have similar stability at their mRNA level. This assumption was examined using available mRNA half life data. Studies on mRNA half lives revealed direction dependent degradation of transcripts and variable half lives of the transcripts of different operons in *E. coli* [Bernstein et al. 2002; Selinger et al. 2003]. To compare mRNA half lives of correlated and inversely correlated gene pairs, assortativity index was measured for each coregulated gene pair. Notably, correlated gene pairs show positive assortative value indicating similarity in their mRNA half lives. On the other hand, gene pairs that are anti-correlated show disassortativity implying that they need not have similar mRNA half life (Figure 7).

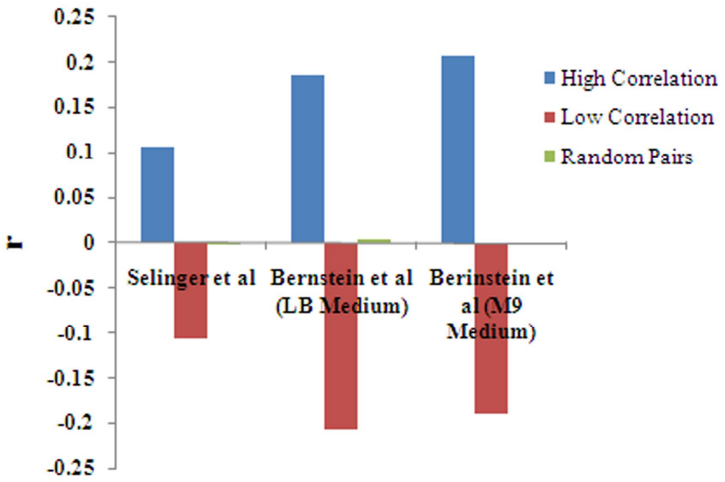

Figure 7: Expression correlation and mRNA half life. Highly correlated gene pairs exhibit assortative behaviour with respect to their mRNA half lives. However, gene pairs that are less correlated are disassortative in terms of their mRNA half lives.

3.2.3 Nature of Anticorrelation

Genes that are highly correlated in expression share a functional pathway, or are required for the temporal responses in an organism. On the other hand, anticorrelated gene pairs might indicate mutually exclusive functions. In the expression dataset analyzed for *E. coli*, there are 33014 gene pairs with expression correlation less than -0.6. It was intriguing to find such a large number of gene pairs

that exhibit strict anticorrelation in expression. However, it is difficult to arrive at a straightforward biological reasoning for such a behavior of gene expression.

A logical assumption to probe for biologically meaningful associations of proteins is that genes that anticorrelate with number of other genes are likely to be linked functionally. Such functionally linked gene pairs anticorrelate in expression with a common set of other genes. This feature termed congruency is indeed applied to genetic interaction network to predict protein-protein interactions [Tong et al. 2004]. In order to determine the functional associations of proteins using anticorrelated gene expression pattern, the number of shared anticorrelated genes for each gene pair was calculated. With the threshold of 10 shared anticorrelated genes for a gene pair, 101797 such protein linkages were obtained. Interestingly, the network thus derived clusters into two large disconnected modules, with 583 (Module1) and 406 (Module 2) proteins in each module. In addition, a small cluster consisting of five genes namely b3893, b3617, b2209, b4443 and b3894 is observed, which were ignored for further analysis (Figure 8).

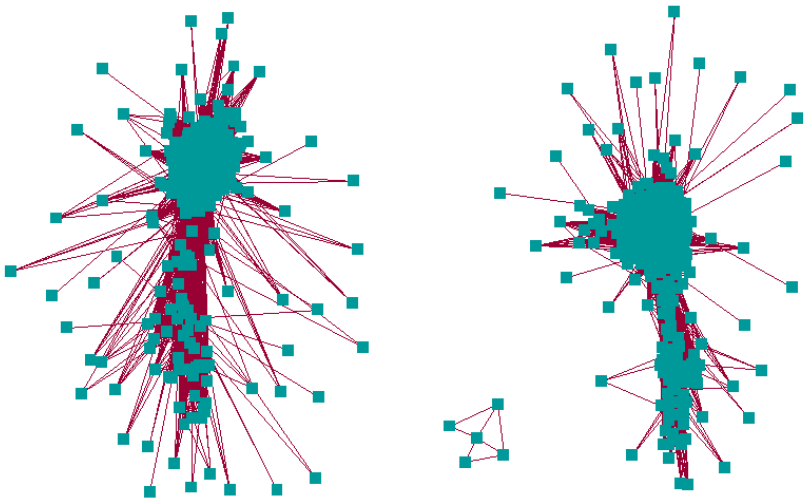

Figure 8: Anticorrelated modules in *E. coli.* Illustration of the two large modules of genes that are identified sing congruency of anticorrelated gene pairs.

Furthermore, characterization of the two modules reveals that Module 1 is enriched for essential genes (169 of the 299 essential genes from KEIO data) and hubs.

Subsequent pathway analysis indicated the enrichment of Module 1 for pathways such as transcription, translation, replication and nucleotide metabolism. On the other hand, the Module 2 is enriched for the genes from xenobiotic metabolism (Table S2 in Appendix II).

Notably, genes within the two modules highly correlate in expression and the genes between the modules anticorrelate in expression. This suggests a tight regulation between the expressions of two sets of genes in *E. coli*, the expression of one set denoting the higher probability of repression of the other set of genes. In order to test for the conservation of such a regulatory pattern in other organisms, gene expression in *Shewanella oneidensis* was considered for which the compiled expression data was available in MMMD database. In addition, orthologs of *S. oneidensis* in the two anticorrelated modules of *E. coli* were identified using bi-directional blast approach. It is observed that about 78% of genes from module 1 and 14% of module 2 are conserved in *S. oneidensis.* Subsequently, expression correlations for identified orthologs in *S. oneidensis* were compared with that of the *E. coli* genes. Though the orthologs of Module 1 genes are correlated in their expression in *S. oneidensis* as well, high correlation between genes of Module 2 as seen in *E. coli* is not observed for *S. oneidensis.* In addition, anticorrelation of Module 1 and Module 2, as observed in *E. coli*, is not conserved for the orthologs in *S. oneidensis* (Figure S2 in Appendix II). It appears that the regulatory mechanism for the identified modules is unique to *E. coli* and not conserved in *S. oneidensis*.

3.3 Discussion

Systems level analyses of *E. coli* gene expression were performed by coupling available microarray data with protein interaction networks, mRNA half life and metabolic pathways. The genes of *E. coli* can be profiled into three classes depending on their expression. The class 'Widely Expressed' is enriched for hubs and essential genes, and is highly conserved across genomes. 'Conditionally Expressed' genes are responsive against growth requirements and therefore have higher expression variance. The class 'Less Expressed' is less conserved and codes for comparatively stable transcripts. By calculating assortativity, it was shown that the gene pairs which

have correlate in expression also show similarity in their mRNA half lives. Using anticorrelated gene pairs, two anticorrelated gene modules were identified in *E. coli*. Interestingly, the orthologues of these modules in *S. oneidensis* do not show a similar regulatory pattern indicating organism specific regulatory mechanisms.

The derivatives such as expression correlations from gene expression datasets provide useful information about gene-gene relationships. An interesting aspect in this regard is to understand pathways that are co-regulated. Expression data can also be applied to understand higher orders of gene regulation. Gene expression dynamics can be translated into conditional/context dependent protein interactions which provide useful insights into temporal responses of an organism depending on growth environment.

3.4 Methods

3.4.1 Gene Expression Data

Expression data was downloaded from Many Microbe Microarray Database [http://m3d.bu.edu/, Faith et al. 2008] which consists of expression information for 4297 genes of *E. coli* in 466 growth conditions.

In order to determine whether a gene is expressed in a given condition, the median was calculated for the distribution of expression intensities of all the genes in the condition. A gene i with expression intensity X_i is considered expressed in condition j if $X_i > Median_j$ [Hegde et al. 2008]. Using this criterion, a binary profile denoting the presence or absence of the genes of *E. coli* across 466 growth conditions was constructed. Essential genes and non-essential genes of *E. coli* were obtained from KEIO collection [Baba et al. 2006] and Posfai et al. [2006] respectively. The average number of conditions in which essential or non-essential genes are expressed is 89% and 29% respectively. A gene is classified as 'Widely expressed' if it is expressed in more than 88% conditions, 'Less expressed' if the expression is in less than 29% conditions and 'Conditionally expressed' otherwise.

In the functional linkages network predicted using genome-context methods [Yellaboina et al. 2007], top 30% high degree nodes are defined as hub proteins. Phyletic retention was calculated by bi-directional blast of *E. coli* protein sequences

against 362 bacterial genomes with e-value cutoff of e^{-04}. The data for mRNA half lives were obtained from Bernstein et al. (2002) and Selinger et al. (2003). Orthologs of *Mycoplasma genitalium* were identified using bi-directional blast with e-value cutoff of e^{-04}. Network centrality measures were calculated according to [Manimaran et al. 2009]. Pathway classification for *E. coli* genes were downloaded from KEGG database [Kanehisa and Goto 2000].

Similar analysis of expression correlation as that in *E. coli* was performed in the microbe *Shewanella oneidensis*. Gene expression in *Shewanella oneidensis* for 207 growth conditions was downloaded from M3D database. Using bi-directional blast with e-value cutoff of e^{-04}, 1779 orthologs of *E. coli* proteins were identified in *S. oneidensis.*

3.4.2 Assortativity

Assortativity between gene expression correlations and mRNA half lives were calculated using the formula [Newman 2002]:

$$r = \frac{\left(\frac{1}{M}\right)\sum_{i=1}^{M} j_i k_i - \left[\frac{1}{M}\sum_{i=1}^{M}\frac{1}{2}(j_i + k_i)\right]^2}{\left[\frac{1}{M}\sum_{i=1}^{M}\frac{1}{2}(j^2{}_i + k^2{}_i)\right] - \left[\frac{1}{M}\sum_{i=1}^{M}\frac{1}{2}(j_i + k_i)\right]^2}$$

3.5 References

1. Baba T, Ara T, Hasegawa M, Takai Y, Okumura Yet al. (2006) Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol*. **2:** 2006.0008.

2. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D et al. (2007) NCBI GEO: mining tens of millions of expression profiles--database and tools update. *Nucleic Acids Res*. **35:** D760-D765.

3. Bernstein JA, Khodursky AB, Lin PH, Lin-Chao S and Cohen SN (2002) Global analysis of mRNA decay and abundance in *Escherichia coli* at single-gene resolution using two-color fluorescent DNA microarrays. *Proc Natl Acad Sci U S A*. **99:** 9697-9702.

4. Bhardwaj N and Lu H (2005) Correlation between gene expression profiles and protein-protein interactions within and across genomes. *Bioinformatics.* **21:** 2730-2738.

5. Blattner FR, Plunkett G 3rd, Bloch CA, Perna NT, Burland V et al. (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*. **277:** 1453-1462.

6. Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J et al. (2003) ArrayExpress--a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.* **31:** 68-71.

7. Brazma A, Robinson A, Cameron G and Ashburner M (2000) One-stop shop for microarray data. *Nature*. **403:** 699-700.

8. Champion MM, Campbell CS, Siegele DA, Russell DH and Hu JC (2003) Proteome analysis of *Escherichia coli* K-12 by two-dimensional native-state chromatography and MALDI-MS. *Mol Microbiol.* **47:** 383–396.

