THE GLOBAL ORGANIZATION AND TOPOLOGICAL PROPERTIES OF *DROSOPHILA MELANOGASTER*

A Thesis submitted to the College of
Graduate Studies and Research
in Partial Fulfillment of the Requirements
for the Degree of Master of Science
in the Department of Chemical Engineering at
the University of Saskatchewan
Saskatoon

By
Thanigaimani Rajarathinam

PERMISSION TO USE

In presenting this thesis, in partial fulfillment of the requirements for a degree of Master of Science from the University of Saskatchewan, the researcher agrees that the Libraries of this University may make it freely available for inspection. The researcher further agrees that permission for copying of this thesis in any manner, completely or in part, for scholarly purposes may be granted by the professor who supervised this thesis work. In his absence, the Head of the Department or the Dean of the College of Graduate Studies and Research at the University of Saskatchewan may also give this permission. Any copying, publication or use of this thesis or parts thereof for financial gain without the researcher's written permission is strictly prohibited by law. Due recognition shall be given to the researcher and to the University of Saskatchewan in any academic use, which may be made of any material in this thesis.

Please address all requests for permission to copy or to make any other use of material in this thesis completely or in part to:

Head of the Department of Chemical Engineering

University of Saskatchewan

Saskatoon, Saskatchewan

Canada S7N 5A9

ABSTRACT

The fundamental principles governing the natural phenomena of life is one of the critical issues receiving due importance in recent years. Most complex real-world systems are found to have a similar networking model that manages their behavioral pattern. Recent scientific discoveries have furnished evidence that most real world networks follow a scale-free architecture. A number of research efforts are in progress to facilitate the learning of valuable information by recognizing the underlying reality in the vast amount of genomic data that is becoming available. A key feature of scale-free architecture is the vitality of the highly connected nodes (hubs).

This project focuses on the multi-cellular organism *Drosophila melanogaster*, an established model system for human biology. The major objective is to analyze the protein-protein interaction and the metabolic network of the organism to consider the architectural patterns and the consequence of removal of hubs on the topological parameters of the two interaction networks.

Analysis shows that both interaction networks pursue a scale-free model establishing the fact that real networks from varied situations conform to the small world pattern. Similarly, the topology of the two networks suffers drastic variations on the removal of the hubs. It is found that the topological parameters of average path length and diameter show a two-fold and three-fold increase on the deletion of hubs for the protein-protein interaction and metabolic interaction network, respectively. The arbitrary exclusion of the nodes does not show any remarkable disparity in the topological parameters of the two networks. This aberrant behavior for the two cases

underscores the significance of the most linked nodes to the natural topology of the networks.

ACKNOWLEDGEMENTS

I would like to express earnest gratitude to my respected supervisor, Dr. Yen-Han Lin for his consistent guidance and support all through the duration of this thesis work. His critical appraisal and observation have been priceless to every stage of this work.

I wish to extend appreciation to my advisory committee members, Dr. Richard Evitts and Dr. Fangxiang Wu for their valuable suggestions and productive feedback. I am also indebted to my external examiner, Dr. Lope Tabil, Jr. for his editing proficiency, meticulousness and encouragement.

DEDICATION

To

My father for his unwavering emotional and monetary support throughout my life;

Thanks, Dad.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AD	Activation Domain
ADP	Adenosine diphosphate
ATP	Adenosine triphosphate
BD	Binding Domain
cDNA	complementary DNA
DNA	Deoxyribonucleic Acid
EC	Enzyme Commission
GRN	Gene Regulatory Network
H^+	Hydrogen ion
KEGG	Kyoto Encyclopedia of Genes and Genomes
LAN	Local Area Network
MAPK	Mitogen-Activated Protein Kinase
m-RNA	messenger-RNA
RNA	Ribonucleic Acid
TCA	Tri Carboxylic Acid

LIST OF SYMBOLS

N	Number of nodes or vertices in a network
P, p	Probability distribution
$G_{n,p}$	Erdos and Renyi model
C_n, C	Clustering coefficient of node ‘n’
C_{in}	Product metabolites
C_{out}	Substrate metabolites
k	Degree or number of connections of each node
γ	Exponent in the power law relation
$p(k)$	Probability that a node has ‘k’ links

CHAPTER 1

INTRODUCTION

The understanding of the phenomenon of life demands a strong discernment of its doctrines (*refer to glossary*). Several components are responsible for the function and maintenance of cellular processes in any living species.

1.1 The Cell

The cell is the building block and basic subunit of any self-regulating living system. It is responsible for carrying out the different processes that occur in an organism. There are different kinds of these classes of cells, each with their own function. Examples of tissues containing different types of cells include bone, blood, muscle, skin and hair. A cell contains several key elements including nucleus, mitochondrion, ribosome, vacuole, centriole and cytoplasm (*refer to glossary for information on organic components of the cell*). Each element of the cell is responsible for a particular task in the mechanism of the organism [1]. For instance, nucleus, endoplasmic reticulum, vacuole and ribosome are involved in the developmental process of an organism. A typical animal cell, containing these elements that assist in the safe and proper execution of the mechanism of the biological species, is shown in Figure 1.1.

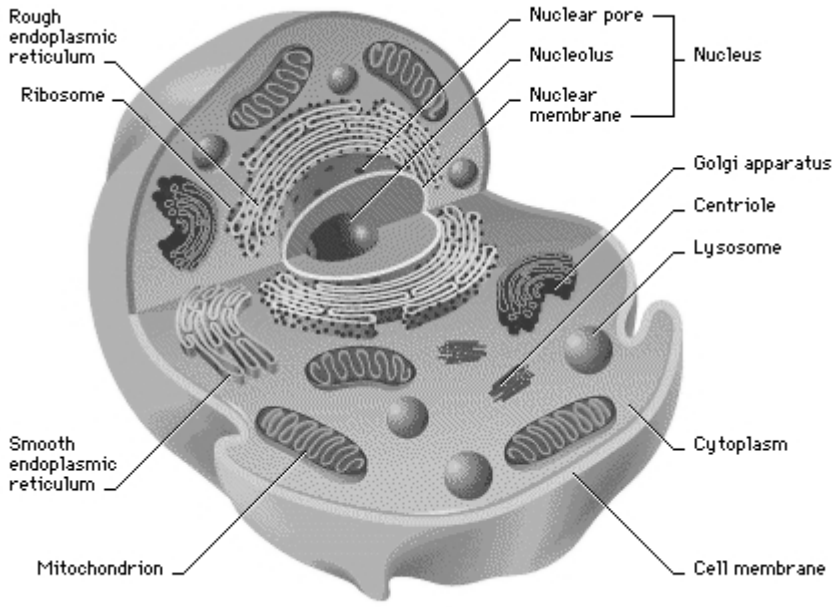

Figure 1.1: An illustration of the components of an animal cell (eukaryote) [2]

1.2 Genetics of an Organism

A chromosome is the visible state of genetic material during a phase of division of the cell. Deoxyribonucleic acid (DNA) is the major storehouse of genetic material.

The Central Dogma of Biology

The DNA molecule carries and transfers the genetic information in all living organisms. It replicates its information through a process that involves many enzymes. The DNA is then transcribed into Ribonucleic acid (RNA) by specific enzymes called RNA polymerases through a process known as *transcription*. Some viruses use RNA instead of DNA as their genetic material. The final product is a messenger-RNA (m-RNA). It is used as a template that directs the synthesis of protein by ribosomes.

Translation occurs at the ribosome where the m-RNA is used to specify the sequence of amino acids in the polypeptide chain [3]. Some proteins are synthesized in specific cells. This progression of events is often referred to as the Central Dogma of Molecular Biology [3]. A pictorial summary is given in Figure 1.2.

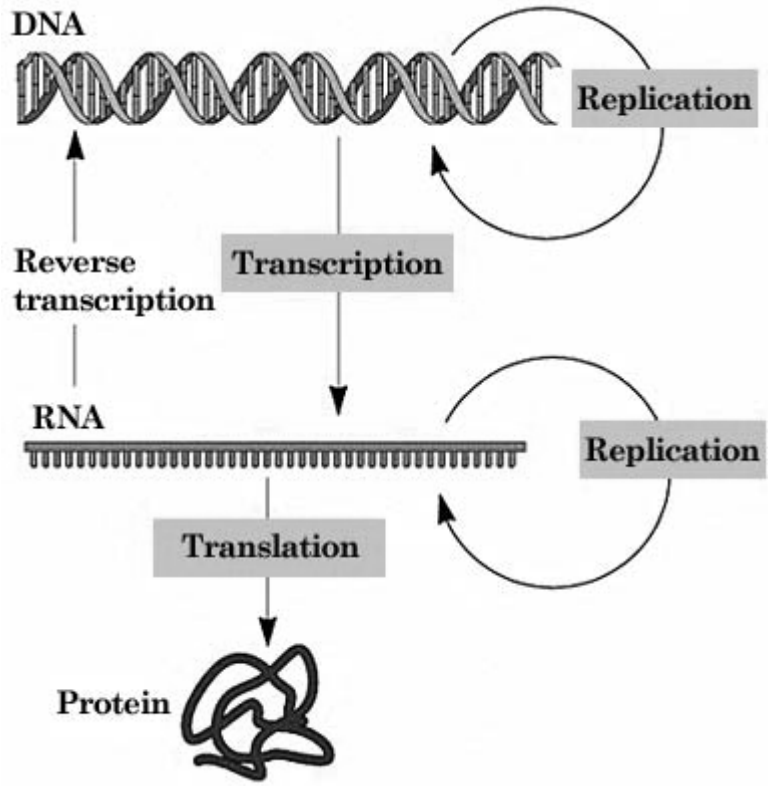

Figure 1.2: A graphical representation detailing the steps involved in the Central Dogma of Molecular Biology [4]

1.3 Biochemical Networks

Genomic study has enabled scientists to relate genes and proteins by virtue of the similarity in their genomic sequences. Nonetheless, metabolites cannot simply be correlated with genes or proteins. Hitherto, the most effective way to do this association is to utilize reference databases that accumulate the acknowledged information about

metabolic reactions, i.e., their substrates, products and enzymes. These databases serve as the basis for identifying and validating drug targets depending on the knowledge of the biochemical pathways in which potential target molecules operate within cells. For this reason, the study of biochemical pathways is the focus of drug discovery research and is central to the approach of many genomic and pharmaceutical companies. Biochemical processes arbitrate the interaction of cells with their environment and are accountable for most of the information processing that occurs inside them. The three main categories of biochemical processes are described below.

1.3.1 Metabolic Pathways

Living cells extract, convert and accumulate energy from nutrients obtained from food sources by the process of metabolism [5]. Metabolism collectively refers to all the physical and chemical activity, which occurs in the cells of the organism that may either discharge energy from nutrients or utilizes energy to generate new substances like proteins for continual growth and functioning, essential to maintain life. Metabolic pathways are sequences of chemical reactions each catalyzed by an enzyme that enable the formation of certain product molecules from other small substrates [6]. A typical metabolic pathway is shown in Figure 1.3. Metabolites are usually small molecules while enzymes are proteins.

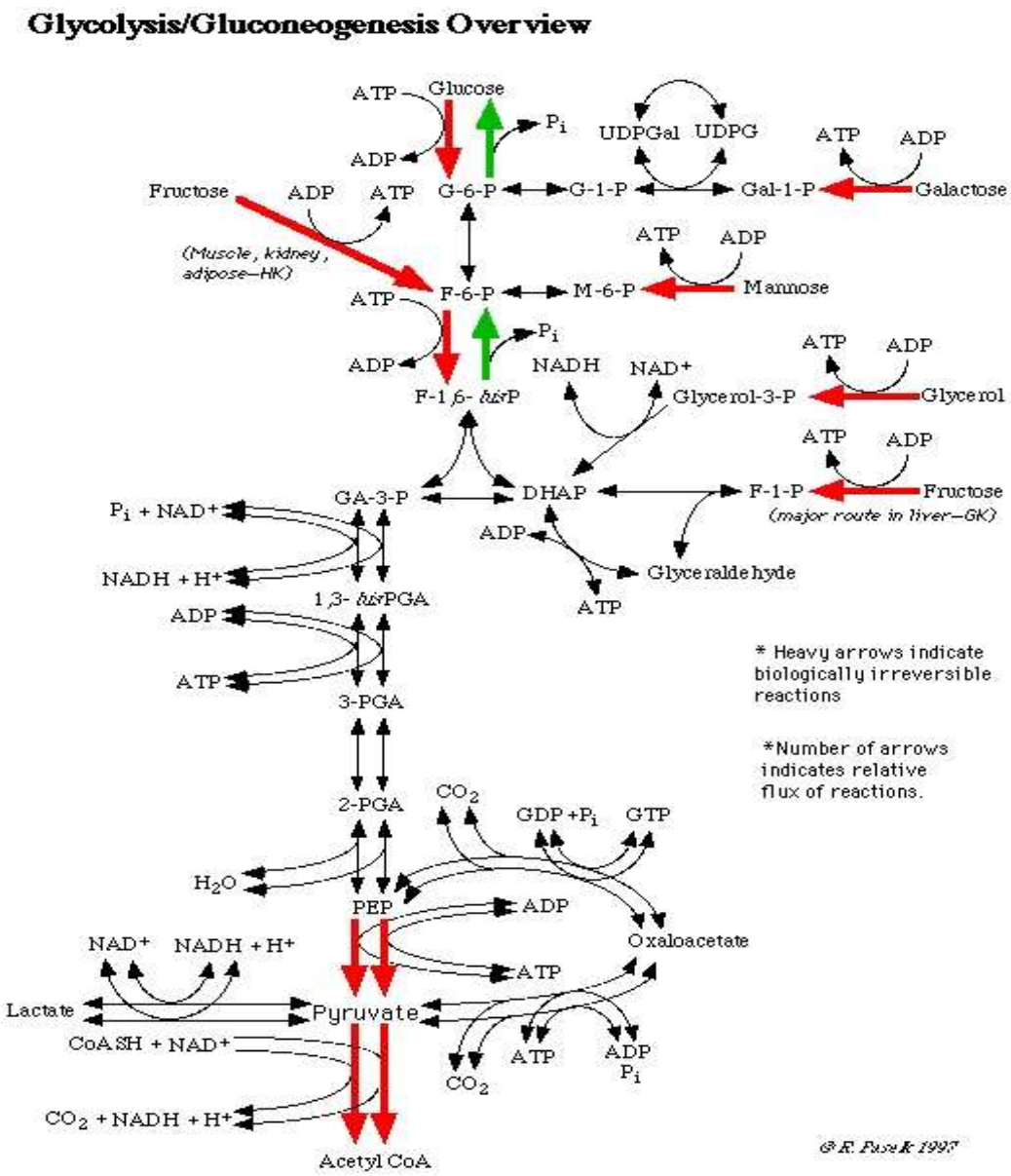

Figure 1.3: Glycolysis/gluconeogenesis pathway (carbon metabolism) [7]