9. Faith JJ, Driscoll ME, Fusaro VA, Cosgrove EJ, Hayete B et al. (2008) Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental metadata. *Nucleic Acids Res.* **36:** D866-870.

10. Hegde SR, Manimaran P and Mande SC (2008) Dynamic changes in protein functional linkage networks revealed by integration with gene expression data. *PLoS Comput Biol*. **4:** e1000237.

11. Jeong H, Mason SP, Barabási AL and Oltvai ZN (2001) Lethality and centrality in protein networks. *Nature.* **411:** 41-42.

12. Kanehisa M and Goto S (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28:** 27–30.

13. Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV et al. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.* **14:** 1675-1680.

14. Manimaran P, Hegde SR and Mande SC (2009) Prediction of conditional gene essentiality through graph theoretical analysis of genome-wide functional linkages. *Mol Biosyst.* **5:** 1936-1942.

15. Miller MB and Tang YW (2009) Basic concepts of microarrays and potential applications in clinical microbiology. *Clin Microbiol Rev.* **22:** 611-633.

16. Newman MEJ (2002). Assortative Mixing in Networks. *Phys Rev Lett*. **89:** 208701.

17. Pósfai G, Plunkett G 3rd, Fehér T, Frisch D, Keil GM et al. (2006) Emergent properties of reduced-genome *Escherichia coli*. *Science.* **312:** 1044-1046.

18. Selinger DW, Saxena RM, Cheung KJ, Church GM and Rosenow C (2003) Global RNA half-life analysis in *Escherichia coli* reveals positional patterns of transcript degradation. *Genome Res.* **13:** 216-23.

19. Stoughton RB (2005) Applications of DNA microarrays in biology. *Annu Rev Biochem.* **74:** 53-82.

20. Yellaboina S, Goyal K and Mande SC (2007) Inferring genome-wide functional linkages in *E. coli* by combining improved genome context methods: Comparison with high-throughput experimental data. *Genome Res* **17:** 527–535.

21. Zhou X, Kao MC and Wong WH (2002) Transitive functional annotation by shortest-path analysis of gene expression data. *Proc. Natl. Acad. Sci. U S A*, **99:** 12783-12788.

Chapter 4

Prediction of Genome-wide Protein Functional Linkages in *Mycobacterium tuberculosis*

4.1 Introduction

Increased rate of tuberculosis (TB) infection and the emergence of multi and extensively-drug-resistant tuberculosis [WHO Report 2010] have necessitated urgent efforts towards enhanced understanding of its causative agent, *Mycobacterium tuberculosis*. The complete genome sequence of the common laboratory strain *M. tuberculosis* H37Rv was unraveled in 1998 [Cole et al. 1998] and the sequence was later re-annotated [Camus et al. 2002]. The genome annotation combined with comparative genomic studies has revealed several novel features including gene families that are unique to this bacterium such as PE and PPE genes and the eukaryotic like serine threonine protein kinases (STPKs). The genome also harbors unique genes involved in lipid biosynthesis, drug resistance and pathogenesis [Cole et al. 1998].

Owing to its distinct characteristics and the disease causing ability, *M. tuberculosis* has been studied widely across the world. For example, different experiments have accumulated information about the biochemical and structural properties of a number of proteins of this organism [Smith 2003; Hett and Rubin 2008]. There is also large amount of data on mycobacterial gene expression under a variety of growth conditions [Barrett et al. 2009]. Genome-wide surveys of essential genes of *M. tuberculosis*, both *in vitro* and those potentially involved in pathogenesis, have been carried out [Sassetti et al. 2003; Sassetti and Rubin 2003; Rengarajan et al. 2005]. Similarly, large-scale proteome profiling to classify *M. tuberculosis* proteins into different cellular compartments has been carried out [Mawuenyega et al. 2005]. Such varied studies have helped in understanding not only the unique genetic makeup of the bacillus, but also the possible roles of individual genes during different steps of pathogenesis.

M. tuberculosis has also been the centre of attention for a few studies executed at the systems level. In this regard, there have been attempts to model gene regulation, protein interactions and metabolic pathways of the organism. Functional linkage maps of *M. tuberculosis* have been defined by using genome context methods [Strong et al. 2003]. In another study, Balaszi et al have assembled gene regulation information and studied transcriptional changes that might mediate switch to

dormancy [Balázsi et al. 2008]. Flux balance analysis (FBA) on the model of mycolic acid biosynthesis pathway has revealed potential drug targets pertaining to this pathway [Raman et al. 2005]. Nonetheless, the systems level understanding of *M. tuberculosis* remains inadequate. One of the major obstacles being that a large fraction of genes are either putative or are unannotated, and thereby coherence among different pathways that contribute to virulence remains to be defined systematically. There is therefore a pressing need to integrate different approaches to understand tuberculosis and perceive mechanisms by which *M. tuberculosis* enters a dormant phase, or emerges out of it. A new promise in this aspect is the availability of a number of genome sequences of clinical strains and the data gathered by individual and high-throughput experiments which permit studies at the systems level.

One of the promising approaches to understand complex functional associations of the molecules and their organization is by analyzing genome-wide protein:protein interactions by means of graph theoretical representations. In this regard, there have been a few attempts to generate protein interaction maps of *M. tuberculosis*, all of them being *in silico* predictions. Strong et al used genome context methods to derive a functional linkage map of 4,886 interactions among 1,958 proteins followed by clustering of the network in order to reconstruct some of the biochemical pathways in *M. tuberculosis* [Strong et al. 2003]. Another study involved translating high confidence interactions of *E. coli* to *M. tuberculosis* [Cui et al. 2009]. A set of 6,091 interactions among 793 proteins was obtained in this study to identify proteins involved in signaling pathways. The interaction database STRING [Jensen et al. 2009] houses protein interactions of *M. tuberculosis* derived by literature curation and other methods. The highest confidence interactions in STRING which are supported by multiple methods or curation include around 6,403 interactions between 1,653 proteins. However, none of the above sources represents a comprehensive collection of *M. tuberculosis* protein interactions.

This study attempts to combine genome context methods, namely phylogenetic profile [Pellegrini et al. 1999; Enault et al. 2003], gene distance [Korbel et al. 2004] operonic co-occurrence [Dandekar et al. 1998; Overbeek et al. 1999] with the available high throughput gene expression studies for the prediction of genome-wide functional

linkages. The predictive features of these were combined using a Support Vector Machine (SVM). These functional linkages were analyzed in terms of graph theoretical properties and functional pathways.

4.2 Results

4.2.1 Generation of the Protein Functional Linkages

One of the studies earlier showed that the genome context methods, namely phylogenetic profile, operonic frequency and gene distance can be effectively combined using a SVM to predict protein functional linkages in *E. coli* [Yellaboina et al. 2007]. Apart from the genome context methods it is interesting to note that the proteins which have functional relations are also known to show good correlation in their gene expression [Bhardwaj and Lu 2005]. Therefore, correlation in gene expression among the genes is included as a feature for SVM training.

It is well known that large-scale gene expression datasets are expected to be noisy [Marshall 2004]. However, the expression levels between gene pairs, if functionally related, are likely to be correlated across different experimental conditions. In general, no gene pair is anticipated to exhibit high expression correlation except when the two genes are coregulated thereby suggesting a functional relationship between the two. This conjecture was sought to be tested using known operonic gene pairs. Significantly, the operonic gene pairs show high expression correlation compared to the same number of randomly generated non-operonic pairs (Figure 1). This observation, therefore, strengthened the confidence in using the microarray data for interaction predictions. Thus, genome context methods have been supplemented with correlations in gene expression to generate genome-wide protein functional linkages map of *M. tuberculosis*.

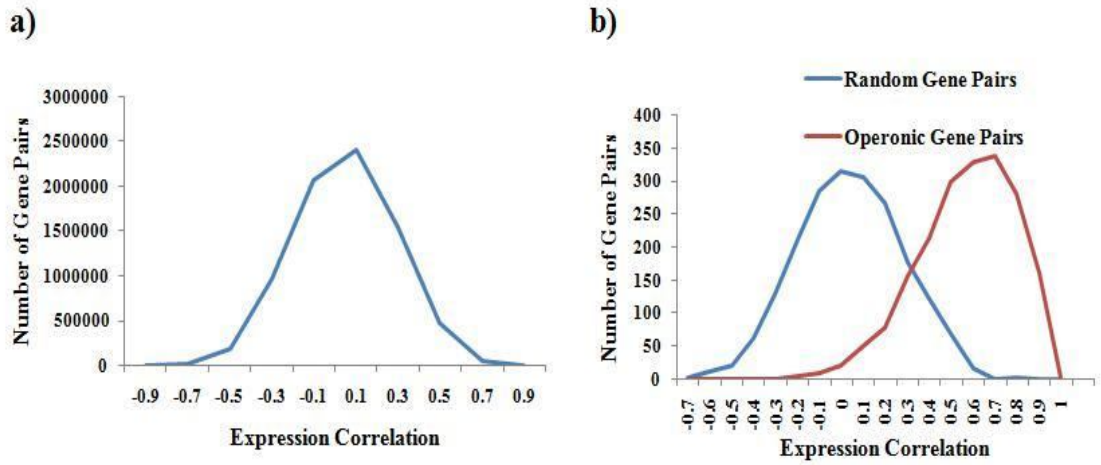

Figure 1: Pearson correlation coefficient between expression values of pairs of genes. a) The genome-wide gene expression correlations of all the gene pairs and b) expression correlation between operonic gene pairs and randomly paired non-operonic gene pairs. As anticipated, it is evident that the genes on operons exhibit higher correlation in their expression compared to the non-operonic gene pairs.

In order to test whether the positive and negative pairs used in training the SVM show characteristic distribution for the data features chosen, students' t-test was performed on the available positive pairs and equal number of randomly chosen negative pairs. Figure 2 depicts the distribution of vectors of the gene pairs used in training, which show a distinctly different distribution with a p value of $2.2e^{-16}$. All the data features chosen are therefore capable of distinguishing between positive and negative pairs, suggesting their potential application in the prediction of functional linkages. Thus, after optimizing training for the SVM using these features, prediction accuracy of 88% with 76% sensitivity was obtained. This optimized model was chosen for prediction of functional linkages on the genome-wide scale.

4.2.2 The Functional Linkages of *M. tuberculosis* Proteins

The predicted protein functional linkages network has 32,546 interactions among 3,571 proteins (Table S1 in Appendix III). The largest connected component of the network comprises of 95% of the nodes and has a diameter of 12. The network shows scale free property with the degree exponent of 1.67. The overall topological parameters of the network are summarized in Table 1.

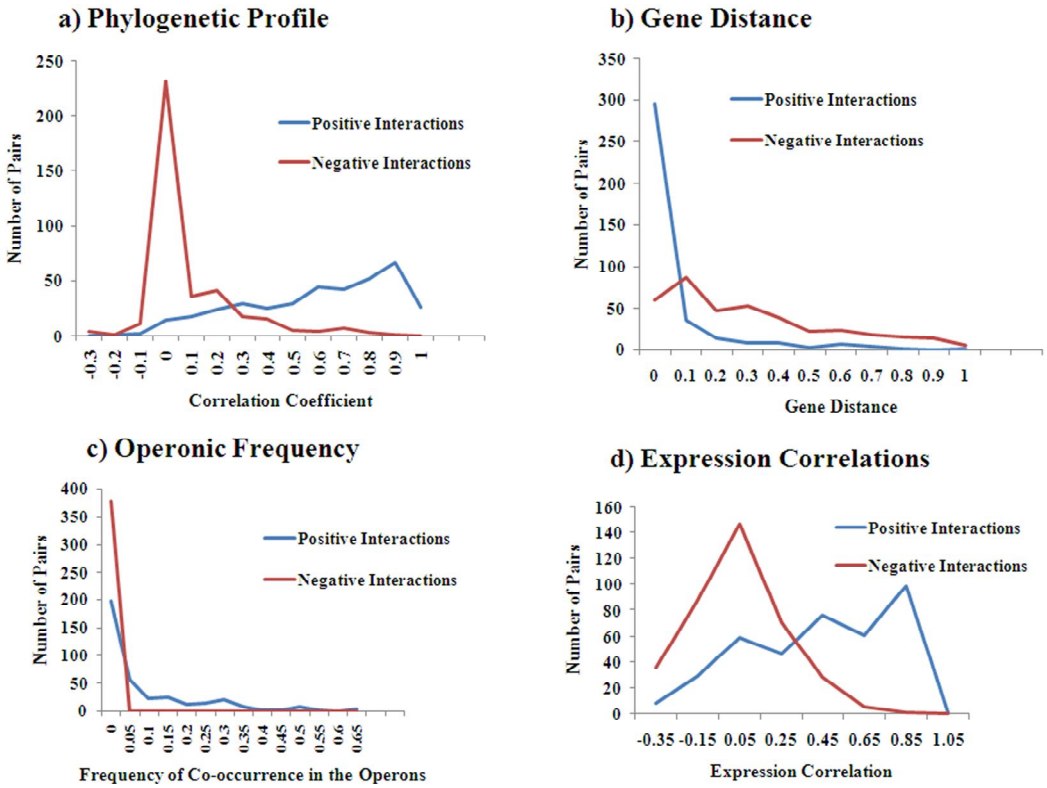

Figure 2: Plots indicating the distinctive behavior of the positive and negative protein interaction pairs with respect to genome context methods and gene expression correlations. All the four features show statistically significant and distinct distribution for the positive and negative pairs.

Number of Interactions	32546
Number of Nodes	3571
Percentile Core Nodes	95%
Average Degree	19.2
Degree Exponent	1.67
Diameter	12
Average Clustering Coefficient	0.22

Table 1: Topological Properties of the predicted protein functional linkages

Comparison of the network with previously derived interaction maps shows around 30% overlap (Table S2 in Appendix III). Though the overlap appears less, it is observed that protein interaction maps obtained from different sources generally have fewer interactions in common due to inherent bias in the method used or the noise [Yellaboina et al. 2007].

The average degree and clustering coefficient of the proteins classified into different metabolic pathways is shown in Figure 3. Proteins belonging to translation pathway, such as the ribosomal proteins and other translation related proteins, have high degree as well as high clustering coefficient. On the other hand, proteins of the "cell growth and death" pathway have high degree but are less clustered. The membrane transport proteins on the other hand show less degree but are highly clustered. Thus, varied relationship between clustering and degree in different metabolic pathways is apparent from the data.

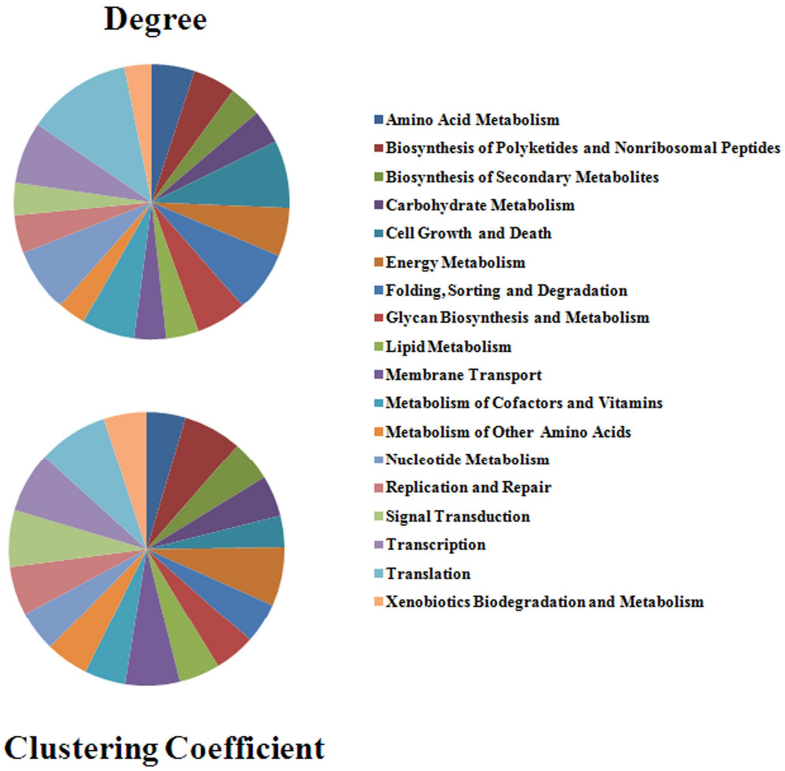

Figure 3: Average degree and clustering coefficient of the proteins across metabolic pathways. Proteins show high degree or high clustering coefficient depending on their function. Translation pathway proteins have high degree as well as high clustering coefficient whereas membrane transport proteins are highly clustered with less connectivity.

Proteins coded by essential genes in biological networks are known to exhibit high network centrality measures compared to their counterparts [Manimaran et al. 2009]. This observation was tested in *M. tuberculosis* network for the experimentally proposed essential genes [Sassetti et al. 2003]. From the derived functional interaction network, three centrality values, namely degree, closeness and betweenness, were calculated for each of the genes, and the genes were further divided into three categories: high centrality (top 30% nodes), medium centrality (between 30-70% nodes) and low centrality (others). The proportion of essential genes for these three centrality parameters was plotted as shown in Figure 4. It is evident from the figure that there is a strong correlation between network centrality and gene essentiality in the proposed network.

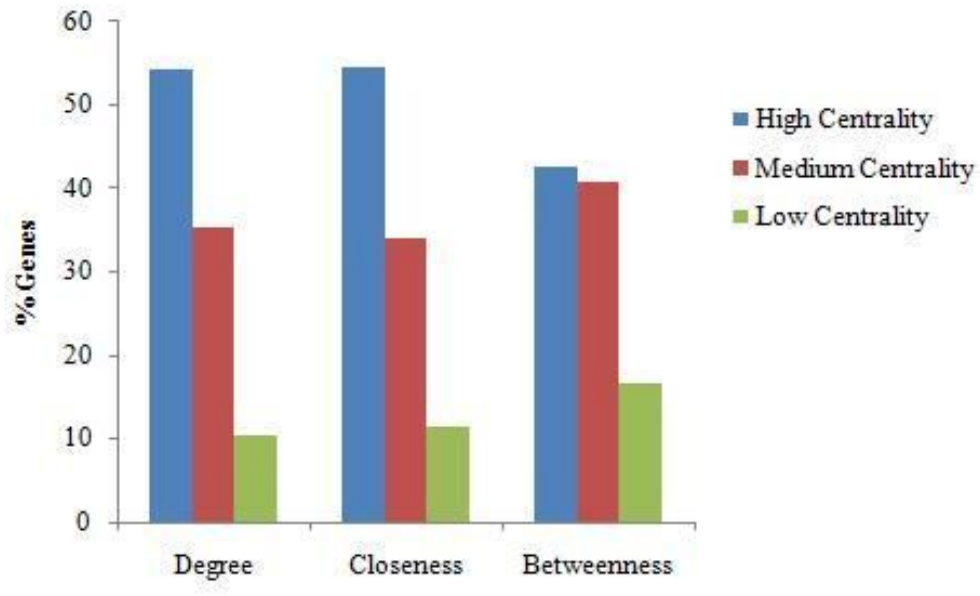

Figure 4: Proportion of Essential Genes in the bins of decreasing centrality values. For all the three centrality values calculated, namely degree, closeness and betweenness, it is clearly seen that there exists a good correlation between centrality and lethality.

The proteins of information pathway such as DnaG, DnaB, Rho and ribosomal proteins; proteins belonging to intermediary metabolism and respiration such as ATP synthase subunits, purine and pyrimidine biosynthesis proteins; proteins in the amino acid biosynthesis pathways and cell division proteins such as FtsX, FtsZ and FtsH;

and proteins of cell wall formation such as MurA, MurB, MurC and MurD show very high network centrality values. Interestingly, PPE4 and PPE46 of the PPE family proteins have high centrality values, suggesting that these proteins might also be constituents of the essential gene set of *M. tuberculosis*. The complete list of proteins with high centrality values is given in Table S3 in Appendix III.

Proteins involved in closely related functions, or those involved in the same biochemical pathway, are known to cluster together to form functional units in an interaction map [Rives and Galitski 2003]. Clustering of the network indicated that it is divided into 184 clusters which were further annotated using the pathway information taken from Sanger and TubercuList databases (Table S4 in Appendix III). The largest cluster is enriched with cell wall and cell processes, virulence related proteins and a large number of conserved hypothetical proteins. The ribosomal proteins and other translation related proteins cluster along with DNA replication and repair proteins. A large number of PE-PPE proteins were found to be associated with virulence proteins. It is likely therefore that these proteins play an important role in virulence determination, as has been suggested earlier [Cole et al. 1998; Banu et al. 2002]. Thus, the sub-division of the interaction map into clusters based on the nature of their interaction might lead to a better understanding of cross-talks between different functional categories.

4.2.3 Web User Interface for *M. tuberculosis* Functional Interactions

Predicted protein functional linkages are made available on the user interface at http://www.cdfd.org.in/MtbPPI/ wherein user can download the flat file for protein interactions. In addition, individual protein search option is given in which case all the interacting partners along with their functional annotation will be displayed on the screen. Option for expression correlation searches are also provided for a pair of gene in *M. tuberculosis.* All the associated data files for prediction such as the list of selected genomes and description of gene expression conditions can be retrieved from the site as supplementary files. Figure 5 is the snapshot of the interface.