The glycolysis pathway converts glucose into pyruvate with the simultaneous production of a relatively small amount of Adenosine triphosphate (ATP). Glycolysis can be carried out even in the absence of oxygen and hence, is an essential pathway for those organisms that ferment sugars. For example, the glycolysis pathway is used by yeast to produce the alcohol found in beer.

1.3.1.1 Types of metabolic pathways

A metabolic pathway may be defined as a series of chemical reactions resulting in either the formation of metabolic product (s) to be used or stored by the cell (metabolic sink) or the initiation of another metabolic pathway (then labeled as a flux causing step). They contribute to two kinds of processes:

Catabolism is the oxidative degradation of molecules into simpler ones. Some of the commonly found catabolic pathways are glycogenolysis (conversion of glycogen into glucose), glycolysis (conversion of glucose into pyruvate and ATP) (refer to Figure 1.3) and protein catabolism (hydrolysis of proteins into amino acids).

Anabolism is the reductive synthesis of molecules to produce building blocks and compounds from simpler precursors. A few examples of anabolic pathways include glycogenesis (process of glycogen synthesis from glucose molecules) and gluconeogenesis (formation of glucose, especially by the liver, from non-carbohydrate sources such as amino acids).

It is imperative to comprehend that the pathways, whether catabolic or anabolic in nature, are coordinated (with support from hormones) by the energy requirements (e.g., development and absorption) and physiological actions of an organism.

1.3.1.2 Energy generation

Metabolic pathways can be envisioned as a sequence of enzyme-catalyzed reactions. A cyclical process of energy conversion occurs in cells of living organisms

during metabolism. Chemical energy is manufactured from nutrients during catabolism and this energy, in turn, is used to produce new molecules during anabolism from the same type of nutrients, to maintain the structure and function of an organism. Both types of reactions are essential for thriving metabolism. Enzymes carry out the blending of an energy yielding reaction with an energy consuming one. Enzymes of metabolic pathways are able to capture this energy in small quantities and store it in the form of internal high-energy compounds drastically reducing the amount of energy lost as heat [8].

1.3.1.3 Reaction networks as graphs

The complex web of reactions involves substrates reacting with one another in the presence of enzymes to yield products. Each reaction generates a product that may become the substrate for another reaction. The metabolic compounds involved in the mechanism of an organism can be characterized by association or network graphs. The compounds are the nodes and the interactions, which are reactions leading to another compound that may be either substrate or product, represent the resultant links. The discernment of this obscure world is facilitated by such a construction and is explored exhaustively in ensuing chapters.

1.3.2 Gene Regulatory Pathways

A biological organism does not build all the proteins that it is competent to produce at all times. As an alternative, it becomes accustomed to the environment and constructs only those genetic products that are essential for its continued and secure existence in a particular environment. Gene regulation allows the organism to sense its environment and respond appropriately by expressing the suitable set of genes needed

for that precise setting. A gene regulatory pathway can be called the on-off switch of a cell operating at the gene level. Gene regulatory networks (also genetic regulatory networks or GRN) are a collection of DNA segments in a cell that interact with one another and with other substances in the cell, thereby governing the rates at which genes in the network are transcribed into m-RNA. Transcription factors - proteins that promote or repress transcription either directly or indirectly - unite the regulatory DNA elements. They dynamically coordinate the level of expression for each gene in the genome by controlling the nature of that gene to be transcribed into RNA. Each RNA transcript then functions as the template for synthesis of a specific protein by the process of translation [9].

A transcription factor regulatory network is shown in Figure 1.4. A GRN may include one or more input signaling pathways, regulatory proteins that incorporate the input signals, RNA, several target genes and proteins produced from those target genes. Besides, such networks frequently include vibrant feedback loops that offer further regulation of structural design and output [9].

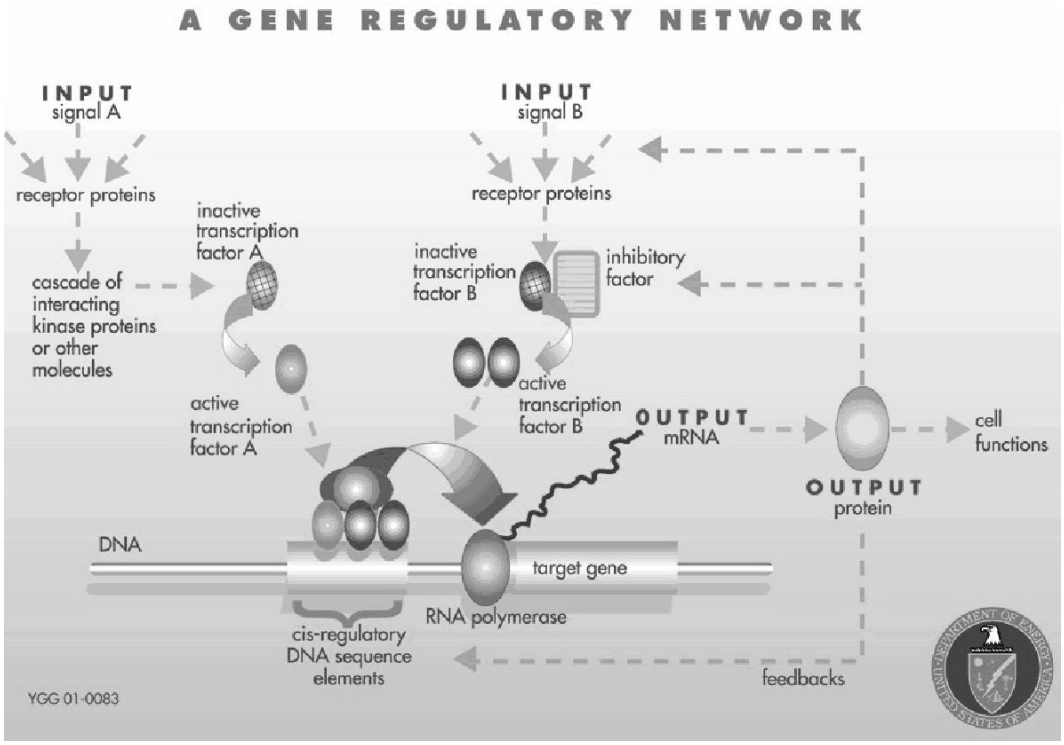

Figure 1.4: Gene regulatory network [10]

1.3.3 Signal Transduction Pathways

Signal transduction is the process of flow of signals by which an extra-cellular signal (typically a hormone or neurotransmitter) interrelates with a receptor at the cell surface. This interrelation causes a change in the functioning of the cell (e.g., activating glucose uptake or initiating cell division) by permitting signals in the form of small ion movement, in or out of the cell. These ion movements effect changes in the electrical potential of the cells that, in turn, proliferates the signal along the cell [11]. More composite signal transduction involves the coupling of ligand-receptor interactions to many intracellular events including phosphorylations by serine/threonine kinases.

Signals are transduced by modification of the protein activity or location by another protein. Protein phosphorylations cause a change in enzyme activities and protein conformations. The ultimate outcome is a shift in cellular activity and transformation of genes expressed within the responding cells [11]. Therefore, signal transduction networks can be expressed to be pathways of molecular interactions that provide communication between the cell membrane and intracellular end-points, leading to some modification in the cell. As an example, the Mitogen-Activated Protein Kinase (MAPK) pathway of *Drosophila melanogaster* is shown in Figure 1.5.

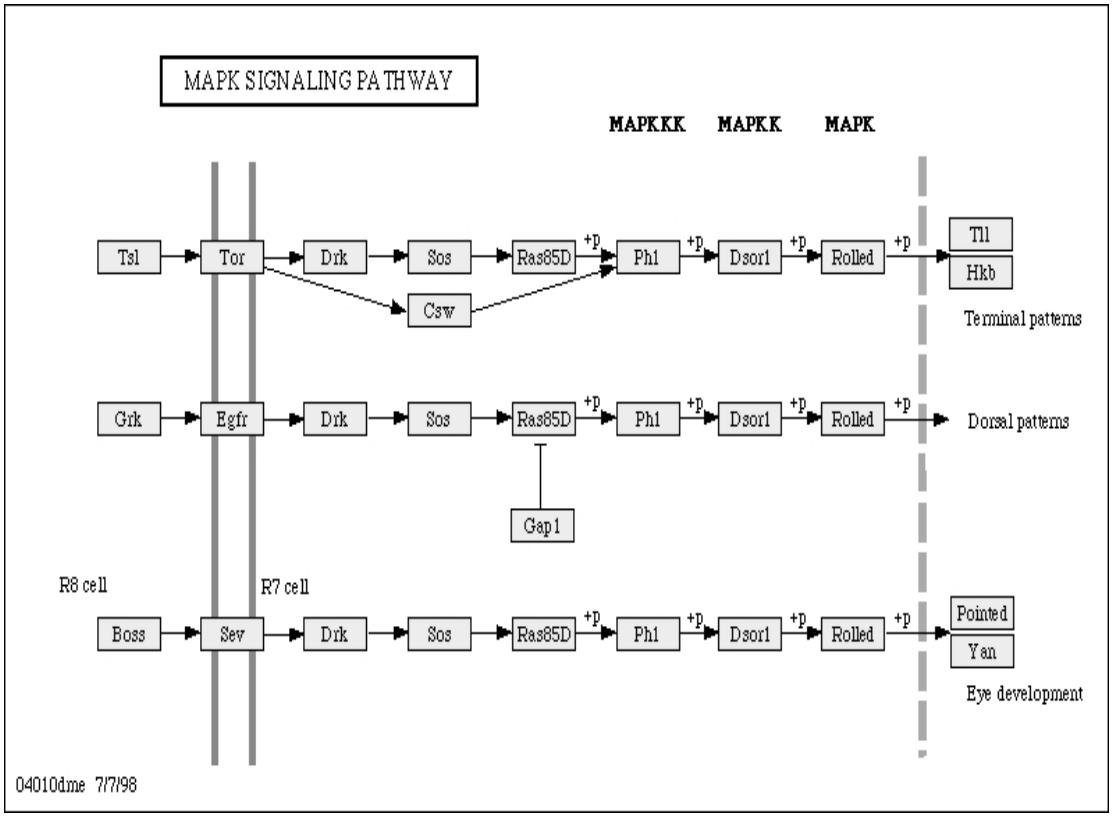

Figure 1.5: A schematic sketch of the MAPK signaling pathway in *Drosophila melanogaster* [12]

Metabolic, gene regulatory and signal transduction pathways are integrated and reliant on one another. For instance, gene regulation circuits are fed by external signals

transmitted by signal transduction pathways. The composite nature of these systems makes their precise understanding difficult.

Although it is in its seminal stages, systems biology is becoming a major field drawing significant consideration from researchers and business enterprises. The study of pathway information is of singular interest to biotechnology and pharmaceutical companies because it represents a level of assessment that is steps ahead of conservative static studies. The definitive goal of systems biology is to be able to replicate such biological systems and then to envisage the effects of particular perturbations. In the innovative model of genomic-based drug discovery and development, it is likely that a few good tools applied to the right pathways will pilot to valuable discoveries. Consequently, systems analysis focused institutions emphasize cross-disciplinary training for aspiring investigators in order to ease their entry into the field.

1.4 Proteins

Proteins are, beyond doubt, the physical starting point of life. They are the most important biochemical agents in the body. Their name comes from the Greek word *proteios*, which means *prime* or *chief* that provides an allusion to their importance.