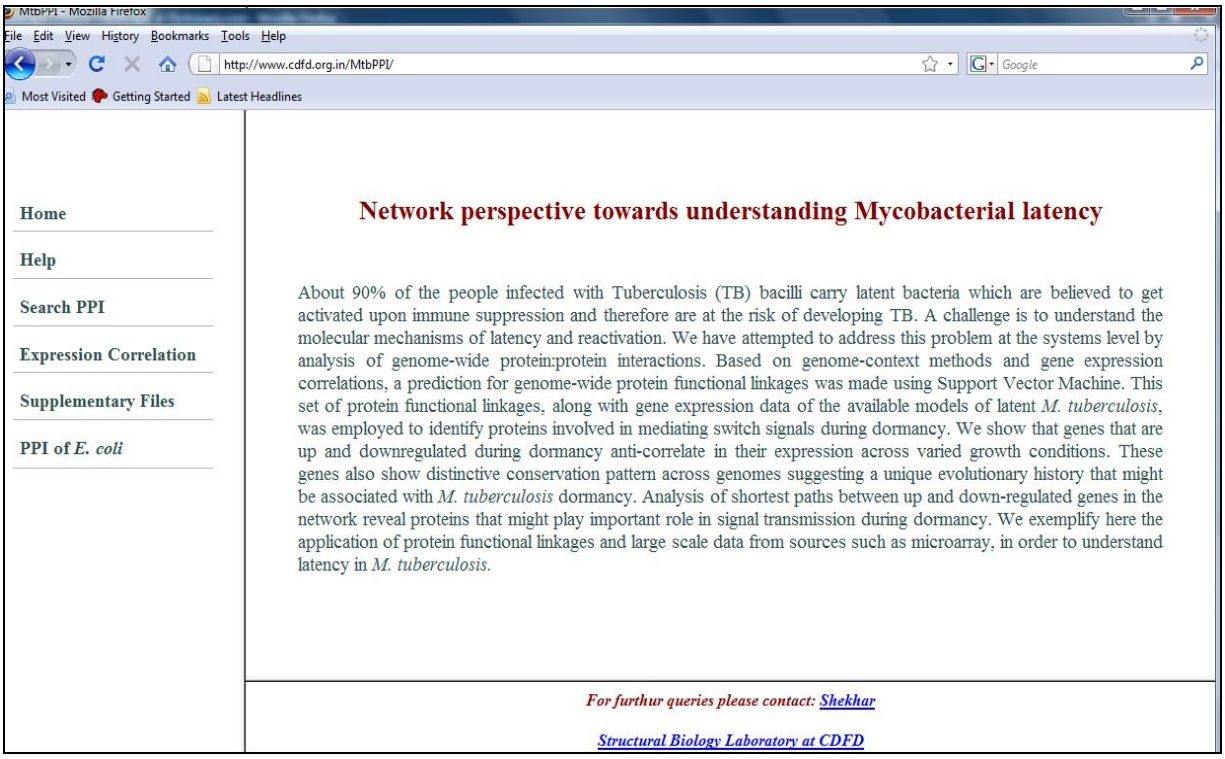

Figure 5: Snapshot of the interface for accessing protein interactions in *M. tuberculosis*. Search options for protein interactions and expression correlations can be made using *M. tuberculosis* protein names or the respective Rv numbers.

4.3 Discussion

M. tuberculosis causes tuberculosis which claims about two million deaths annually [WHO report, 2010]. Systems level approaches have thus become inevitable to understand the disease causing ability of *M. tuberculosis*. In this study, protein functional linkages for *M. tuberculosis* proteins were predicted using genome context methods namely: phylogenetic profile, gene distance, operonic frequency, along with gene expression correlations. All these features are combined using support vector machine. A genome-wide network of protein functional linkages was derived which showed scale-free property and small world behaviour. The centrality measures calculated on the network imply correlation between gene essentiality and higher network centrality. Furthermore, the network was clustered to identify functionally segregated protein modules. Predicted network, along with associated data files, is

made available on the web site. Such a network of protein interactions will hopefully be a useful resource for tuberculosis research and drug discovery.

4.4 Methods

4.4.1 Positive and Negative Interaction Pairs

The known interacting protein pairs were obtained by a combination of text mining methods and bi-directional blast against high confident protein interactions for *E. coli* listed in EcoCyc database [Keseler et al. 2009]. Those obtained by text mining were retrieved from literature by the use of natural language processing methods (Goyal and Mande, unpublished results). Bi-directional BLAST was carried out for each pair listed in the EcoCyc database against the *M. tuberculosis* H37Rv genome sequence. Only those pairs were further considered which showed a score better than e^{-10} for both the proteins.

The hypothesized non-interacting data set was obtained by the method described in [Yellaboina et al. 2007]. Briefly, the protein pairs which are not colocalised in the same subcellular compartment were considered to be non-interacting. The protein localization was predicted using the SIGCLEAVE tool available at http://mobyle.pasteur.fr/cgi-bin/Mobyle Portal. The top scoring 809 proteins with predicted secretary signal sequence within the first 50 residues of the N-terminus were considered to be extracellular. There are 161 proteins which do not possess any known signal sequence along their entire length and were considered to be cytoplasmic. Such negative interacting protein pairs were generated by randomly pairing the predicted cytoplasmic and extracellular proteins.

4.4.2 Selection of the Genomes

The sequences of 763 bacterial genomes were downloaded from NCBI ftp site (ftp://ncbi.nih.gov/genomes/Bacteria). In the initial filter, bacteria with linear or multiple genomes were removed. This resulted in a set of 669 genomes for further analysis. Homologous genes of all the known open reading frames of *M. tuberculosis* were searched against these 669 genomes using BLASTp with e-value cutoff of e^{-04}. For the species with complete genome sequences of more than one strain, the one

which shared maximum number of ORFs with *M. tuberculosis* was chosen. This resulted in a list of 481 genomes for further consideration (Table S5 in Appendix III).

4.4.3 Prediction Features

Phylogenetic Profile: BLAST with e-value cutoff of e^{-04} was used to obtain bit scores for the ORFs of *M. tuberculosis* against 481 selected genomes. The resulting profile was doubly normalized as in [Yellaboina et al. 2007]. Pearson Correlation Coefficient (PCC) was calculated for each gene pair and used as a feature for training the SVM.

Intergenic Distance: For using minimum gene distance as a training feature, top 100 organisms sharing maximum number of ORFs with *M. tuberculosis* were considered. For each genome, distances (in base pairs) of transcriptional start sites between all the gene pairs were calculated in both clockwise and anticlockwise directions and the minimum of these was normalized by the total genome length. A similar profile was constructed for *M. tuberculosis* genome as well. For each gene pair of *M. tuberculosis,* the minimum distance in its genome and its orthologs in other genomes was considered as a feature vector.

Frequency of co-occurrence in the predicted operons: Operon predictions for 267 organisms were obtained from [Yellaboina et al. 2007]. The frequency of co-occurrence of protein pairs as operonic across all the genomes was calculated.

Expression Correlations: Gene expression data for *M. tuberculosis* was downloaded from NCBI-Geo [Barrett et al. 2009]. The expression values for the multiple trails were normalized. The conditions which had expression variance of more than 5 were not considered for the analysis. The expression ratio in 154 growth conditions for each gene was compiled and Pearson Correlation Coefficient was derived for each gene pair. The list of selected conditions is detailed in Table S6 in Appendix III.

4.4.4 Protein Interactions Prediction

Support Vector Machine (SVM) tool LibSVM [Chih-Chung and Chih-Jen L 2001] was used for the prediction of genome-wide functional linkages. All possible gene pairs in *M. tuberculosis* carried the feature vector labels, namely correlation coefficient of the phylogenetic profile, minimum intergenic distance, frequency of co-

occurrence in predicted operons and expression correlations. The machine was trained on positive and negative data sets using these data features. Since the number of expected interacting pairs was likely to be much lower than the non-interacting pairs, the ratio of negative interacting pairs and the positive interacting pairs was increased for each trial. Each test included five fold cross validation followed by calculations of sensitivity and specificity. The interactions were predicted with each of the model files obtained and the final network was selected based on sensitivity, specificity and the accuracy of prediction. The final predictions are available on the web server (http://www.cdfd.org.in/MtbPPI/).

4.4.5 Network Analysis

Different topological parameters of the network that included degree exponent, clustering coefficient and diameter were calculated according to [Dorogovtsev and Mendes 2003]. Centrality measures were calculated as in [Manimaran et al. 2008]. Clusters in the network were detected using the Infomap tool [Rosvall and Bergstrom 2008]. Functional annotations of *M. tuberculosis* proteins were derived from KEGG [Kanehisa and Goto 2000], TubercuList (http://genolist.pasteur.fr/TubercuList/) and Sanger (http://www.sanger.ac.uk/) databases. Clustering coefficient or degree for a pathway is considered to be high if the average clustering coefficient or the average degree for the proteins in the pathway was more than the average of all the proteins.

4.5 References

1. Balázsi G, Heath AP, Shi L and Gennaro ML (2008) The temporal response of the *Mycobacterium tuberculosis* gene regulatory network during growth arrest. *Mol Syst Biol* **4:** 225.

2. Banu S, Honoré N, Saint-Joanis B, Philpott D, Prévost MC et al. (2002) Are the PE-PGRS proteins of *Mycobacterium tuberculosis* variable surface antigens? *Mol Microbiol*. **44:** 9-19.

3. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D et al. (2009) NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Res* **37:** D885-90.

4. Bhardwaj N and Lu H (2005) Correlation between gene expression profiles and protein-protein interactions within and across genomes. *Bioinformatics* **21:** 2730-2738.

5. Camus JC, Pryor MJ, Médigue C and Cole ST (2002) Re-annotation of the genome sequence of *Mycobacterium tuberculosis* H37Rv. *Microbiology* **148:** 2967–2973.

6. Chih-Chung C and Chih-Jen L (2001) LIBSVM: a library for support vector machines,. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

7. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C et al. (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393:** 537-544.

8. Cui T, Zhang L, Wang X and He ZG (2009) Uncovering new signaling proteins and potential drug targets through the interactome analysis of *Mycobacterium tuberculosis*. *BMC Genomics* **10:** 118.

9. Dandekar T, Snel B, Huynen M and Bork P (1998) Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci* **23:** 324-328.

10. Dorogovtsev SN and Mendes JF (2003) Evolution of Networks: From Biological Nets to Internet and WWW. Oxford: Oxford University Press.

11. Enault F, Suhre K, Abergel C, Poirot O and Claverie JM (2003) Annotation of bacterial genomes using improved phylogenomic profiles. *Bioinformatics* **19:** i105-i107.

12. Hett EC and Rubin EJ (2008) Bacterial growth and cell division: a mycobacterial perspective. *Microbiol Mol Biol Rev* **72:** 126-56.

13. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C et al. (2009) STRING 8—a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res* **37:** D412-D416.

14. Kanehisa M and Goto S (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* **28:** 27-30.

15. Keseler IM, Bonavides-Martínez C, Collado-Vides J, Gama-Castro S, Gunsalus RP *et al.* (2009) EcoCyc: a comprehensive view of *Escherichia coli* biology. *Nucleic Acids Res* **37:** D464-D470.

16. Korbel JO, Jensen LJ, von Mering C and Bork P (2004) Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs. *Nat Biotechnol* **22:** 911-917.

17. Manimaran P, Hegde SR and Mande SC (2009) Prediction of conditional gene essentiality through graph theoretical analysis of genome-wide functional linkages. *Mol Biosyst* **5:** 1936-1942.

18. Marshall E (2004) Getting the Noise Out of Gene Arrays. *Science* **306:** 630-631.

19. Mawuenyega KG, Forst CV, Dobos KM, Belisle JT, Chen J et al. (2005) *Mycobacterium tuberculosis* functional network analysis by global subcellular protein profiling. *Mol Biol Cell* **16:** 396-404.

20. Overbeek R, Fonstein M, D'Souza M, Pusch GD and Maltsev N (1999) The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci USA* **96:** 2896-2901.

21. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D and Yeates TO (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci USA* **96:** 4285-4288.

22. Raman K, Rajagopalan P and Chandra N (2005) Flux balance analysis of mycolic acid pathway: targets for anti-tubercular drugs. *PLoS Comput Biol* **1:** e46.

23. Rengarajan J, Bloom BR and Rubin EJ (2005) Genome-wide requirements for *Mycobacterium tuberculosis* adaptation and survival in macrophages. *Proc Natl Acad Sci U S A* **102:** 8327-8332.

24. Rives AW and Galitski T (2003) Modular organization of cellular networks. *Proc Natl Acad Sci U S A* **100:** 1128-1133.

25. Rosvall M and Bergstrom CT (2008) Maps of random walks on complex networks reveal community structure. *Proc Natl Acad Sci U S A* **105:** 1118-1123.

26. Sassetti CM, Boyd DH and Rubin EJ (2003) Genes required for mycobacterial growth defined by high density mutagenesis. *Mol Microbiol* **48:** 77-84.

27. Sassetti CM and Rubin EJ (2003) Genetic requirements for mycobacterial survival during infection. *Proc Natl Acad Sci U S A* **100:** 12989-12994.

28. Smith I (2003) *Mycobacterium tuberculosis* pathogenesis and molecular determinants of virulence. *Clin Microbiol Rev* **16:** 463-496.

29. Strong M, Graeber TG, Beeby M, Pellegrini M, Thompson MJ et al. (2003) Visualization and interpretation of protein networks in *Mycobacterium tuberculosis* based on hierarchical clustering of genome-wide functional linkage maps. *Nucleic Acids Res* **31:** 7099-7109.

30. WHO Report on Multidrug and extensivelydrug-resistant TB (M/XDR-TB) (2010) 2010: Global Report on Surveillance and Response, World Health Organization.

31. Yellaboina S, Goyal K and Mande SC (2007) Inferring genome-wide functional linkages in *E. coli* by combining improved genome context methods: Comparison with high-throughput experimental data. *Genome Res* **17:** 527-535.

Chapter 5

Understanding Communication Signals during Mycobacterial latency through Predicted Genome-wide Interactions of Proteins

5.1 Introduction

One of the enigmatic features of tuberculosis is that only about 5-10% of the infected individuals develop active tuberculosis [Bloom and Murray 1992]. In rest of the cases, *M. tuberculosis* persists in a dormant or a non-replicative state in human tissues for a prolonged time with a potential to resume growth when conditions favour [Stewart et al. 2003]. It is observed that the bacilli that are in the latent state in the human granulomas are metabolically less active with little exposure to oxygen [Stewart et al. 2003]. In addition to hypoxia, NO production from the host macrophages is an important factor in maintaining bacteria in the latent state [Voskuil et al. 2003]. Despite having known several genes that are induced during mycobacterial latency, the complex cascade of regulatory events that specifically direct and maintain bacteria in a latent phase in response to host environments are least understood. Importantly, resistance of dormant bacilli to majority of the drugs impedes control and eradication of tuberculosis [Stewart et al. 2003]. Therefore, it is important to understand the signals that drive the transition between active and dormant phenotypes and the associated regulatory mechanisms.

Towards addressing the phenomena associated with the switch between latent and actively replicating phases, several experimental studies have attempted to simulate the dormancy phase using the *in-vitro* models to generate gene expression profiles. Such conditions include hypoxia [Sherman et al. 2001; Park et al. 2003; Bacon et al. 2004; Muttucumaru et al. 2004; Voskuil et al. 2004], NO treatment [Voskuil et al. 2003], stationary phase [Voskuil et al. 2004] and nutrient deprivation [Betts et al. 2002]. Along with the *in-vitro* data, murine models of *M. tuberculosis* dormancy have also been studied [Schnappinger et al. 2003; Karakousis et al. 2004].

Since persistence is a collective response to multiple factors, a single model is unlikely to represent latency completely. Therefore, integration of all the models becomes inevitable to understand latency. Also, it is interesting to address the phenomenon of latency in *M. tuberculosis* through the analysis of functional interaction network. The predicted protein interaction map was therefore used to demarcate the proteins that might play important role in the dormant phase of *M. tuberculosis*.

5.2 Results

5.2.1 Persistence in *M. tuberculosis*

Since individual experimental models might not capture all the dormancy features of *M. tuberculosis*, and moreover considering that the microarray experiments might be affected by intrinsic noise, the commonality among these was considered to identify key genes regulating latency. In this regard, the gene expression datasets of different persistence models of *M. tuberculosis* were used to identify up and down-regulated genes during early persistence. The list of different models and the number of up and down-regulated genes are detailed in Table 1. Intriguingly, the overlap among the differentially regulated genes between different models of *M. tuberculosis* dormancy is not very high. For example, murine model of *M. tuberculosis* latency has 7%, 19%, 6% and 3% overlap with the other available models of NO treatment, hypoxia, stationary phase growth and starvation model respectively (Table S1 in Appendix IV). Therefore, the genes that are common to at least 5 of the 12 expression conditions related to latency were considered for analysis. Following this criterion, there were 50 genes that show increase in expression levels and 34 genes that are down-regulated during latency (Table S2 in Appendix IV).

In order to understand if the factors leading to dormancy are shared by different prokaryotic species, evolutionary conservation of the 84 genes was probed by constructing a binary phylogenetic profile of these genes across 481 genomes followed by counting the numbers of genomes harboring these genes. Interestingly, the 50 genes that are upregulated in *M. tuberculosis* dormancy are far less conserved across species than the 34 genes that are downregulated. As few as 16 upregulated genes are present in less than 145 of the 481 genomes. On the other hand, as many as 26 down regulated genes are present in at least 336 genomes (Figure 1). The downregulated genes represent those involved in basic cellular processes, such as replication, transcription and translation. The rates of these basic cellular processes being significantly slowed during dormancy offers a possible explanation that the downregulated genes are common to many species. The remarkable observation, however, that the upregulated genes are less conserved than the downregulated genes suggests a possible unique mechanism of dormancy adapted by *M. tuberculosis*. The 50

upregulated genes therefore appear to constitute a "dormancy signal" that is unique to *M. tuberculosis*. Such a dormancy signal might then be transmitted to the genes involved in basic cellular processes in order to slow down the overall metabolic rates.

Experimental Condition	Up-regulated Genes	Down-regulated Genes
O_2 Depletion		
Park et al, Mol Microbiol, 2003	161	71
Sherman et al, PNAS, 2001	135	79
Muttucumaru et al, Tuberculosis, 2004	358	381
Bacon et al, Tuberculosis, 2004	144	55
Voskuil et al, Tuberculosis, 2004 i) NRP Day 6 II) NRP Day 8	 94 116	 279 350
Stationary Phase		
Voskuil et al, Tuberculosis, 2004 i) Stationary Phase Day 6 ii) Stationary Phase Day 8	 15 39	 139 222
NO Model		
Voskuil et al, J. Exp. Med, 2003	223	181
Starvation Model		
Betts et al, Mol Microbiol, 2002	170	211
Murine Model		
Schnappinger et al, J. Exp, Med, 2003	154	90
Karakousis et al, J. Exp. Med, 2004	253	49

Table 1: List of publications related to dormancy models used in this study and the number of up and down-regulated genes in each model.

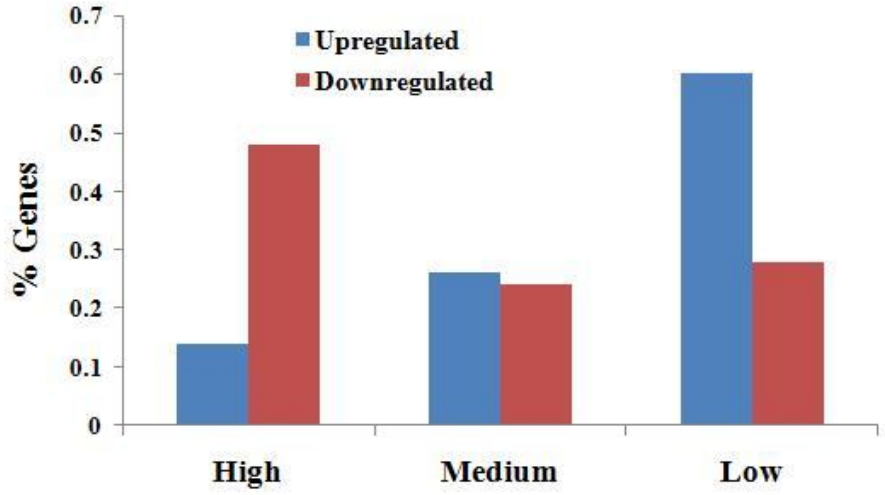

Figure 1: Phyletic retention of up and down-regulated genes. The genes that are upregulated during latency are less conserved compared to the genes that are downregulated.

Although the 84 genes showed coordinated regulation during latency-like conditions, it was interesting to probe if similar regulation controlled their expression even under other experimental conditions. The expression correlation among the 50 upregulated genes and the 34 downregulated genes across 154 growth conditions was then computed. Interestingly, the expression values among the 50 upregulated genes correlate highly (Figure 2). Similarly, those among the 34 downregulated genes also show strong correlation. On the other hand, the up and the down-regulated genes exhibit inverse correlation. This observation suggests that the 84 genes are coordinately regulated not only under latency-like conditions, but also are regulated in a controlled manner under most other conditions of growth. The 84 genes thus form a regulon-like structure with an extensive cross talk in their expression. Such a crosstalk can possibly be uncovered using the functional interaction network.

Furthemore, the protein functional linkages were integrated with the available gene regulatory information in order to understand the crosstalk between the 84 latency-related genes [Balázsi et al. 2008]. One of the important observations upon such an integration of the two networks was that DosR, a well studied dormancy associated protein, regulates the expression of about 25 upregulated genes (Figure 3). In addition, the genes regulated by DosR also interact extensively among themselves

forming a clique-like architecture in the network. Moreover, important latency genes such as *hspX*, *pfkB*, *Rv2030c* and *Rv2028c* are additionally regulated by sigma factor SigC, implying a multifarious regulatory circuit of latency. On the other hand, Rv3676, a transcriptional regulatory protein of the cAMP receptor protein (CRP) family [Kumar et al. 2010], regulates the expression of downregulated genes such as *Rv1566c*, *Rv1158c*, *hupB*, *lprK* and *mce1D*. Thus, a central regulatory circuit controlled by DosR, with degeneracy offered by transcription factors such as SigC and CRP, combined with an extensive network of interactions among the 84 genes appears to control latency in *M. tuberculosis*.

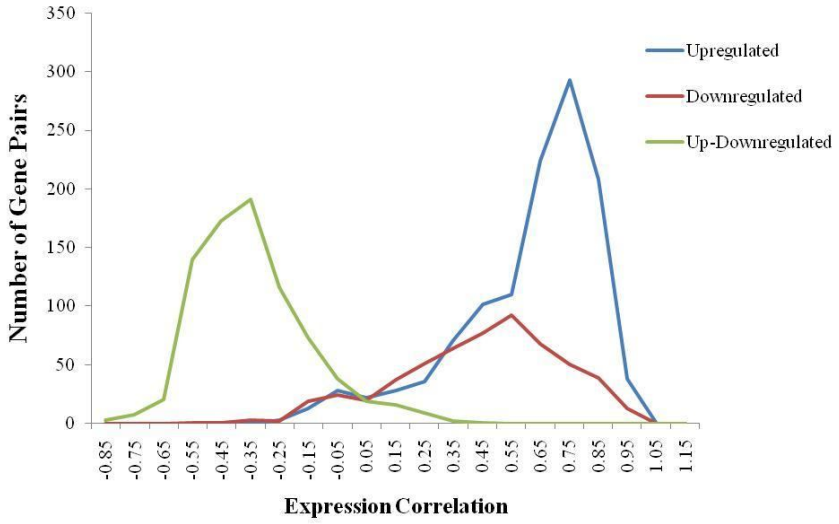

Figure 2: Expression Correlation between up and downregulated genes. Upregulated genes correlate in their expression and a similar trend is observed for downregulated genes. However, there is inverse correlation in expression between up and downregulated genes.