1.4.1 Essence of Proteins

Proteins are macromolecular compounds constructed from one or more un-branched chains of amino acids to become polymers. The human body requires 20

types of amino acids to form the various proteins. A protein may contain 200 to 300 amino acids but some are much smaller and a few very much larger. The biggest to date is titin, a protein found in skeletal and cardiac muscle that contains about 27,000 amino acids in a single chain [13].

1.4.1.1 Amino acids - the structural building blocks of proteins

As mentioned earlier, proteins are composed of chains of amino acid sequences. A sequence of many such amino acids is referred to as a polypeptide. The complete product, either one or more chains of amino acids, is called a protein. Conjugated proteins, in addition, contain other kinds of molecules. For example, glycoproteins contain carbohydrates, nucleoproteins comprise nucleic acids and lipoproteins include lipids.

There are 20 different α-amino acids pertinent to the building of mammalian proteins and each can be distinguished by the R-group substitution on the α-carbon atom (except in the case of glycine where the R-group is hydrogen). This carbon, as revealed in Figure 1.6, is the α-carbon.

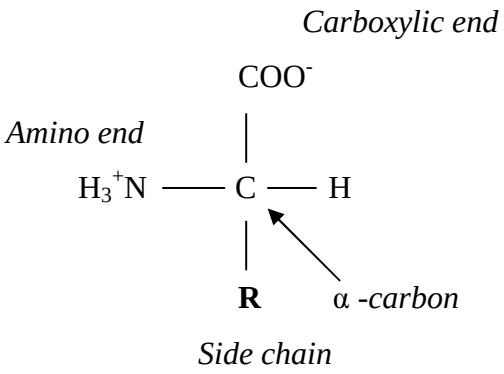

Figure 1.6: The general structural formula representation for an amino acid

The α-amino acids in peptides and proteins (excluding proline) consist of a carboxylic acid (-COOH) and an amino (-NH_2) functional group attached to the same tetrahedral carbon atom. The carboxyl or amino group may be ionized to form NH_3^+ or COO^- as an upshot of a condensation reaction between the amino group of one amino acid and the carboxyl group of another [14].

Each amino acid is different and therefore has its own unique properties. There are two classes of amino acids dependent on the nature of the R-group. Hydrophobic amino acids repel the aqueous environment and hence, dwell mainly in the interior of proteins. This class of amino acids does not ionize or participate in the formation of hydrogen bonds. Hydrophilic amino acids which are present on the exterior surfaces of proteins or in the reactive centers of enzymes tend to interact with the aqueous environment, frequently leading to the formation of hydrogen bonds [14].

Quite a few amino acids are found either in free or combined states (i.e., not associated with peptides or proteins). These non-protein associated amino acids perform specialized functions, apart from the development of peptides and proteins, such as the action of tyrosine as a neurotransmitter in the formation of thyroid hormones or glutamate.

1.4.1.2 Protein structure

There are four discrete aspects to the structure of a protein as explained in Figure 1.7.

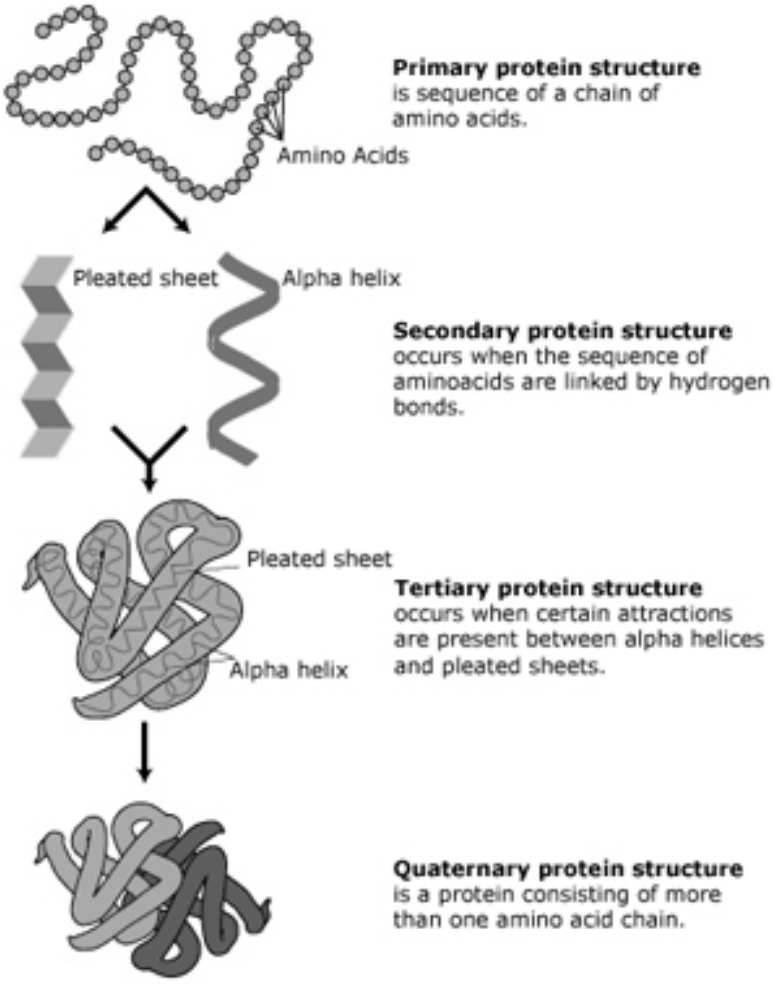

Figure 1.7: The different structures of a protein molecule [15]

Besides these levels of structure, proteins may shift between numerous parallel structures at the time of execution of their biological function.

1.4.2 Characteristics of Proteins

Proteins are involved in almost all biological activities of the cell. Every structural or enzymatic function in the living cell depends on proteins. They perform a wide variety of functions in the cell serving as enzymes, structural components or signaling molecules.

1.4.2.1 Functions of proteins

Proteins:

- Give structure to hair, skin and bones down to the cellular level
- Allow the transport of chemicals within and outside the cell. Proteins are interspersed throughout the cell membrane to attract important nutrients and permit their passage through the membrane
- Act as hormones in the body, coordinating all bodily processes at the molecular level
- Operate as antibodies in the support of the immune system. Antibodies attach themselves to foreign intruders (e.g., viruses) and incapacitate them so that they can be disposed
- Are the transcription factors that trigger and disable genes to guide the differentiation of the cell and its subsequent responsiveness to signals reaching it
- Act as enzymes, facilitating all chemical production in the body during the building of chemical compounds

1.4.2.2 Interaction of proteins

Proteins are seldom known to function in isolation. All proteins in any cell are united through a widespread web of connections, where non-covalent interactions are endlessly generated and dissociated. A variety of forces that is accountable for these interactions includes electrostatic forces, van der Waal's forces, hydrogen bonds and hydrophobic effects. It is extensively believed that the non-affinity towards water impels the interaction of protein pairs while hydrogen bonds and electrostatic interactions govern the specificity of the interface [16]. Proteins can adjust their structures according to the requirements and attach themselves to various molecules dependent on the task to be performed. Water is typically barred from the interfacial area [16]. However, among the various kinds of interactions like protein-protein, protein-DNA and protein-RNA, the protein-protein interactions mediate almost all biological processes central to life. Hence, their study is critical in identifying the fundamental concepts about interacting proteins.

1.4.2.3 Protein-protein interactions

An understanding of the interacting protein pairs offers insight into the function of important genes [17]. The description of the protein-protein interactions with the help of potent bioinformatics tools can be utilized to expose relevant pathways. The interpretation of these biological pathways involved in disease and drug response enhances the knowledge of the system under investigation. It enables researchers to categorize the genes and the proteins, which were not formerly related with disease or drug response, to aid in the development of fresh therapeutics.

Protein-protein interactions are highly significant as they provide functional information about one another. Biologically significant protein-protein interactions vary from the more general class of physical interactions. The principal difference is that both proteins must be in their suitable states (e.g., covalently modified state, conformational state, cellular location state, etc.) in a biological interaction [18]. This regulation of protein states through protein-protein interactions forms the basis for many dynamic biological processes that occur inside the cells [18]. As a result, the understanding of these interactions warrants information about the protein states.

A decisive goal of studying protein-protein interaction is to establish, assign and/or typify the function of a protein [19]. The physical interaction between a new protein and another protein having a known function can serve as a cursor that the two proteins may have a common function [20]. In addition, physical interactions are undoubtedly of enormous utility during the study of single proteins or defined biological processes. However, they do not echo the vast amount of information that has been amassed in the biological literature.

These protein-protein interactions, whether physical or biological, are represented in the form of maps called network graphs, where the proteins are the nodes and their ensuing interactions with other proteins are denoted by links connecting the two proteins. Mathematical standards can then be applied to study such networks. The principles of network theory are explained in Chapter 2. Protein interaction maps concentrate mainly on physical contacts without obvious chemical conversions. Understanding such protein-protein interactions is a crucial component of integrated biology. The availability of the genome sequences of hundreds of species, ranging from bacteria to human beings, poses two major challenges to biologists in the elucidation of this blueprint of life, namely

- Recognizing the function of each gene product and

- Understanding phenotypes through the biological interactions of the gene products

1.4.2.4 Detection, utility and complexity of protein-protein interactions

Large scale and high throughput experimental techniques have been developed to address these questions by acquiring data on the whole genome, instead of just a few genes [21]. The main methods available to identify protein interactions are two-hybrid analysis, mass spectrometry and mutation studies. The classical view of protein function focuses on the action of a single protein molecule, its biochemical activity or molecular function. An expanded view defines the function of a protein in the context of its network of interactions. Each protein interacts with various partners that also interact with other proteins. Any protein would fail to implement its specified function unless it binds to other biomolecules [20]. Most of the cellular processes are coordinated by these specific protein-protein interactions. All these interactions connect the proteins into an extensive and intricate web. Every protein function in the context of this web of interacting molecules and its interactions with other molecules define how its biochemical activities are exploited and regulated in related biological processes.

A large proportion of known protein-protein interactions have been detected by genome-scale two-hybrid assays. When researchers explore a gene and its function, it is a primary requisite to learn about the gene involved in the problem that is under investigation. In any case, once researchers have targeted the right gene, they are competent enough to duplicate and modify it. This alteration enables them to study its properties based on a deep-rooted understanding of its construction and working. This basic approach also applies to many single proteins. Biochemists can analyze a cell, extract the protein of interest, purify it using a succession of chromatography methods

and study its properties using well-established methods. In each of these cases, a single gene or a single protein remains stable and constant.

Nevertheless, interacting protein pairs are a class apart. Protein-protein interactions are almost momentary by definition. Any protein-protein interaction has the central purpose of producing some type of regulatory change in response to ecological circumstances. For this reason, the associations between proteins involved in these interactions are not as well built as the bonds between bases in DNA or between amino acids in a single protein. Hence, the immense complexity involved in the generation of this vital information highlights the significance of these protein-protein interactions.

The following chapter deals with the background information about the representation of interactions between metabolic compounds and protein pairs. A fundamental review of graphical network study is offered to develop a rigid understanding of the topic.

CHAPTER 2

NETWORKS AND BIOLOGY

This chapter demonstrates the transition from the biological world to the realm of networks. The concept of networks is important to this thesis work. Biology rejoins toward the latter part of the chapter by which time, the reader should have developed a strong integrating bond between the two interlinked fields.

2.1 Visualization of Networks

Please refer to glossary for the description of the various terms involved in the study of networks.

Networks as graphs

Networks play a vital role in more than a few facets of life. A network is an interconnected or interrelated chain, group or system. A network can be envisaged as a connection among people or objects. An example of an uncomplicated networking arrangement may be a set of computer terminals connected to one another. The computer systems are called as nodes and the wires connecting them are termed as links or connections. A link represents a pair wise relationship. A local area network (LAN) provides networking capability to a group of computers in close proximity to each other, such as in an office building, school or home. The smallest network can have

exactly two computers and a large network can accommodate thousands of computer systems.

A network is useful for sharing diverse resources like files, printers, games or other applications and may be connected to a multitude of other networks, thereby forming a vast chain of unified information systems. This network web can be depicted as a graph where the nodes are connected by the edges for the purpose of application of mathematical principles to evaluate their intrinsic characteristics.

2.1.1 Simple Illustration of Networks

A real-life instance of networking can be observed at a point-of-sale terminal in a certain section of a supermarket that forms part of a larger network comprising all the computers in the supermarket. This LAN, in turn, is linked as a component of a wider network, to the bank's network because credit card authorization is to be obtained from it. This simple illustration may appear like a complex one.

2.1.2 Complex Networks

On a broader perspective, social ties - familial and professional, the World Wide Web, network of scientific papers connected by citations, electrical power grids, transportation systems and biological networks are examples of real-world complex networks.

To elaborate on a few of the above examples, in a social network, the people are the nodes and the relationship between one another form the links [22]. The power stations and the electrical cables connecting them operate as nodes and links, respectively in an electrical power grid system [22]. In a biological network, the compounds (metabolites) perform as nodes and the reactions that convert it to other compounds denote their links [23]. Thus, various complex systems that exist are also instances of networking. The following section explores the complex world of networks to study their diverse features and applications.

2.2 Classification of Networks

Networks that exist are of varied nature. Complex networks are modeled using realistic concerns. A brief description of each of the basic type of networks is given below.