An interesting outcome of the network analysis pertains to the dormany signal detected by the 50 upregulated genes, and its transmission to the 34 downregulated genes. DosR interacts with two-component sensory kinases DosS (Rv3132c), DosT (Rv2027c) and Rv0845. It has been reported previously that DosR acts as a cognate response regulator of both DosS and DosT, which sense hypoxia and NO [Roberts et al. 2004]. Interestingly, a classification based on the region around phosphorylated sensor kinases of *M. tuberculosis* assigns DosS, DosT and Rv0845 to the same class [Tyagi

and Sharma 2004]. Whereas Rv0845 is a possible nitrate/nitrite sensor protein, DosS and DosT function as redox and hypoxia sensors respectively [Kumar et al. 2007]. It is therefore likely that DosS, DosT and Rv0845 sense the dormancy signal and trigger a regulatory cascade controlled by DosR.

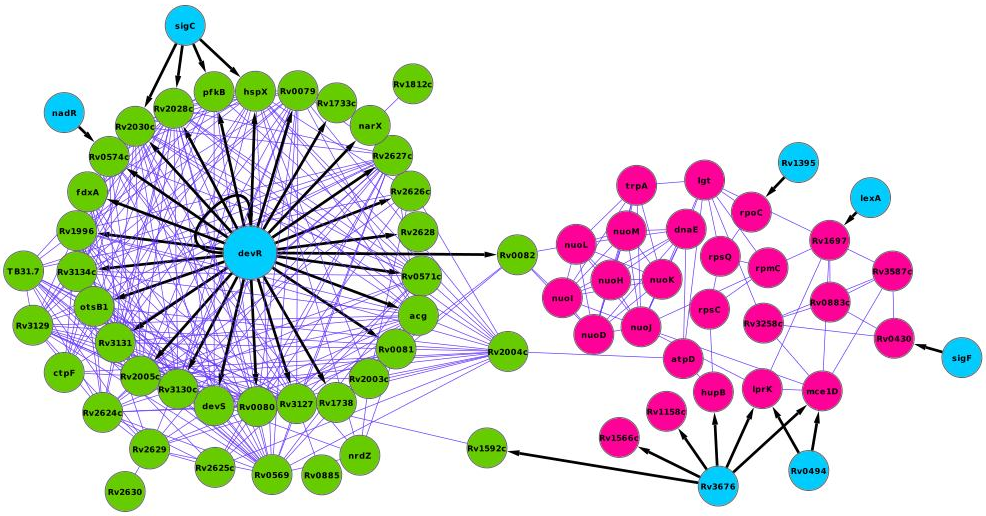

Figure 3: Graphical representation of functional interactions and regulatory interactions of up and downregulated genes. DosR acts as a master regulator, regulating the expression of a number of upregulated genes. A probable oxidoreductase Rv0082 under the regulation of DosR associates with the subunits of aerobic respiratory machine NADH oxidoreductase complex. Upregulated proteins are colored in green, downregulated proteins are colored in red and transcription factors are colored in blue. Directed edge denotes gene regulation and undirected edge is a protein functional linkage.

The gene regulatory network also suggests that a hypothetical oxidoreductase, Rv0082, is regulated by DosR with increase in its expression level under dormancy conditions. The functional interaction network further suggests that Rv0082 might associate with three of the subunits of NADH-ubiquinone oxidoreductase complex (NDH-1). The subunits of NDH-1 complex are involved in respiration and not surprisingly form a clique in the functional interaction network. Moreover, these are downregulated during latency. The expression correlation between NDH-1 genes spanning from *Rv3145-Rv3158* and ATP synthase genes spanning from *Rv1304-*

Rv1311 are correlated in their expression (Figure S1 in Appendix IV). This suggests a succeeding downregulation of ATP synthesis upon downregulation of NDH-1 genes. Interestingly, NDH-1 proteins interact with ribosomal proteins RpsC and RpsQ, and DNA polymerase DnaE in the subnetwork of functional interactions. The succession of protein connectivity in the subnetwork therefore suggests that DosR possibly communicates latency signals to the respiratory chain through Rv0082, resulting in the shutdown of respiration mediated by NDH-1 followed by growth suppression. Since NAD^+ pool is essential for TCA cycle and other biosynthetic pathways, it appears that the downregulation of NDH-1 respiration is critical in the early stages of latency, which might lead to the arrest of cell growth and division subsequently. Interestingly, the mutants of NADH dehydrogenase I of *Escherichia coli* show competitive disadvantage in the mixed stationary phase cultures [Zambrano and Kolter 1993]. The inability of these mutants to adapt to the stationary phase might arise due to their inefficiency in transmitting signals downstream to slowdown cellular growth processes. Thus, protein interactions and gene regulatory information support the complex regulatory hierarchy of the genes involved in latency.

In order to further address the complexity of the cross-talk between the up and the downregulated genes in the network, the network of the 84 genes was expanded to construct a “Dormancy Core” comprised of directly interacting proteins of the up and the downregulated genes. The “Dormancy Core” was divided into up or downregulated modules depending on the association of the proteins with the up or the downregulated genes. The number of nodes in the upregulated and downregulated modules was 172 and 632 respectively. There are 29 proteins which are common to both these modules (Figure 4). These 29 proteins might participate in direct signalling between the up and the down regulated modules.

Examination of the topological properties of this subnetwork interestingly showed that the nodes in the downregulated module possess higher degree compared to the up-regulated module (Figure S2(a) in Appendix IV). As the downregulated module is enriched by proteins of cell growth and division, it is not surprising that these nodes possess higher degree since high degree nodes are more likely to perform essential growth functions [Jeong et al. 2001]. Also, the downregulated module nodes

are closer to other proteins in the network as revealed by their closeness centrality compared to the up-regulated module (Figure S2(b) in Appendix IV). This suggests their important role in mediating the information flow in the network (Figure S2(c) in Appendix IV). However, there is no apparent difference in the clustering coefficients of these two modules (Figure S2(d) in Appendix IV). Thus, the known centrality characteristics are able to distinguish between the upregulated and downregulated modules in the network.

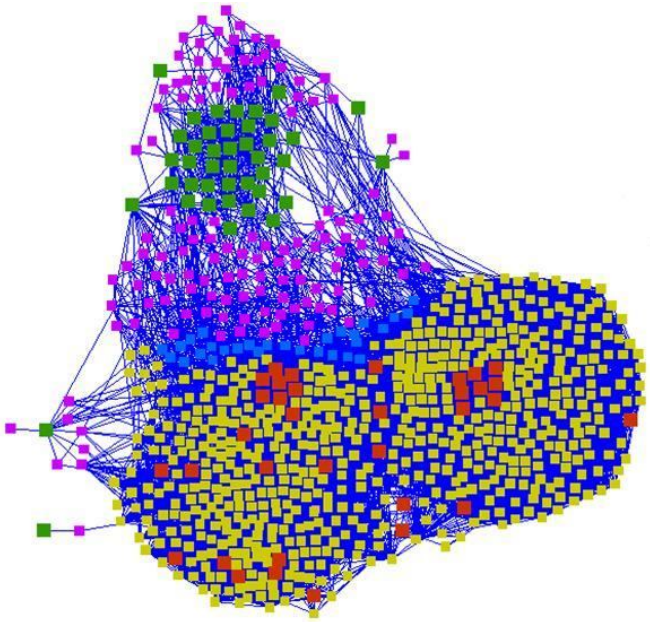

Figure 4: Schematic representation of the up and down-regulated modules. Node color represents the class as follows: Green – Upregulated proteins; Red – Downregulated proteins; Purple – First neighbors of the upregulated proteins; Yellow – First neighbors of the downregulated proteins; Blue – Proteins interacting with both upregulated and downregulated proteins.

The pathway mapping of up and downregulated modules (Figure S3 in Appendix IV) indicates that the genes from Information pathways are down-regulated during the dormant phase of *M. tuberculosis*. The examples are replication proteins such as DnaA, DnaB, DnaN and GyrA, translation initiation factors such InfA and InfB, proteins of the ribosomal complex, repair and recombination proteins. Interestingly, the subunits of ATP synthase belong to the downregulated module, suggesting an important role of electron transport during dormancy. The down-regulated module also contains Fad proteins which are involved in the degradation of

fatty acids, NADH dehydrogenase subunits, proteins involved in polyketides and non-ribosomal peptide synthesis, and cell envelop proteins from the families Lpr and Mur. The up-regulated module, on the other hand, includes the master regulator of dormancy DosR, DosS which functions coordinately with DosR, nitrate reductases NarG, NarJ and NarX, chaperones such as HspX and HtpG, polyketide synthetases such as MbtB and MbtC. Thus, the analysis of the module composition suggests possible pathways that are activated and repressed during the dormant phase of *M. tuberculosis.*

5.2.2 Shortest Paths Analysis in the Dormancy Module

An interesting aspect of dormancy is the communication between differentially regulated pathways [Boshoff and Barry 2005; Höner zu Bentrup and Russell 2001]. In this direction, the task was to identify the possible communication route between the up and downregulated nodes in the dormancy module. All the possible shortest paths were traced from the 50 upregulated nodes to the 34 downregulated nodes and the most probable shortest paths were derived (Methods). The intermediates in the most probable paths were then ranked based on their frequency of occurrence and the top 25% of the intermediates were selected (Table S3 in Appendix IV). These proteins, as it appears, might be involved in transmitting dormancy signals from the upregulated core proteins involved in dormancy, to the downregulated growth module. Below are the few examples that are discussed for their possible role in modulating dormancy.

Some of the well known dormancy related proteins such as DosR, DosS and HspX occur frequently in the most probable paths calculated. DosR and DosS are the two-component regulatory proteins that have been shown previously to activate a number of proteins in response to the onset of dormancy [Park et al. 2003]. HspX, a chaperone, is one such protein from the *dosR* regulon that shows significant induction during dormancy [Hu et al. 2006]. Another protein coded by *Rv2621c* is a hypothetical transcriptional regulatory protein which interacts with six of the 50 up-regulated nodes. It connects the down-regulated module through SseC2 which is a conserved hypothetical protein thought to be involved in sulphur metabolism. Both these proteins fall frequently in the most probable paths calculated. Figure 5 is a graphical

presentation of the interactions of these two proteins. Intriguingly, the interacting partners of SseC2 include WhiB1 and WhiB2, which are transcription factors known to be involved in septation and cell division [Gomez and Bishai 2000]. They require (4Fe-4S) for the catalytic activity and can function as protein disulfide reductases. The *whiB* homologue of *C. glutamicum* is critical for survival after oxidative stress [Kim et al. 2005]. Another interacting partner of SseC2 is Rv2175c, which is a transcription factor and a substrate of the PknL Ser/Thr kinase. The vicinity of *Rv2175c* to the *dcw* (division cell wall) cluster to which *pknL* belongs [Narayan et al. 2007] suggests its possible role in regulating cell growth and division. SseC2 also associates with antigen 85B (FbpB) which is a mycolyl transferase involved in cell wall biosynthesis [Belisle et al. 1997]. The cutinase precursor Cfp21, which promotes mycobacterial survival and virulence [West et al. 2008], is another interactor of SseC2. Other secreted proteins such as Cfp21, Cfp2, Rv3194c and Rv2672 also show interaction with SseC2. Thus, the cascade laid by Rv2621c and SseC2 in connecting essential proteins of dormancy to the cell division and growth proteins of the downregulated module appears to be important during switch from/to dormancy.

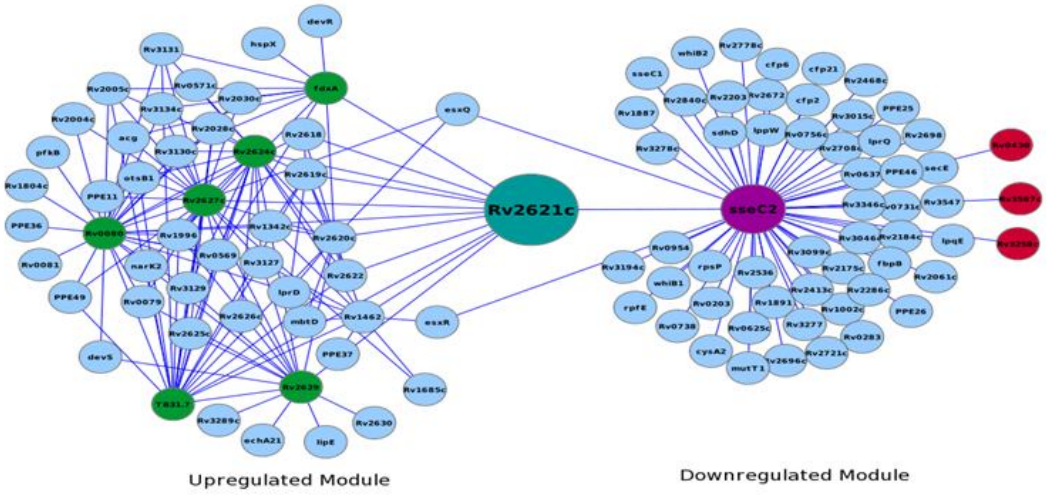

Figure 5: A sub-network depicting the connectivity mediated by proteins Rv2027c and SseC2**.** Both these proteins occur most frequently in the paths between up and downregulated genes. The upregulated proteins associating with Rv2027c are colored in green and the downregulated proteins associating with SseC2 are colored in red.

Another interesting protein of a two-component system is Rv1626, the crystal structure of which suggests its possible role in transcriptional antitermination [Morth et al. 2004]. In the functional linkage network it interacts with PknA, a Ser/Thr protein kinase and NarK2, a nitrate/nitrite reductase. Both these proteins are members of the up-regulated module. Interestingly, Rv1626 interacts with several down-regulated module proteins, some of which are adenylate kinase Adk, tryptophan synthase TrpA, ribosomal proteins RpsA and RplT, and a two-component transcriptional regulator MtrA.

Another example of an important protein in the most probable paths is EsxR, a secreted ESAT-6 like protein. Interestingly, one of the proteins it connects in the down-regulated module is RpfA, which is a resuscitation promoting factor required for resuscitation from dormant state [Kana et al. 2008]. Notably, Rpfs have been shown to promote growth in *Micrococcus luteus* and are important for virulence [Mukamolova et al. 2002]. Thus, the connectivity mediated by EsxR and RfpA between up and downregulated modules appears to be important. The subunits of ATP synthase AtpH, AtpG and AtpC also appear in the most frequent paths. Notably, an inhibitor diarylquinoline targets the proton pump of ATP synthase in *M. tuberculosis* [Andries et al. 2005]. Based on these observations, the proposed hypothesis is that an alteration in the mode of respiration possibly serves as a signal for growth and it is logical to assume that the proteins of ATP synthase communicate such signals. In order to further determine the controllers of the dormancy network, a Boolean model simulation with logical relations derived using gene expression correlations was performed by Dr. Hannah Rajasingh, TCS, Hyderabad.

5.3 Discussion

Understanding the persistent stage of *M. tuberculosis* has proved challenging and has profound implications in containing the disease as the current anti-tuberculosis drugs target only the cells that are actively growing [Stewart et al. 2003]. Moreover, the micro-environment of the granulomas, where latent *M. tuberculosis* resides, is impermeable to the drugs. Inevitably, the systems level understanding of

persistence to describe key players in this process, and their association with other proteins is not understood in great detail.

There are 84 differentially regulated genes which are common among many gene expression studies of various dormancy models of *M. tuberculosis*. The analysis interestingly reveals that these 84 genes are coordinately regulated not only under dormancy-like conditions, but rather form a regulon-like structure. Among these, the 34 downregulated genes show high evolutionary conservation. A logical argument here is that the evolutionary conservation of these genes is due to their participation in basic cellular processes. In contrast, the dormancy signal in *M. tuberculosis* appears to be unique, as evidenced by far less conservation of the upregulated genes. It might therefore appear that different bacteria have adopted different mechanisms of entering dormancy, leading eventually to shutting down of the highly conserved basic metabolic processes.

Having identified two distinct clusters of genes, 50 upregulated and 34 downregulated, it is important to understand how these might be coordinately regulated. Boolean modeling was therefore used to examine such transition to dormant phase in *M. tuberculosis* through the 84 differentially regulated genes. In Boolean modeling, an attractor state is the terminal vertex of the state transition graph, i.e. the state to which the system will converge given a certain input state. The attractor state can either be a single steady state, or a set of states through which the system cycles [Albert et al. 2008]. The latency model developed in this study attempts to understand the key regulators, which when turned on, results in an attractor state that mimics the latent phase of the pathogen. The latent state which is derived upon the activation of certain transcription factors is determined by comparing the states of the genes in the model with prior knowledge of their upregulation or repression during latency. Boolean modeling is therefore an attractive approach to address coordinated regulation among the 84 genes.

In this direction, the results of the Boolean model simulation of the latency subnetwork were provided by Dr. Hannah Rajasingh, TCS, Hyderabad. The results imply that the model converges to an attractor cycle in which about 92% of the upregulated genes remain active when four of the transcription factors in the model

namely, DosS (Rv3132c), DosR (Rv3133c), Rv0081 and CRP (Rv3676) are activated at the input. This suggests that these transcription factors are required to be expressed in order to maintain other members of the dormancy network in an active state. Several experimental studies have indicated involvement of a number of transcription factors in the regulation of initiation of the dormancy state [Stewart et al. 2003; Chao and Rubin 2010]. Therefore, the above mentioned transcription factors are not necessarily the only regulators of these processes. However, what the model indicates is that they are the minimum set of transcription factors required to obtain a system that exists in a dormancy-like state. While other factors are most likely involved in establishing the latent condition and adapting to external disturbances from the host macrophage, these four regulatory proteins appear to be the core set of regulators for initiating and maintaining signals for latency.

Two of the regulators predicted to be important for *M. tuberculosis* latency by Boolean modeling are DosR and DosS. Together they form a two-component regulatory system in which DosS is the sensory kinase and DosR is the corresponding response regulator [Roberts et al. 2004]. These regulators have earlier been implicated as key mediators of latency in several experimental studies [Stewart et al. 2003]. For example, rapid induction of *dosR* and *dosS* is observed upon reduced oxygen tension [Sherman et al. 2001]. Similarly, targeted disruption of *dosR* revealed that most of the genes that are induced by hypoxia are regulated by DosR [Park et al. 2003]. In addition, DosR and DosS are the members of the 'dormancy regulon' identified upon NO treatment [Voskuil et al. 2003]. Furthermore, significant load of bacilli and hypervirulence was observed in a SCID mouse model which was infected with *M. tuberculosis* with *dosR* deletion [Parish et al. 2003]. Activation of DosR is mediated by DosS, which has been established as a redox sensor with O_2, NO and CO as modulatory ligands [Kumar et al. 2007]. Hence, the two-component system DosR-DosS appears to play a major role in initiating and maintaining latency. The results using Boolean modeling and network analysis reinforce the importance of these two proteins in establishing latency in *M. tuberculosis.*

Another regulatory protein predicted by Boolean modeling to be important for latency is Rv3676, a cAMP receptor family protein (CRP), which is a global regulator

of number of pathways. Computational identification of possible regulatory sites for CRP has earlier revealed a number of genes, some of which are implicated in starvation and hypoxic conditions [Bai et al. 2003]. Therefore, the association of Rv3676 in latency appears significant. Although there are no reports on the direct involvement of this protein in latency, it will be interesting to study the regulatory role of Rv3676 in this context. Thus, the hypothesis is that Rv3676, by virtue of being a global regulator, might influence cross talk among the genes involved in latency.

The fourth important transcriptional regulator suggested by Boolean modeling in *M. tuberculosis* latency is Rv0081, which is a regulatory protein from the ArsR/SmtB family [Campbell et al. 2007]. *Rv0081* is the first gene in the operonic locus *Rv0081-Rv0088,* which codes for the components of formate dehydrogenase complex. *Rv0081* has been observed to be upregulated in multiple latency models, and is also shown to be regulated by DosR in the transcription regulatory network of *M. tuberculosis* [Balázsi et al. 2008]. Interestingly, the functional interaction network predicted by us in this study places Rv0082 at the intersection of upregulated and downregulated gene clusters. By being on an operon it might be assumed that Rv0081 controls the transcription of *Rv0082*. Thus, the predicted functional interaction network and Boolean modeling together suggest important roles of Rv0081 and Rv0082 in communicating latency signals between modules of upregulated and downregulated genes in *M. tuberculosis* latency.

The network based approach supplemented with Boolean modeling in elucidating crosstalk between the upregulated and downregulated genes leads us to propose a fascinating hypothesis of the latency process. What is observed in the current study, and is well known, is that DosR plays an important regulatory role in the dormancy switch. The dormany signals sensed by two-component sensor kinases, DosS, DosT and possibly by Rv0845 is transmitted to DosR which is a cognate response regulator. In the downstream, the signal is relayed through Rv0081 to the respiratory chain mediated by the Rv0082. The information flow from Rv0082 then triggers switching off ATP synthesis, leading eventually to significant slowing down of replication, transcription and translation processes. Thus, through an intricate communication signal, the basic cellular processes such as cell division and growth are

shut down. Some of the hypotheses proposed in this work will obviously need to be tested experimentally.

5.4 Methods

Functional annotations of *M. tuberculosis* proteins were derived from KEGG [Kanehisa and Goto 2000], TubercuList (http://genolist.pasteur.fr/TubercuList/) and Sanger (http://www.sanger.ac.uk/) databases. The shortest paths were calculated using Dijkstra's algorithm [Dijkstra 1959]. Since there can be more than one shortest path for a pair of nodes, most probable paths were derived by considering most frequently occurring proteins in the shortest paths at each position. Sub-networks were visualized and analyzed using Cytoscape 2.4.1 [Shannon et al. 2003].

Combined subnetwork was constructed by merging protein functional linkages and gene regulatory interactions [Balázsi et al. 2008]. If the interaction for two proteins is represented as both protein functional linkage as well as gene regulatory interaction, the latter was considered for the analysis.

5.5 References

1. Andries K, Verhasselt P, Guillemont J, Göhlmann HW, Neefs JM et al. (2005) A diarylquinoline drug active on the ATP synthase of *Mycobacterium tuberculosis. Science* **307:** 223-227.