2.2.1 Regular Network

A regular network is one, where every node in the network is connected to the same number of nodes, i.e., it has the same number of links. The diagram shown in Figure 2.1 illustrates a crystal lattice structure.

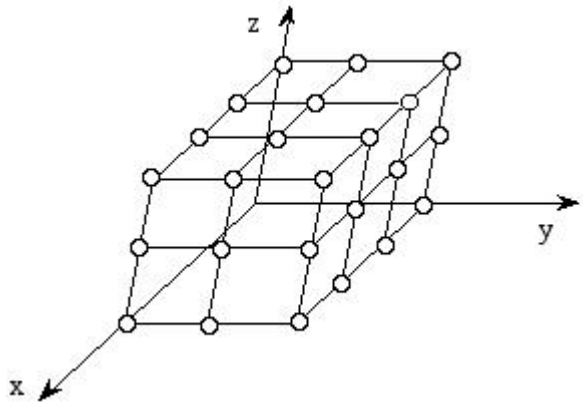

Figure 2.1: Crystal lattice structure demonstrating a regular networking arrangement [24]

Every node of this 3-dimensional design as indicated by small circles, is connected to the same number of nodes (six) or, in other words, has the same number of partners. Hence, a regular network is referred to as an *ideal* network. Some examples of this type of network include chains, grid structures and crystal lattices.

2.2.2 Random Network

The first venture into the dominion of networks was undertaken by Erdős and Rényi and developed later during the 1960's [25, 26, 27]. They pioneered a new concept, known as random networking, and presented it to the scientific community. According to their model, a group of N nodes is joined by links with probability p that are positioned between the pairs of nodes selected homogeneously in a random fashion creating a network containing approximately $p*N*(N-1)/2$ links. Of the several versions of this model, the commonly used representation is denoted as $G_{n,p}$. In this model, each possible edge between two nodes occurs with an independent probability p and absent with the probability $1-p$ [28]. In other words, it is the collection of graphs of N vertices

where each graph emerges with a probability corresponding to its number of edges. Complex systems are modeled by connecting nodes with randomly placed links as shown in Figure 2.2.

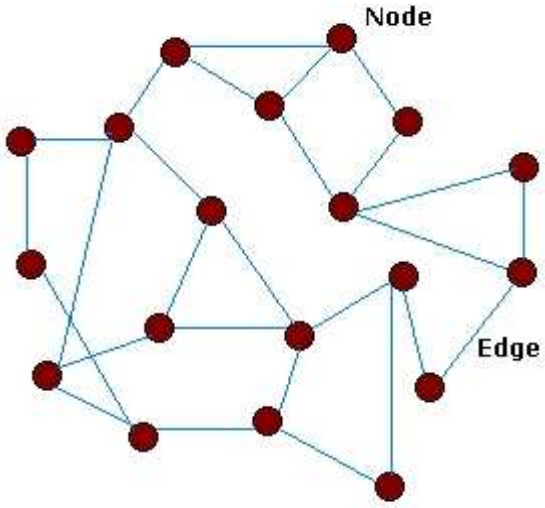

Figure 2.2: A random network where most nodes have identical number (three) of links

2.2.2.1 Giant component of the network

The original papers by Erdős and Rényi [25, 26, 27] have a multitude of interesting properties. A component can be defined as a subset of nodes in the graph that can be traversed by other nodes using a path through the network [28]. For very small degrees, a large number of nodes are disconnected from each other, as there are few edges in the graph. As the size of the graph becomes greater, the component size remains constant. The large component, which contains the majority of nodes in the graph, contains a set portion of the total number of nodes that scales linearly with the size of the whole graph above a threshold value. This large component is the *giant component* of the graph. The giant component in a random model is indicative of the behavior of real-world networks [28].

An example of a random network might be the number of customers who visit a particular store in one day and the frequency with which they meet a person at the cosmetics, fresh vegetable or meat counters. The measurement of this frequency distribution of visits to the various counters follows a classical bell curve or a Poisson distribution [29], as shown in Figure 2.3.

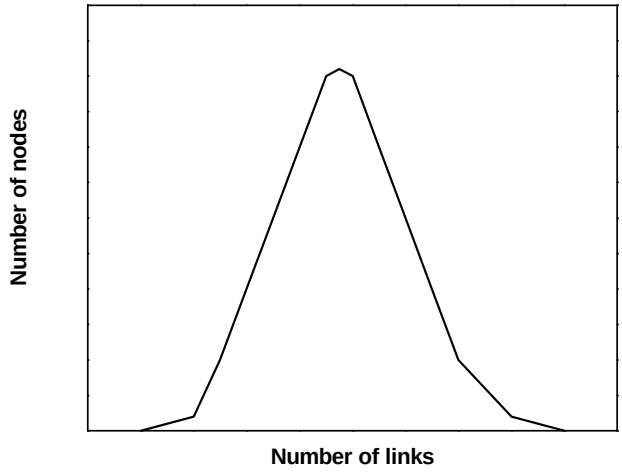

Figure 2.3: The Poisson distribution of a random network

2.2.2.2 Attributes of a random network

In a random network, very few nodes are well connected. Despite the randomly placed links, all nodes have “more or less” the same number of links. The average number of links is representative of the majority of nodes. This explains the bell shape of the distribution [22]. The peak point of the curve is the average number of links characterizing most nodes in the network. Hence, this network is also referred to as a *typical* network. The homogeneous nature of a random network makes it tolerant to

disruptions. Alternatively, it is prone to random malfunction as every node is equally as essential as another is. The failure of one node may cause the network to break into small clusters, thereby causing the network to lose its significance [22]. Thus, in a random network, it is not possible to construct any form of valuable and adaptive system. Another facet of random networks is the possibility of very long path lengths between any two nodes. A node may have to traverse a multitude of intermediate nodes to get to its target node. This signifies that, in terms of communication flow, the network does not have an efficient pathway. Mathematicians had been following the random network theory as the base for all computations about systemic pathways and distribution modeling until the scale-free network model was proposed in 1998.

2.2.3 Scale-free Network

In 1998, Watts and Strogatz published a groundbreaking paper that proposed another type of network pattern in addition to the already existent models [30]. This new representation called scale-free network model, fell between the classes of regular and random networks. Their discovery caused researchers to reassess the established norm of network modeling.

Watts and Strogatz rewired a regular network with a probability $P = 0$ to an intermediate level such that it did not reach a probability $P = 1$ as for a random network, where P is the probability of finding a random distribution [30]. This modification resulted in a middle ground with a probability between 0 and 1 (Figure 2.4). This intermediary probability is the range of scale-free networking patterns.

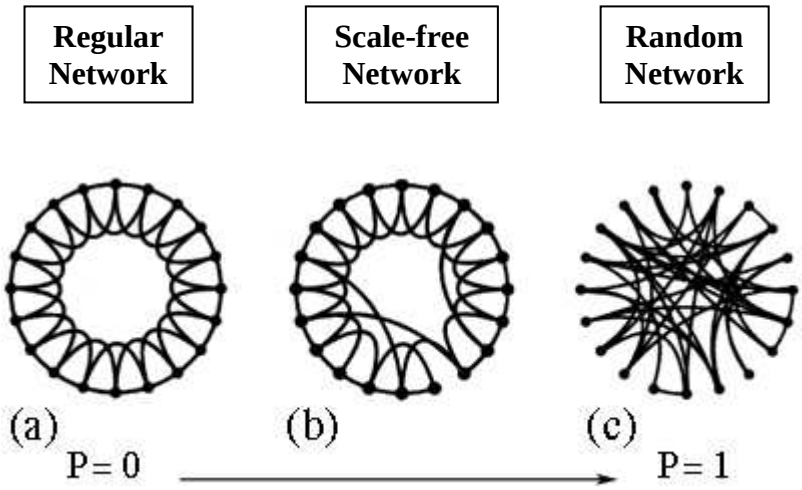

Figure 2.4: A sketch indicating the probability values for the three types of networks [30]

2.2.3.1 Characteristics of scale-free networks

A scale-free network (Figure 2.5) is entirely divergent from a regular and random network. In this type of network, a few nodes have a large number of links and tend to dominate the rest of the nodes having lesser number of links [30]. The nodes having a large number of links are called as hubs.

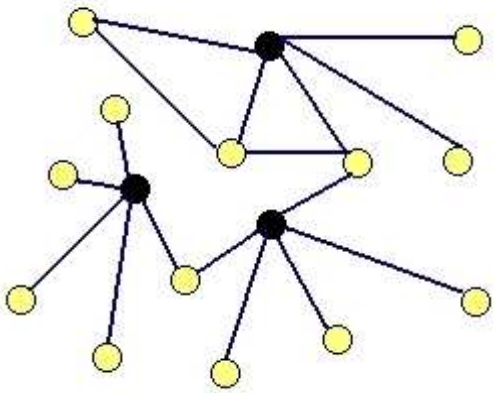

Figure 2.5: A scale-free network (the dark colored nodes are the hubs having a large number of links when compared to the other nodes)

Due to the absence of any specific dimension that can be used to classify such networks, they are referred to as scale-free networks. The presence of many different scales, though not one being typical, has led to researchers also referring to these networks as *scale-rich* networks. A notable feature of these networks is that a node would require only a few steps to reach another node. In other words, it is very easy to travel from one node to another within the network. This *small world* nature of the networks has enabled scientists to design many real-world networks based on this model.

The distribution of the nodes in a scale-free network declines with an increase in the number of links and is found to decay as a power law according to the relation:

$$p(k) \sim k^{-\gamma} \qquad (1)$$

where "~" denotes proportionality, $p(k)$ signifies probability and "k" represents a specific number of links [31]. The exponent γ is found to have a value of 2.1-2.4 for

three varied cases of the World Wide Web [32, 33], the metabolic networks [23] and the Internet [34]. Scale-free networks generally have an exponent value of 2.0 - 3.0.

The allocation of nodes does not follow a Poisson distribution as for the random model; instead, it is characterized by a decreasing function (Figure 2.6a).

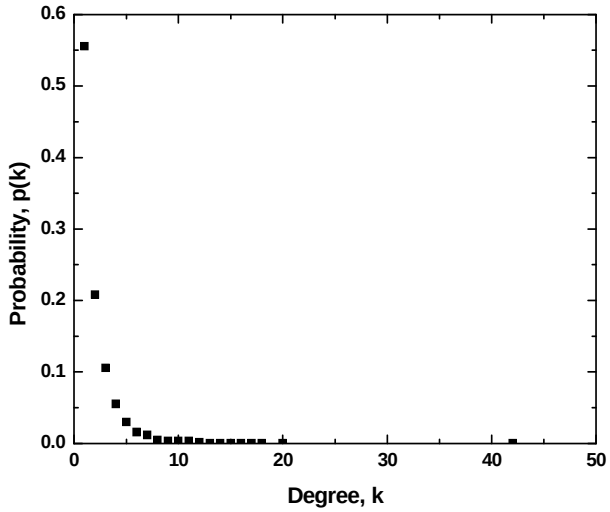

Figure 2.6a: Distribution model of a scale-free network on linear scale [30]

The plot of the probability distribution, on a log-log scale, gives a linear correlation (shown in Figure 2.6b) establishing the fact that the distribution follows a power law [29].

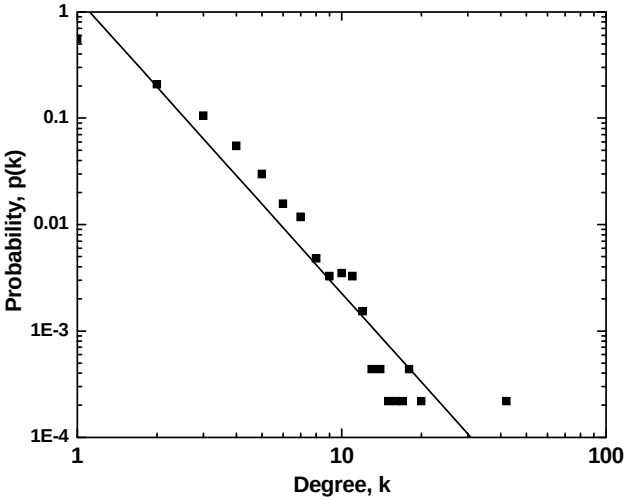

Figure 2.6b: Power law distribution of a scale-free network on logarithmic scale [30]

The diminishing nature of the curve is for the reason that there are many nodes with a small number of connections and hence, the values close to 1.0 at the start of the curve. As the number of connections is increased, the nodes become very sparsely populated and the curve begins to fall as there are only a small number of well-connected nodes in the network [30]. This theory has been proven and quite a few research efforts later, it can be concretely stated that most real world networks are part of the scale-free community.

Although scale-free property is a ubiquitous phenomenon, it cannot be truly called a universal one. Some structures are better explained using a random network model. The power grid system of Western United States [35, 36] and the graph of company directors [37] seem to have degree distributions with a purely exponential tail.

2.2.3.2 Scale-freeness, an ingrained phenomenon

The motivation of scale-free networks dates back to the renowned experiments demonstrating “six degrees of separation”, conducted in the 1960’s by the renowned psychologist Stanley Milgram [38, 39]. Six degrees of separation is the theory that anyone on earth can be connected to any other person through a chain of acquaintances that has no more than five intermediate human beings. Milgram proposed his theory encompassing only people in the United States of America. Nonetheless, the concept has recently taken the form of an online game where people all over the world try to direct a message from a source to its target by forwarding the electronic mail to persons whom they think have a better chance of acquaintance with the target. On average, the shortest chain is suggested to consist of only six persons.

In 1929, the Hungarian writer Frigyes Karinthy in a short story called *Chains* first proposed this theory. She may not have known the enormity of her proposal at that time. Several decades later, this very notion has been extended to the World Wide Web [40]. The conclusion is that, despite the fact that there is perpetual growth in the number of web pages everyday, every page on an average is only nineteen links away from any other page on the web [40].

2.2.3.3 Mechanism of the scale-free model

The scale-free model is based on two critical principles. Erdős and Rényi [25, 26, 27] had assumed that the full set of nodes in a network would be available before modeling. However, in reality, networks may possibly evolve with time. An example is the World Wide Web where a large number of new web pages are added almost

everyday and these newer nodes have to be taken into account at the time of modeling. The algorithm of the scale-free model is given below [40, 41].

Growth

It has been found that new nodes show an inclination to attach to older nodes than contemporary nodes. Although there are varying theories as to the classification of the “old” node, it is generally accepted that nodes tend to link to other nodes that have two or more times their own number of connections. Hence, the older nodes attain the remarkable ability to acquire new links and grow to become highly connected. These nodes become hubs of the network. This feature is a classic case of *rich gets richer*. Consequently, an evolving network may become a scale-free network.

Preferential attachment

Not all nodes in a network can be assumed equal. A verification of this discovery can be witnessed from the fact that new web pages are likely to carry links to older, well-established nodes like Yahoo and MSN than any relatively new, unknown web page. This partisan treatment meted out to the older nodes (in this example, the older nodes are well-established) also serves to produce a scale-free network pattern. For this reason, it can be seen that the hubs of the network avoid linking with the other hubs as they demonstrate a penchant to bond with the new nodes. Besides, nodes may age and stop receiving links. This saturation causes other mature nodes to become the dominant hubs of the network.

Nevertheless, the networks used for study during this thesis work are non-evolving as the data of the interactions is obtained from specific experimental results that are available in databases. These experimental results may be updated later but at

the time of investigation, the most up-to-date information available has been used to model the systems as a non-evolving network.

2.2.4 Comparison of Random and Scale-free Network

A major advantage of the scale-free network over the random network is the shorter path length needed to get from one node to the other. Another inherent feature of the network is that it is unaffected by the random node failures [42]. The network may lose a few nodes due to random malfunction but the network as a whole is largely intact. Although this resistance to random errors is a positive feature, in the sense that an error in a single power station may not black out several cities, there is another face to the story. This network cannot withstand a premeditated attack on their hubs [42]. A synchronized attack on the highly linked nodes may destroy the network. Such a targeted attack on the World Wide Web may paralyze Internet traffic. As scandalous as the idea may sound, it is easier said than done.

Clustering

Apart from the aforementioned differences, the real-world scale-free network deviates from the random graph model in a clustered way. Real-world networks show strong clustering or network transitivity as shown by Watts and Strogatz [30] and Watts [43], whereas, the random network model does not. A network is clustered if the probability of two nodes being connected by an edge is higher when the nodes have a common neighbor, i.e., another node in the network to which both nodes are attached. This is measured by using a clustering coefficient, C_n that is the average probability that two neighbors of a specific node ‘n’ are also neighbors of one another [44]. In other words, the clustering coefficient is a measure of the interrelatedness of the neighbors of an entity. The mean clustering coefficient for the complete small world network is then defined to be the average of C_n over all nodes in the network.

The clustering coefficient fluctuates between a value of zero and one. An assessment near zero denotes that most of the nodes linked to any given node ‘n’ are not connected to each other, i.e., the system is better modeled using the random network theory. On the contrary, a value near one means that the neighbors of any given node ‘n’ will have a large propensity to be connected to one another. The cliquishness of the network decreases with probability and reaches a maximum value in a regular network. An entity must have at least two neighbors to compute the clustering coefficient. For the many real-world networks, the clustering coefficient is found to be a high value [30]. Nevertheless, for a random graph there is no dominant value of probability of two nodes being connected if they have a mutual neighbor than if they do not. This means that the clustering coefficient, C, for a random graph is $C=p$, where p is the probability distribution [30].

The following section incorporates the network theory with the natural science of living organisms.

2.3 Scale-Free Nature in Biology

The design governing the actual world networks has been a subject of intense scrutiny over the last few years. The graph of interactions in a metabolic network is an important tool to study the fundamental structure of a living being and its continued sustenance in response to most external stimuli.

2.3.1 Metabolic Networks

The network architecture pursued by the metabolic network of microorganisms has been discovered to be of a scale-free nature [23, 45]. The topological properties that have been determined signify that most metabolic networks possess a heterogeneous arrangement of nodes [46]. Five researchers from the University of Notre Dame and Northwestern University Medical School tested 43 biological organisms chosen from eukarya, archaea and bacteria indicative of all the three domains of life [23]. In their depiction, a metabolic network is constructed of nodes representing the substrates that are linked to one another through connections that signify the metabolic reactions. The metabolic reactions are the actual steps that pave the way for the conversion of a specific substrate to its target node (the product of the reaction). Their primary objective was to unearth the topology governing these metabolic networks. They used a graph theoretic approach [47] for the representation of the biochemical reactions to show the probability that a given substrate taking part in a certain number of reactions followed a power law distribution. They provided tangible evidence to corroborate that the 43 organisms tested from the three domains of life adhered to this trend. This homogeny in architecture transpires irrespective of their individual structural blocks or species-specific reaction pathways and hence, a scale-free model could best describe the metabolic network architecture of biological organisms.

2.3.1.1 Consistent scale-free behavior

As referred to in the introductory note to networks, a salient feature of the small world network arrangement is the relatively small number of steps required to traverse from one node to another through the existing links. The biochemical pathway represents the links required for the conversion of one compound to another in a metabolic network. For non-biological networks, the average degree of connectivity of

any node is constant resulting in the increase of the average path length (diameter according to [23]), which is defined as the shortest pathway averaged over all pairs of nodes [40, 48]. Extension of this maxim to the metabolic network signifies that a complex bacterium should have a larger diameter than a simpler one. However, they found that the diameter of the metabolic network for the 43 organisms remained constant. They concluded that as the complexity of the organism increases, the substrates, in turn, link to more substrates to maintain a constant diameter. The total number of substrates in the organism also has a bearing on the network. As the number of substrates in the organism increases, some substrates attain more links, i.e., they participate in more number of reactions [23].

2.3.1.2 Susceptibility to coordinated attacks

Jeong *et al*. [23] further tested the true scale-free nature of the network by investigating the susceptibility of the system to a coordinated attack. During the removal of the most connected substrates from the network, they found that the diameter of the network increased and the network disintegrated into small clusters. Conversely, the network was found to be highly defiant - indicated by the minor variations in the diameter - against random errors simulated by the elimination of substrates in no specific order. The aforementioned work of Jeong *et al*. paved the way for contemporary associates of the research community to build upon their discovery. A number of research efforts have enhanced our knowledge of metabolic networks; nonetheless, the initial contribution of Jeong *et al*. shall remain as one of the most extraordinary exploits of systems biology.

2.3.2 Developments in Metabolic Network Analysis

The year 2003 spawned a few modifications to the publication of Jeong *et al*. [23]. With newer genomes being sequenced with the passage of time, two researchers Ma and Zeng analyzed the metabolic network of 80 fully sequenced genomes. Rather unsurprisingly, they ascertained that the overall network design for all the organisms adhered to a scale-free nature [49].

2.3.2.1 Role of current metabolites

However, a striking difference between this effort and that undertaken by Jeong *et al*. in 2000 is that the latest effort showed some errors that went unnoticed in the prior work. The effort of Jeong *et al*. [23] did not account for the role of the current metabolites during the calculation of the path lengths. Jeong *et al*. determined the diameter of the network for the 43 organisms as approximately the same value. They reported that most of the metabolites in the network could be converted to one another in about three steps. This result is astonishing because of the actual long pathways that are required for the production of many metabolites. Ma and Zeng noted that during their computation, Jeong and co-researchers had included numerous current metabolites as nodes of the network [49].

Due to this reason, an impractical calculation of the path length is generated in many cases. To refer to an example given by Ma and Zeng, let us consider Adenosine triphosphate (ATP) and Adenosine diphosphate (ADP) as nodes and include current metabolites as cofactors in the network. A cofactor is any substance that is required to be present, in addition to an enzyme, to catalyze a particular reaction. Jeong *et al*. [23] had calculated the number of reaction steps needed for the conversion of glucose to

pyruvate as two. This evaluation is biochemically unrealistic as it actually takes nine reaction steps for the conversion of glucose to pyruvate [49]. Similar discrepancies existed due to a variety of other current metabolites.

2.3.2.2 Reconstruction by Ma and Zeng

With the intention of rectifying these errors, Ma and Zeng reformed the metabolic network of 80 fully sequenced genomes. The identification of the current metabolites and their removal from the network also eliminated the connections concerning them. Reversibility of reactions was also accounted and using a graphical representation, they found that variations existed in the network structure of the three domains of organisms [49].

Vitality of network scale

Ma and Zeng discovered that the average path length increases with the network scale. Parasites, which are organisms that live off the host species leading a parallel life inside the host feeding off their own energy, contain lesser number of nodes (node number less than 300) and their metabolic network is not well linked consisting of many small clusters. This results in shorter average path length [49]. These results conformed to the view that parasites have lost a large number of genes during evolution in order to adapt to the changing environments [50]. For networks with a larger scale, an unambiguous relation cannot be deciphered because the average path length diverges greatly even for networks with a similar number of nodes in different organisms.

Significance of complexity

They also found that eukaryota and archaea have a longer average path length than bacteria. Although all organisms have a similar fundamental structure, they demonstrate quantitative diversity in their metabolic network architecture as portrayed by the topological parameters of the network. This diversity echoes the various evolutionary cycles that each organism has undergone over time [49]. The differing topological parameters for the three domains indicate the mixed compactness and centrality of the metabolic pathways [51]. As suggested by Ma and Zeng, a better depiction can be derived by exploring the reactions and pathways to gain a concrete understanding of the biological significance of the basic structural variations that subsist in each organism.

2.4 Protein-Protein Interaction Network of Microorganisms

The study of protein evolution has piloted the ability to identify that some proteins may be related to others. After establishing the fact that scale-free topology is inherent in the metabolic network of biological organisms [23], Jeong and other researchers from the University of Notre Dame sought to utilize this principle to examine the protein-protein interaction network of microbiological organisms, specifically that of *Saccharomyces cerevisiae*.

2.4.1 Direct and Indirect Interactions

Proteins may have either direct or indirect interactions with one another. In a direct or physical interaction, two protein chains bind to each other. Indirect association refers to proteins being a member of the same functional module (e.g., transcription

initiation complex and ribosome). A protein of this nature may not directly bind to another protein. These interactions echo the dynamic state of the cell and their existence depends on the particular environment or developmental status of the cell. However, the coupling of existing and potential interactions together defines the protein-protein interaction network within the genome of a given organism.

The assignment of gene function to newly sequenced genes as part of genome projects would not be feasible without the trappings to recognize the similarity in amino acid sequences. The ability to appreciate the nature of protein evolution allows the biotechnologist to develop novel and technologically useful proteins *in vitro*. The presence of some proteins with a large number of interactions may be due to a specific structural composition that is different from other less connected proteins [52]. This perception is central to the design of long-lasting immunizations and associated drug treatments for human diseases.

2.4.2 Protein Interaction Map of *Saccharomyces cerevisiae*

Jeong *et al*. [53] demonstrated the role of a protein as an element in a network of protein-protein interactions with their publication of the protein-protein interaction map of *Saccharomyces cerevisiae*. The proteins are the nodes and the interactions between them represent the links. The strength of each interaction is supposed to be one if a link exists and zero otherwise.

Architecture of the protein-protein interaction network

It was found that the protein-protein interaction network of *Saccharomyces cerevisiae* follows a heterogeneous scale-free architecture [54]. Two separate

simulations were performed to verify the effects of random and coordinated attack on the interaction network. The computational removal of the well-connected yeast proteins caused the diameter, i.e., the average shortest distance between all pairs of proteins in the largest cluster of the network to increase steadily [53]. On the contrary, the removal of arbitrarily chosen yeast proteins did not affect the topological parameter of the network [53].

Corroboration of results

This resistance against random exclusion has been shown to be in agreement with the results from mutagenesis experiments [53]. They recognized that the organism is able to withstand the eradication of an extensive number of sparsely connected proteins. They deduced that if topology is responsible for this tolerance, then the essentiality of the proteins should substantiate their finding. They correlated information from the interacting proteins with the phenotypic effects of their removal from the yeast proteome. The lethality of a protein depends on the number of connections it shares in the protein-protein interaction network. The proteins with a large number of connections are found to be highly essential and their removal disturbs the network topology, proving lethal to the network. Although there are several proteins with lesser number of links, their elimination did not produce such adverse consequences. Jeong *et al*. [53] concluded that the most connected proteins, which play a central role in the network design, are three times more essential than the proteins that interact with only a few other neighbors.