2. Bacon J, James BW, Wernisch L, Williams A, Morley KA et al. (2004) The influence of reduced oxygen availability on pathogenicity and gene expression in *Mycobacterium tuberculosis. Tuberculosis* **84:** 205-217.

3. Bai G, McCue LA and McDonough KA (2005) Characterization of *Mycobacterium tuberculosis* Rv3676 (CRPMt), a cyclic AMP receptor protein-like DNA binding protein. *J Bacteriol.* **187:** 7795-804.

4. Balázsi G, Heath AP, Shi L and Gennaro ML (2008) The temporal response of the *Mycobacterium tuberculosis* gene regulatory network during growth arrest. *Mol Syst Biol* **4:** 225.

5. Belisle JT, Vissa VD, Sievert T, Takayama K, Brennan PJ et al. (1997) Role of the major antigen of *Mycobacterium tuberculosis* in cell wall biogenesis. *Science* **276:** 1420-2.

6. Betts JC, Lukey PT, Robb LC, McAdam RA and Duncan K (2002) Evaluation of a nutrient starvation model of *Mycobacterium tuberculosis* persistence by gene and protein expression profiling. *Mol Microbiol* **43:** 717-731.

7. Bloom BR and Murray CJ (1992) Tuberculosis: commentary on a reemergent killer. *Science* **257:** 1055-1064.

8. Boshoff HI and Barry CE 3rd (2005) Tuberculosis - metabolism and respiration in the absence of growth. *Nat Rev Microbiol* **3:** 70-80.

9. Campbell DR, Chapman KE, Waldron KJ, Tottey S, Kendall S et al. (2007) Mycobacterial cells have dual nickel-cobalt sensors: sequence relationships and metal sites of metal-responsive repressors are not congruent. *J Biol Chem*. **282:** 32298-310.

10. Chao MC and Rubin EJ (2010) Letting sleeping dos lie: does dormancy play a role in tuberculosis? *Annu Rev Microbiol* **64:** 293-311.

11. Dijkstra EW (1959) *Numerische Math* **1:** 269–271.

12. Gomez JE and Bishai WR (2000) whmD is an essential mycobacterial gene required for proper septation and cell division. *Proc Natl Acad Sci U S A* **97:** 8554-8559.

13. Hu Y, Movahedzadeh F, Stoker NG and Coates AR (2006) Deletion of the *Mycobacterium tuberculosis* alpha-crystallin-like *hspX* gene causes increased bacterial growth *in vivo. Infect Immun* **74:** 861-868.

14. Höner zu Bentrup K and Russell DG (2001) Mycobacterial persistence: adaptation to a changing environment. *Trends Microbiol* **9:** 597-605.

15. Kana BD, Gordhan BG, Downing KJ, Sung N, Vostroktunova G et al. (2008) The resuscitation-promoting factors of *Mycobacterium tuberculosis* are required for virulence and resuscitation from dormancy but are collectively dispensable for growth in vitro. *Mol Microbiol* **67:** 672-684.

16. Kanehisa M and Goto S (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* **28:** 27-30.

17. Karakousis PC, Yoshimatsu T, Lamichhane G, Woolwine SC, Nuermberger EL et al. (2004) Dormancy phenotype displayed by extracellular *Mycobacterium tuberculosis* within artificial granulomas in mice. *J Exp Med* **200:** 647-657.

18. Kim TH, Park JS, Kim HJ, Kim Y, Kim P et al. (2005) The *whcE* gene of *Corynebacterium glutamicum* is important for survival following heat and oxidative stress. *Biochem Biophys Res Commun* **337:** 757-764.

19. Kumar A, Toledo JC, Patel RP, Lancaster JR Jr and Steyn AJ (2007) *Mycobacterium tuberculosis* DosS is a redox sensor and DosT is a hypoxia sensor. *Proc Natl Acad Sci U S A* **104:** 11568-11573.

20. Kumar P, Joshi DC, Akif M, Akhter Y, Hasnain SE et al. (2010) Mapping conformational transitions in cyclic AMP receptor protein: crystal structure and normal-mode analysis of *Mycobacterium tuberculosis* apo-cAMP receptor protein. *Biophys J* **98:** 305-314.

21. Morth JP, Feng V, Perry LJ, Svergun DI and Tucker PA (2004) The crystal and solution structure of a putative transcriptional antiterminator from *Mycobacterium tuberculosis. Structure* **12:** 1595-1605.

22. Mukamolova GV, Turapov OA, Kazarian K, Telkov M, Kaprelyants AS et al. (2002) The *rpf* gene of *Micrococcus luteus* encodes an essential secreted growth factor. *Mol Microbiol* **46:** 611-621.

23. Muttucumaru DG, Roberts G, Hinds J, Stabler RA and Parish T (2004) Gene expression profile of *Mycobacterium tuberculosis* in a non-replicating state. *Tuberculosis* **84:** 239-246.

24. Narayan A, Sachdeva P, Sharma K, Saini AK, Tyagi AK et al. (2007) Serine threonine protein kinases of mycobacterial genus: phylogeny to function. *Physiol Genomics* **29:** 66-75.

25. Parish T, Smith DA, Kendall S, Casali N, Bancroft GJ et al. (2003) Deletion of two-component regulatory systems increases the virulence of *Mycobacterium tuberculosis*. *Infect Immun.* **3:** 1134-40.

26. Park HD, Guinn KM, Harrell MI, Liao R, Voskuil MI et al. (2003) *Rv3133c/dosR* is a transcription factor that mediates the hypoxic response of *Mycobacterium tuberculosis*. *Mol Microbiol* **48:** 833-843.

27. Roberts DM, Liao RP, Wisedchaisri G, Hol WG and Sherman DR (2004) Two sensor kinases contribute to the hypoxic response of *Mycobacterium tuberculosis*. *J Biol Chem* **279:** 23082-23087.

28. Schnappinger D, Ehrt S, Voskuil MI, Liu Y, Mangan JA, Monahan et al. (2003) Transcriptional Adaptation of *Mycobacterium tuberculosis* within Macrophages: Insights into the Phagosomal Environment. *J Exp Med* **198:** 693-704.

29. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT et al. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13:** 2498-2504.

30. Sherman DR, Voskuil M, Schnappinger D, Liao R, Harrell MI et al. (2001) Regulation of the *Mycobacterium tuberculosis* hypoxic response gene encoding alpha-crystallin. *Proc Natl Acad Sci U S A* **98:** 7534-7539.

31. Stewart GR, Robertson BD and Young DB (2003) Tuberculosis: a problem with persistence. *Nat Rev Microbiol* **1:** 97-105.

32. Tyagi JS and Sharma D (2004) Signal transduction systems of mycobacteria with special reference to *M. tuberculosis. Current Science* **86:** 93-102.

33. Voskuil MI, Schnappinger D, Visconti KC, Harrell MI, Dolganov GM et al. (2003) Inhibition of respiration by nitric oxide induces a *Mycobacterium tuberculosis* dormancy program. *J Exp Med* **198:** 705-713.

34. Voskuil MI, Visconti KC and Schoolnik GK (2004) *Mycobacterium tuberculosis* gene expression during adaptation to stationary phase and low-oxygen dormancy. *Tuberculosis* **84:** 218-227.

35. West NP, Wozniak TM, Valenzuela J, Feng CG, Sher A et al. (2008) Immunological diversity within a family of cutinase-like proteins of *Mycobacterium tuberculosis*. *Vaccine* **26:** 3853-3859.

36. Zambrano MM and Kolter R (1993) *Escherichia coli* mutants lacking NADH dehydrogenase I have a competitive disadvantage in stationary phase. *J Bacteriol* **175:** 5642-5647.

Appendix I

The data is provided in the CD with path: /AppendixI/. Following are the contents of the folder.

1. Figure S1: Schematic representation of the analysis. (a) Pictorial representation of differential gene expression in the network context. Colours red and green represent the nodes expressed uniquely under the defined conditions, whereas blue nodes are expressed under both the conditions. (b) Four way comparison of the networks. UWT represents the wild type, TWT- UV treated wild type, UML- the *lexA* mutant and TML- UV-treated *lexA* mutant.

2. Figure S2: The overlap of the interactions and the nodes in UWT-TWT and UML-TML.

3. Figure S3: Mapping of unique nodes to different metabolic pathways.

4. Figure S4: Path length analysis for the nodes of phosphotransferase system. The absence of CmtB in the UV treated *lexA* mutant network increases the pathlength from YggD, a putative *cmt* operon transcriptional regulator to other phospho transferase system (PTS) enzymes.

5. Table S1: Microarray data processing and Network information for the conditional networks.

6. Table S2: Uniquely expressed genes list in the four-way comparison study.

7. Table S3: Interacting partners of Hda in UV treated wild-type network, classified according to functional classes and with their functions.

8. Table S4: Functions of the high centrality measure nodes in the comparison set UWT-TWT and UML-TML.

9. Table S5: List of Mutation dependent/independent and UV treatment dependent/independent proteins.

Appendix II

The data is provided in the CD with path: /AppendixII/. Following are the contents of the folder.

1. Figure S1: Expression correlation and Protein interactions. Genes coding for the proteins that are functionally linked are correlated in their expression.

2. Figure S2: Gene expression correlations between identified modules in *E. coli* and *S. oneidensis.* The genes in the identified modules in *E. coli* correlate among themselves and anticorrelate between each other. However, this regulation in expression is not observed for the corresponding modules of orthologous proteins in *S. oneidensis.* This figure illustrates the distribution of the expression correlations between the genes of Module1 (2a), Module 2 (2b) and between Module 1 and Module 2 genes (2c).

3. Table S1: List of widely expressed, conditionally expressed and less expressed genes in *E. coli*.

4. Table S2: KEGG pathway mapping of the anticorrelated modules.

Appendix III

The data is provided in the CD with path: /AppendixIII/. Following are the contents of the folder.

1. Table S1: Predicted Protein functional linkages of *M. tuberculosis*

2. Table S2: Comparison of predicted interactions with other available interactions. There is about 30% overlap between predicted interactions and previous reports

3. Table S3: List of high centrality proteins in the network

4. Table S4: List of communities identified in the interaction map using Infomap community detection tool

5. Table S5: List of selected genomes for constructing phylogenetic profile

6. Table S6: List of microarray conditions used for calculating expression correlations.

Appendix IV

The data is provided in the CD with path: /AppendixIV/. Following are the contents of the folder.

1. Figure S1: Expression correlation between genes coding for NDH-1 subunits and the genes coding for ATP synthase subunits. The plot suggests that NDH-1 genes and ATP synthase genes are correlated in their expression.

2. Figure S2: Differential Network Properties of the Modules of up and downregulated genes. Proteins in the downregulated module show high degree centrality ($P < 2.2e^{-16}$), high closeness centrality ($P < 2.2e^{-16}$) and high betweenness centrality ($P < 3.9e^{-09}$) compared to the proteins in the upregulated module. However, there is no apparent difference in their clustering coefficients ($P < 0.44$).

3. Figure S3: TubercuList pathway map of the proteins belonging to up and down-regulated modules. Genes belonging to Information pathway are significantly downregulated during dormancy.

4. Table S1: Comparison of different dormancy models in terms of number of genes up and downregulated in each.

5. Table S2: List of up and downregulated genes during dormancy and their functions.

6. Table S3: List of proteins possibly mediating the dormancy signals between up and down-regulated modules.