2.5 Progression to Higher Strata of Organisms

The first foray made by Jeong and fellow researchers with the publication of the protein interaction map of *Saccharomyces cerevisiae* has sown the seeds for developing

newer research ventures using their breakthrough. This thesis work has been born of the inspiration acquired from the efforts over the last few years.

2.5.1 Complex Biological Organisms

All endeavors to study the network design of biological organisms have been restricted to simple species [53, 55]. A valid reasoning behind this predisposition is that the lesser the number of genes involved, the fewer are the complications to be encountered during the study. Research labs around the world furnish abundant datasets on a regular basis, but the specific analyses of these datasets are underdeveloped due to the compound environment of the biological interaction networks. The nature of research is multifaceted due to the fact that even the simplest unicellular organisms have more than a few hundreds of genes transforming them into a compound life form from the perspective of a data analyst. Nevertheless, the shift of focus to more complex beings would provide a better understanding of the structural properties prevailing in the higher echelons of nature's offspring.

With a vision to gain an advanced perception into the secrets of nature, the multi-cellular organism *Drosophila melanogaster* is selected as the test species for this research assignment. The preference of *Drosophila melanogaster* to other complex organisms for this venture has its justification. The following material describes the key factors that motivated the use of this organism as the test species for this study and describes the general characteristics of the organism that makes it a valuable research class.

2.5.1.1 Importance of *Drosophila melanogaster*

The *Drosophila melanogaster* (black-bellied dew-lover), a dipteran (two-winged) insect, is one of the most precious biological organisms [56]. It belongs to a species in the kingdom “Animalia” of the domain “Eukarya”. It is a tiny, common fly found near unripe and perished fruits and is deemed as a pest in virtually all areas. In modern biological literature, it is often simply called *Drosophila* or fruit fly (common name). Charles Woodworth is credited with being the pioneer in breeding *Drosophila* in quantity and for suggesting that the organism might be used for genetic research [57]. In 1910, fruit flies helped Thomas Hunt Morgan accomplish studies on heredity that placed the small fly in the vanguard of genetic research and still serves as one of the most invaluable research specimens on earth.

2.5.1.2 Habitat

Drosophila melanogaster lives in a wide range of habitats. Its native habitat includes those in the tropical regions but the common fruit fly has been introduced to nearly all temperate regions of the world. The only aspect that limits their habitat is temperature and availability of water. As the meaning of the scientific name implies, it requires moist surroundings to flourish. The development of the fly is extremely dependent on temperature and the adult cannot withstand the colder temperatures of high elevations. Food supply is also limited in these locations and therefore, it cannot survive in cooler climates. In temperate regions where human activity has introduced them, the flies seek shelter in the colder winter months. The *Drosophila* can be found in fruit cellars or other available synthetic structures with a large supply of food [57].

2.5.2 Research Significance of *Drosophila*

The *Drosophila* is being used as a model organism in biological research for nearly a century due to its similarity with the human proteins, some of which may serve as potential drug targets. It lends itself well to behavioral studies. Scientists have discovered an identifiable match between the genetic code of fruit flies and over 60% of known human disease genes. Moreover, about 50% of fly protein sequences are believed to have mammalian analogues. Hence, *Drosophila* is being used as a genetic model for various human diseases including Parkinson's and Huntington's diseases. Some of the features that make the *Drosophila* a versatile research specimen, predominantly in genetics and evolutionary biology are given below [56, 58, 59]:

- Flies are small in size (about 3 mm in length and 2 mm in width) and can be easily grown in the laboratory
- They can be anesthetized easily with simple equipment
- Flies have a short generation time and do well at room temperature with a high productivity (females can lay about 500 eggs in 10 days) (Figure 2.7)
- The care and culture requires little equipment, is low in cost and uses less space even for large cultures
- The study of these proteins may provide knowledge about the possible development of remedial measures for human diseases such as heart disease, cancer or diabetes mellitus
- *Drosophila* are sexually dimorphic, making it easy to differentiate the sexes
- The genetic transformation techniques have been available since 1987
- Its genome has been sequenced in 1998

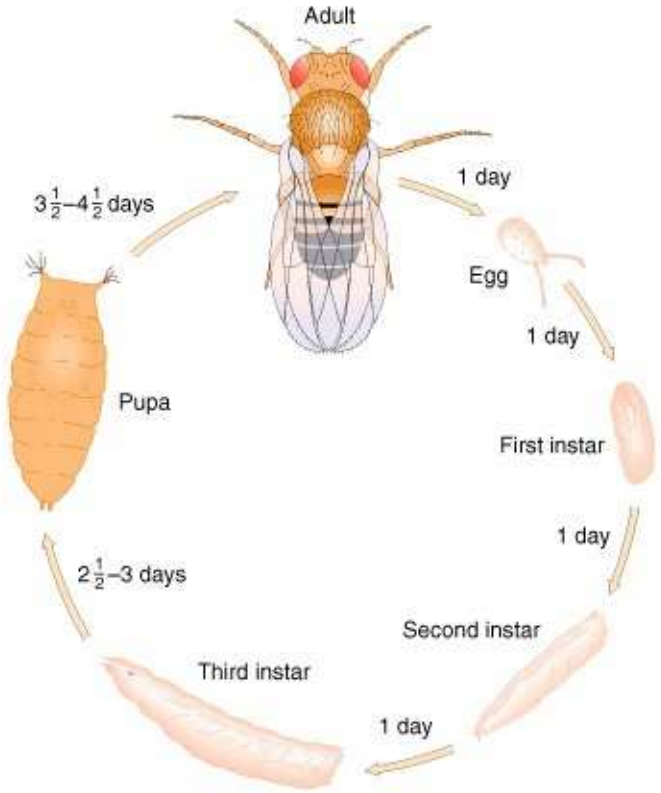

Figure 2.7: The complete life cycle of *Drosophila melanogaster* [60]

2.5.3 Research Objectives

The study of the protein-protein interactions in biological organisms involves a few drawbacks. Firstly, the biological network is enormously complicated. Secondly, the association between proteins is extremely transitory in nature. The ability to overcome these setback factors can generate a vivid description of the interacting protein partners contributing to productive research in drug therapy.

One of the major goals of this project is to study the network topology of the multi-cellular organism *Drosophila melanogaster* at varied levels of complexity. The priority is to gain an understanding of the protein interaction map of the organism [61], to study the network architecture inherent in it and to identify the essential proteins of the protein-protein interaction network [53]. A computation of the topological parameters of the network should assist in this goal.

The field of metabolic engineering is a novel approach to perceive and use metabolic processes. Metabolic engineering seeks to channel resources to intentionally modify the metabolic pathways found in an organism in a fruitful way. This alteration would go a long way in facilitating the understanding and utilization of cellular pathways for chemical transformation and energy transduction. The new awareness would enable the ability to alter biological pathways to produce organic alternatives for less enviable chemical processes, allow for larger agricultural production and provide better understanding of the metabolic basis for some medical conditions that could support in the progress of discovering new remedies [62].

The second phase of this project involves the expansion of this principle to the metabolic network of the organism. The analysis of the network architecture and the identification of the vital metabolites of the network would provide a definitive description of the significance of metabolites to the interaction network. The study of variation in the topological parameters can offer an understanding about the degree of importance of the metabolite being eliminated. This thesis work shall also furnish a comprehensive resource detailing the various enzymes, genes and reactions involved in each metabolic pathway of the organism.

CHAPTER 3

METHODOLOGY

The sequencing of newer genomes has given rise to extensive biological data being available today and the burgeoning career of bioinformatics owes its gratitude to the field of genomics. The surge of interest to utilize this biological data available in public databases has impelled the demand for bioinformaticians in scientific research. Correspondingly, bioinformatics approach forms the foundation for the following methods.

3.1 Analysis of the Protein-Protein Interaction Network of *Drosophila melanogaster*

3.1.1 The Protein Interaction Map

The latter part of 2003 saw the publication of the protein-protein interaction map of *Drosophila melanogaster* through a united effort of Curagen Corporation, Wayne State School of Medicine and Yale University School of Medicine. A brief outline of their work would prove beneficial in understanding the later stages of this thesis.

The two-hybrid analysis system was used to detect the protein-protein interactions of the organism. A concise general description of the two-hybrid system [63, 64, 65] is given below.

The two-hybrid assay to be used consists of two fusion proteins. The objective is to segregate proteins that interact with a bait protein from any organism. The only requirement is that the complementary DNA (cDNA) library that is to act as the possible prey should be from the same or a closely related organism. The arrangement includes two plasmids, one containing the gene for the bait protein and the other holding the cDNA library. A plasmid is an autonomous, circular, self-replicating DNA molecule that carries only a few genes. Once inside the cell, each of these plasmids will express the genes as proteins. The system is built so that any protein encoded by the cDNA library that binds the bait will cause transcription of a reporter gene acting as a pointer to the potential protein-protein interaction.

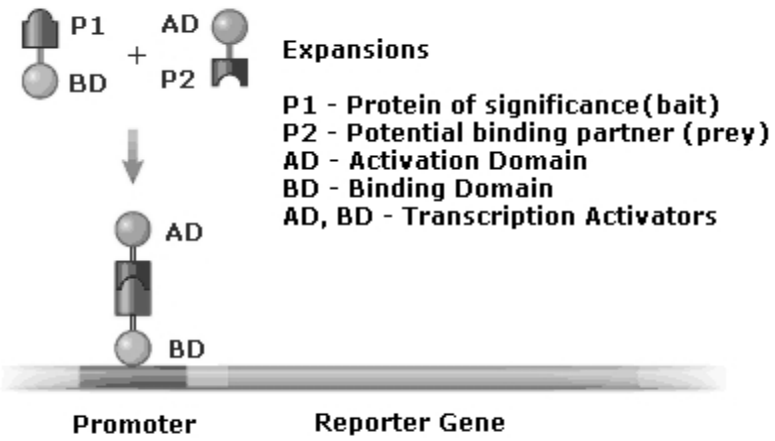

Figure 3.1: A detailed explanation of the two-hybrid system [66]

The cDNA of the protein of significance (P1) or bait is built to include a DNA binding domain (BD). Its potential binding partner (P2) or prey is fused to the coding sequence of an activation domain (AD) (Figure 3.1). If protein P1 interacts with protein P2, a whole and functional transcriptional-activating factor is reconstituted by the fusion of the AD and the BD. This reconstituted factor will stimulate transcription of a reporter gene or genes that is/are under the control of that transcription factor [67]. A reporter gene is one whose protein product may be identified and quantified with consummate ease. Thus, the quantity of the reporter gene formed can be used as a gauge to determine the interaction between the significant protein and its potential binding partner.

Giot *et al*. [61] performed a statistical analysis on the protein-protein interactions detected using the two-hybrid system.

Confidence scores

Giot and co-researchers used the general linear model of the R statistics package [68] to produce a set of confidence values based on the probability of occurrence of each interaction [61]. Using this technique, they identified a set of interactions that included both high confidence and low confidence values.

High confidence interactions

The high confidence interactions are those interactions that have already been published and acknowledged as accurate. Interactions pertaining to two proteins from the same complex have also been integrated into the high confidence dataset [61].

Low confidence interactions

Low confidence interactions are those unlikely to appear *in vivo*. These interactions are obtained during the application of experimental techniques to the proteins but, in reality, they have a very remote possibility of occurrence [61].

Giot *et al*. [61] needed a demarcation value to distinguish between the high confidence and low confidence data. Using the generalized linear model of the R statistical package [68], they obtained a set of confidence scores in the range 0 and 1. The cut-off threshold between the high and low confidence interactions was set at 0.5 to classify the data into two categories. They confirmed the threshold by studying the Gene Ontology [69] annotations of interacting proteins. The confidence score has been established to have a powerful correlation with the depth in the hierarchy at which two proteins share an annotation. The correlation was found to rise for confidence values of 0.5 and higher, thereby justifying the use of 0.5 as the threshold value [61]. The list of interactions with their confidence values is available on the web portal of Curagen Corporation [70].

3.1.2 Preliminary Steps

The protein-protein interaction dataset has to be subjected to a preliminary modification procedure before it can be used for the calculation of the topological parameters. The protein-protein interaction dataset for *Drosophila* is compiled from the *Drosophila* Interaction Database of Curagen Corporation and collaborators from where 41068 protein-protein interactions are retrieved. However, as mentioned earlier, the dataset includes high confidence and low confidence interactions in addition to several redundant and self-interactions.

Study of the data

The high confidence (confidence scores higher than 0.5) protein-protein interaction pairs are selected for this work as they have a greater support for occurrence. A set of 4591 proteins involving 9334 high confidence interactions, exclusive of self-interactions, is used for the protein-protein interaction network. This dataset contains both unidirectional and bi-directional links. A careful study of this dataset reveals that two specific proteins stand apart. Among the proteins that have a one-directional interaction, the two proteins namely CG4039 and CG12918 only have incoming edges. There are several other proteins with unidirectional links but all those proteins have outward-bound edges.

Binary network

The strength of each protein-protein interaction is unknown. The usage of the confidence scores as the weight of the interaction links does not supply any material denotation to them. Several statistical concepts and techniques were analyzed to offer a weight-based interaction study. Nevertheless, it has been decided that the provision of a weight-based system is not affordable at this time. Hence, quite a few futile attempts later, it has been decided that the study would be conducted by treating the system as a binary network. The weight of the edges linking the two proteins is assumed as one when an interaction exists. Conversely, in the absence of an interaction the influence is understood to be zero to produce a binary network.

3.1.3 Topology of the Protein-Protein Interaction Network

The topological parameters of the interaction network offer an insight into the intrinsic architecture of the system. Topological parameters like degree, *k*, and probability distribution, *p*(*k*), of each node can be calculated for the protein-protein interaction dataset. The degree of connectivity '*k*' in a network refers to the number of connections (interactions or edges) contained by a particular node in the network. The degree distribution *p*(*k*) gives the probability that a selected node has exactly '*k*' links in the network. A plot of the probability distribution against the degree is used to determine the architecture of the network. A linear correlation on a log-log scale establishes that the distribution follows a power law.

3.1.3.1 Path length of the network

Analysis of the network involves the calculation of the number of steps required to traverse from one node to another in the protein-protein interaction network. The path length or pathway is defined as a means to quantify the number of links needed to navigate from one node to another within the network. This study would entail a better perception of the various path lengths available to a node and assist in the determination of the shortest path length between any pair of nodes. The evaluation of the shortest path length warrants the use of a mathematical algorithm.

Shortest path length analysis

A universal feature of most real-world compound networks is that nearly all pairs of nodes can be connected by only a few links. A major task in network analysis is to locate the potential connection path lengths between any two proteins in the

interaction network. There may be numerous pathways between two proteins, but the shortest pathway is of greatest interest for network assessment. In order to calculate the shortest distance between two given nodes, a graph search method called ‘breadth first search’ is utilized.

3.1.3.2 Breadth first search algorithm

Principle

The search is performed based on the following principle [47, 71]. In this algorithm, every edge is assumed to have homogeneous weight, i.e., every edge has identical influence. A list containing all nodes in the network is investigated, beginning with any node as the source, to resolve its connections in a series of layers. It is assumed that the search starts at a node ‘n’, which is said to be unexplored at that time. After ‘n’ is visited, all unvisited nodes adjacent to ‘n’ are subsequently traversed. The exploration of the node ‘n’ is said to be complete when the algorithm has visited all the unexplored nodes adjacent to it. The node ‘n’ is marked as explored and is not visited again. The newly traversed nodes have not been investigated and are placed onto the end of the list of unexplored nodes. The first node on this list of unexplored nodes is the next to be explored. Examination continues until all nodes have been searched. If a node has a multitude of untouched neighbors, it would be equally acceptable to visit them in any order. One of the best methods to employ the search would be to visit them in the same order as the stored adjacency list of ‘n’. The working of this algorithm results in the formation of a tree.

Illustration of working of the algorithm

Consider graph G containing a set of nodes V linked by edges E (denoted as G {V, E}) as shown in Figure 3.2a.

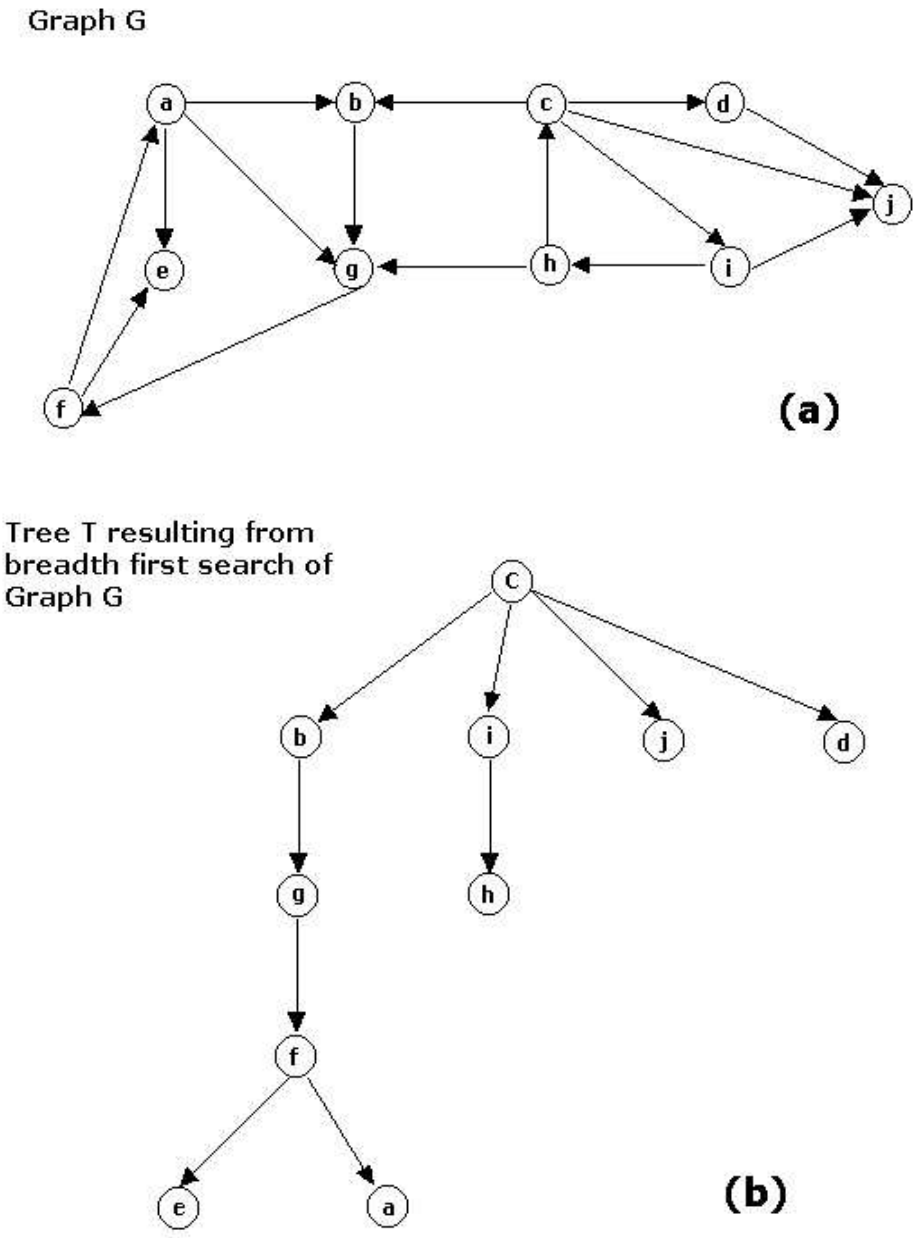

Figure 3.2a & b: Explanation of breadth first search algorithm (graph and tree illustration)

Considering a two-node set ‘g’ and ‘f’ from the tree, each edge can be said to represent from a node visited earlier to the one traversed later (Figure 3.2b). Subsequent

successive edges upwards can only be terminated at ‘c’ (which being the source node in this example has no edge going upward from it).

Therefore, every node in the tree T has a path to ‘c’. A tree is just a connected and acyclic graph. This means that T is at least a connected sub graph of graph G. In any cycle, irrespective of the edges, one direction is upward and the other downward. Nevertheless, in T, each node has at most one upward edge so that T can have no cycles. Therefore, T, in reality, is a tree known as a breadth first search tree. T is also a spanning tree, i.e., if the graph is connected (every node has some path to the source ‘c’) then every node will occur somewhere in T. This can be proved by induction on the length of the shortest path to ‘c’. If a node ‘f’ has a path length L, (i.e., f-g-…-c), then node ‘g’ would have a path length given by (L-1). When ‘g’ is visited, the edge f-g would have been traversed, and if ‘f’ were not already in the tree, it would be added.

The affirmation that the nodes are in this order can be obtained by induction on the layer number. By the induction hypothesis, breadth first search would register all nodes at layer (L-1) before those at layer L. Similarly, all those nodes at layer L are aligned before those at layer (L+1). Every edge of G can be classified into one of three groups. Some edges are in T themselves. Some may connect two nodes at the same layer of T. In addition, the remaining ones connect two nodes on two adjacent layers. It is not possible for an edge to skip a layer. Breadth first search of graph G corresponds to some kind of tree traversal on T. The traversal goes a layer at a time, left to right within a layer (where a layer is defined in terms of distance from the root of the tree). Therefore, the breadth first search tree is a shortest path tree starting from its root node. Every node has a path to the source with path length equal to its level (following the tree itself) and it is not possible for a path to pass over a layer. Hence, in actuality, this is the shortest path.

3.1.3.3 Determination of topological parameters

The Boost Graph Library that contains the breadth first search algorithm is used for the calculation of the shortest path length for every pair of accessible proteins [72]. The shortest path length of the proteins is then employed to compute the average path length and diameter of the protein-protein interaction network.

Average path length

The average path length will provide a description of the number of pathways it would require, on average, for one protein to reach another accessible protein in the interaction network.

Network diameter

Another topological parameter of considerable interest is the network diameter. The diameter of the network is the length of the longest path length among the shortest path lengths computed from the graph search method [47]. This parameter presents information about the length of the longest pathway that is utilized in the interaction network.

3.1.3.4 Robustness and susceptibility of the network

An important task in probing the network is to test its true scale-free character. As the network architecture is of a scale-free nature, it has to be explored for its toughness against random failures and vulnerability to coordinated attacks [42]. The

proteins are ranked based on their degree in descending order. The top 3% of the highly connected proteins are removed sequentially to feign a targeted attack and the effect on average path length and diameter of the protein-protein interaction network is studied. On the other hand, 3% of the proteins are removed at random creating an unintended failure of nodes and the effect on the two topological parameters is calculated again. A graphical plot is used to demonstrate the effect of removal of proteins on the network topology. The molecular function of the hub proteins is retrieved from biological databases and used to identify any possible association to the distinguishing behavior exhibited due to the targeted elimination of those proteins.

The description of the computation of the topological parameters and the study of the robust and susceptible nature of the network with their respective inferences is given in Chapter 4.

3.2 Analysis of the Metabolic Network of *Drosophila melanogaster*

This part of the project uses the same analytical techniques as that exercised for the protein-protein interaction network. Hence, it is assumed that it is dispensable to reiterate through the steps. Any new information that merits mention, to facilitate the understanding of the methodology of analysis of the metabolic network, is given in the following sections of the chapter.

3.2.1 Investigation of the Significance of the Metabolites

KEGG database

The public access to genomic databases serve to analyze, understand and portray the latent patterns entrenched in biological networks. The information retrieved from the KEGG (Kyoto Encyclopedia of Genes and Genomes) database that is a privately owned portal based in Japan is recognized as the basis for this investigation [73, 74].

The KEGG database offers a wealth of genomic assets. The database provides gene files that deliver information about the gene involved in a specific organism and the corresponding enzyme denoted by its name and Enzyme Commission (EC) number. The set of files for *Drosophila melanogaster* are recovered in a rudimentary form. Separate files are retrieved for each metabolic pathway. The reaction files recovered from the KEGG database contain information about the reaction identifier (ID), the reaction involved and the enzyme catalyzing the reaction (denoted by EC number). The files available are based on the type of metabolic pathway.

The major objectives of this work include:

- Construction of a resource, based on the metabolic pathways involved in *Drosophila* using the information retrieved from KEGG
- Analysis of the compounds that play an important role in the metabolism and to determine the network architecture
- Study of the effect of removal of metabolites, in sequence and at random, on the topology of the network

3.2.1.1 Construction of a metabolic pathway based resource

The compilation of the information from the reaction database presents a list of 1017 reactions. The reaction represented by the compound indices such as C00001, C00024, C05345, etc., are used for the study as it provides easier management than the actual names of the compounds themselves. The gene files provide a set of 2969 genes. Incidentally, it should be noted that all the reactions that occur in this organism are reversible. The redundant genes that code for the same enzyme are removed. Using the EC numbers as a correlating factor, the genes and the reactions are listed classifying them according to their metabolic pathways.

3.2.1.2 Identification of the role of metabolites

One of the principal goals of this endeavor is to classify the significant compounds that are involved in the metabolic pathways of *Drosophila melanogaster*. It is of tremendous importance to gain an understanding of the compounds that are vital to the metabolic network. The knowledge about these compounds will lead to a better diagnosis on the obscure web of metabolic networks. To ensure the accomplishment of this objective, the substrates and the products involved in all reactions are isolated. Three separate lists are generated, namely - metabolites acting as substrate, metabolites acting as product and metabolites occurring as both substrate and product in the reaction. Computations performed on the datasets yield an insight into the fact that approximately one third of the metabolites are present in each category.

Graphical plots

The outgoing links, represented by C_{out}, indicate the substrate metabolites. In contrast, the incoming links signifying the metabolites that occur as product in the reaction are designated by C_{in}. A study of the probability distribution of the occurrence of each metabolite will provide a comprehensive view of the metabolic network architecture.

Probability distribution charts

The probability distribution of the metabolites is plotted as $p(k)$ versus 'k' where $p(k)$ is the probability of finding 'k' number of connections. Three different graphs are plotted using the incoming degree, the outgoing degree and a combination of all the degrees to examine the nature of distribution of metabolites in each category.

Frequency distribution of compounds

A frequency distribution histogram is used to substantiate the nature of the network architecture and to determine the impact of the compounds that play a part in the metabolism of the organism.

3.2.2 Examination of the Metabolite-Metabolite Interaction

An investigation of the interaction of the metabolites that may act as substrate or product or both can be utilized to figure out the topological parameters of the network. A sample of the reaction involved in the metabolism is

C05125+C00011<=>C00068+C00022 (2)

The connections (interactions) for this reaction are built as C05125-C00068, C05125-C00022, C00011-C00068, C00011-C00022, C00068-C05125, C00022-C05125, C00068-C00011 and C00022-C00011. On a similar basis, a list of connections is organized for all the reactions. If the reaction contains more than a single mole of compound, e.g., 2C00001, the reaction is modified by substitution of the compound index as (C00001+C00001). The strength of each association is homogeneously maintained as one to produce a binary network.

The downside of current metabolites

The dataset generated above cannot be used *per se*, as the list of connections prepared contains a large number of *current metabolites* [49]. Current metabolites are those cofactors in biochemistry such as ATP, ADP and hydrogen ions (H^+) that should be removed from the dataset before metabolic network analysis. Cofactors are normally used as carriers for transport of electrons and other functional groups to facilitate the catalysis of a reaction [75]. The current metabolites may be explained as being analogous to an external metabolite that takes part in more than a few reactions but does not occur in pseudo steady state in a sub-network [76]. A metabolite is external, if it is well buffered like water, ATP, etc. The fast-paced nature and high yield of metabolites during reactions has resulted in a pseudo steady state assumption that on longer time scales, the concentration of metabolites and the rate of reactions are stable. This condition guarantees that none of the metabolites are produced or consumed in the overall stoichiometry [77, 78].

The calculation of the least number of steps required to get from one compound to another (shortest path length) using this set of data would provide an inaccurate insight into the topology of the network [49]. The current metabolites should be removed before the calculation of the topological factors, as their inclusion would

generate fallacious parameters. Nevertheless, the deletion of the current metabolites and their possible connections cannot be done *per se*. Some of the current metabolites may be primary metabolites acting as either substrate or product. A primary metabolite is essential for regular growth and reproduction. Primary metabolites are those metabolites that occur as the first compound in the substrate or product sequence of a reaction. Such reactions warrant exclusion [49]. Consequently, those reactions in which the current metabolites are present as a primary substrate or product are permitted to be part of the network. The remaining connections that do not involve the current metabolites as primary metabolites (either as substrate or product) are deleted and the links between the substrates and products is reconstructed. This strategy and the removal of redundant connections cause the number of links to be reduced to 3326. This restructured dataset is used for the calculation of the shortest path lengths of the network.

3.2.2.1 Computation of the topological factors

The shortest path length of this metabolic network is calculated using a graph search algorithm. The breadth first search method is used for this purpose [47]. A detailed description of the principle and an illustration of this algorithm are provided in Section 3.1.3.2. The Boost Graph Library that includes the breadth first search algorithm is used for the computation of the shortest path length for every pair of reachable metabolites [72]. The shortest path length of the metabolites is employed to calculate the average path length and the diameter of the metabolite-metabolite interaction network.

3.2.2.2 Simulation of random and coordinated attack on the metabolic interaction network

The metabolic network is examined for its tolerance to random failures and defenselessness against sequential errors [42]. This study will provide a better suggestion about the effects of synchronized and random attacks on the metabolic network. The metabolites are sorted based on their degree (inclusive of incoming and outgoing edges) in declining order. The top 5% of the most connected metabolites are removed successively to simulate a premeditated attack and 5% of the metabolites are removed in an unsystematic fashion imitating an accidental failure of nodes. The outcome on the average path length and the diameter is studied for both cases. A graphical plot is utilized to highlight the effects of eradication of metabolites on the overall topology of the network.

3.2.3 Central Metabolism

The core of metabolism lies in the production of energy. The central carbon metabolism plays a pivotal part through its important conduits of glycolysis, Tricarboxylic Acid (TCA) cycle and pentose phosphate pathways. These pathways portray a critical role in the metabolism of any organism by providing links to various reactions. The compounds involved in the central metabolism coordinate phosphate (energy), carbon, nitrogen, and redox metabolism. They enable the production of energy required for metabolism and aid in the production of other primary and secondary metabolites. The number of steps (reactions) required for compounds to reach another compound involved in the central metabolism serve as an indicator of the degree of possible conversion to metabolites of these key pathways. The degree of interconnectivity illustrates the synchronization of metabolism around these compounds and it is likely that metabolic regulation will revolve around the careful organization of

these metabolites. The shortest biodegradative pathway (least number of steps) can also present information about the number of steps needed for a compound to be biologically degraded [79]. Hence, this quantity is computed for the metabolic network using the breadth first search method. The results obtained during the various stages of this exploratory study are described in Chapter 4.

CHAPTER 4

RESULTS AND DISCUSSION

The results are given in two sections; sections 4.1 and 4.2 provide the analytical results of the protein-protein interaction network and the metabolic interaction network of *Drosophila melanogaster* respectively.

4.1 Protein-Protein Interaction Network

A scale-free network is distinguished from the random network design by a power law distribution of connectivity instead of a Poisson distribution. The graphical plot involving the probability of allocation of nodes and their definite number of links should produce a decreasing function for a network that is scale-free.

4.1.1 Probability-Degree plot for the Protein-Protein Interaction Network

The probability distribution plot of the number of proteins having a definite number (k) of interactions is an excellent indicator of the type of network design inherent in the protein-protein interaction network. As shown in Figure 4.1, using the high confidence interaction dataset, a plot of degree distribution $p(k)$ versus the interactions 'k' on a log-log scale produces a linear correlation hinting that the degree distribution follows a power-law.

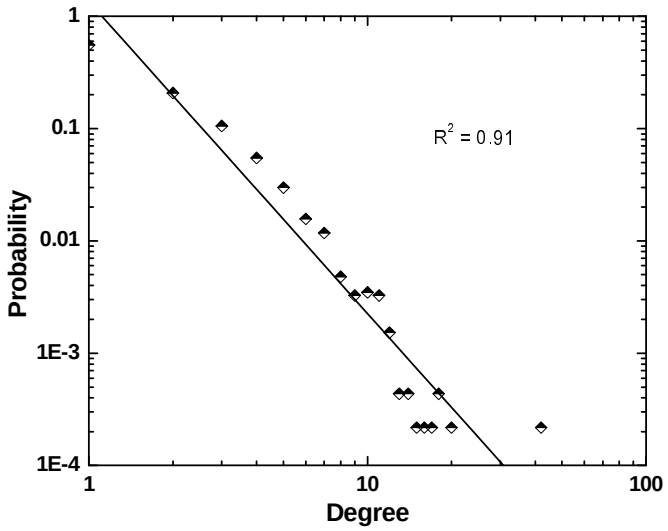

Figure 4.1: The probability distribution-degree plot for the protein-protein interaction network

This result is closely in accordance with the plot shown by Giot *et al*. [61] for the protein interaction map. As the degree of the network increases, the probability of finding a protein with that specified number of interactions begins to fall to a low value. This confirms the salient feature of scale-free networks, i.e., there are several nodes with a low degree and a few dominant nodes (hubs) with a high degree. This result verifies that the protein-protein interaction network of *Drosophila* follows a scale-free architecture.

To exercise some statistical information usage, there are 2549 proteins having a single interaction that constitute more than half (~55%) of the total proteins in the high confidence dataset. Only 1% of the proteins have ten or more interactions. These proteins having a higher number of interactions act as the hubs of the protein-protein interaction network. The exponent γ in the power-law distribution ($y = 1.3511x^{-2.7763}$) is

found to be 2.78. This value is in agreement with those obtained (2.0-3.0) for other networks like the World Wide Web, citations of scientific articles and the Internet that follow a scale-free architecture [31]. The mean degree, i.e., the average number of interaction partners present per protein, of the high confidence dataset is found to be 2.03. This signifies that on average, every protein that is part of the system interacts with two other proteins in the network.

4.1.2 Topological Parameters

The other topological parameters of interest are the average path length and the network diameter. The average path length for the network is calculated as 9.42 and the network diameter is found to be 27 for the high confidence interaction dataset.

The relatively short path lengths typify a small world network. As the network architecture is of a scale-free nature, it has to be tested for its resistance against random failures and vulnerability to coordinated attacks. The proteins are ranked based on their degree in decreasing order for this study. The exclusion of a particular node by chance or by deliberate measure can give rise to contrasting consequences. The severity of these consequences depends entirely on the node that is removed. With the rationale of examining the network for its weakness against sequential attacks and its tolerance to random failures, a milieu simulating a targeted assault where the well-connected proteins (top 3%) are removed in sequential order of their degree is replicated. The second part of the simulation creates a random node failure by removing proteins in an arbitrary fashion. The upshot of these simulations is summarized in the following sections.

4.1.3 Alterations due to Simulated Errors on the Network Arrangement

The consequences of the removal of well-connected proteins and those in a random manner are illustrated as a plot of the corresponding topological parameter versus the number of proteins removed. The ensuing correlations in Figures 4.2 and 4.3 for the 3% (~138 proteins) sequential removal of the most connected proteins, in descending order of their degree, emphasize the critical nature of the highly connected proteins.

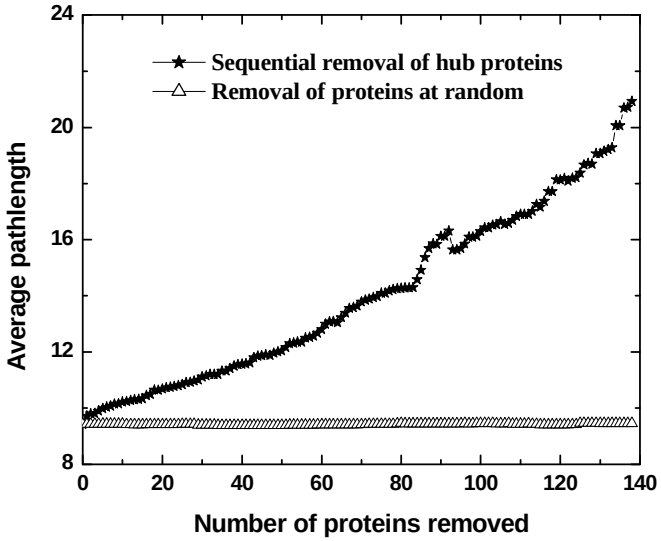

Figure 4.2: Effect of sequential and random removal of proteins on the average path length of the network

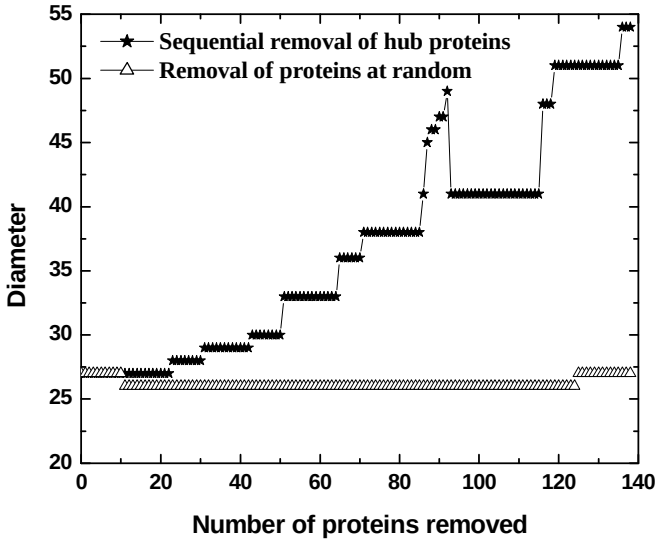

Figure 4.3: Effect of sequential and random removal of proteins on the diameter of the network

For the 138th protein eradicated, the average path length and diameter of the network are determined to be 20.93 and 54, respectively. In comparison with the original topological parameters (average path length = 9.42 and diameter = 27) of the arrangement, it can be observed that the average path length and the diameter of the network are doubled. This offers evidence of the vital character of the hub proteins.

The elimination of hubs changes the topology of the network corroborating their direct relation to the topological parameters that prove lethal to the system. The boost in the average path length signifies that the shortest path length required by a particular protein to get to another protein has increased. This implies that it takes more number of steps to get from a specific protein to its target protein, thereby disturbing the efficient, innate path that had been utilized before the assault on the hubs of the network. The doubling of the diameter highlights the formation of small, secluded clusters of proteins that is very different from the original compact system design. An

interesting feature of the plots is a decrease in both topological parameter values on the removal of CG6998 (Gene *ctp*), a well-linked protein. Figure 4.3 illustrates that the diameter remains consistent for the exclusion of some proteins and then increases. This sequence replicates for the entire 3% exclusion. The removal of Gene *ctp* (when diameter increases to 49) causes the failure of the actual connecting path but the subsequent elimination of other hub proteins result in the mechanism being able to salvage a shorter path. This is a one-time occurrence and typically, the organism is not able to find a shorter recourse to its target protein, as exposed by the escalating nature of the parameter values, once the intrinsic path is lost.

On the contrary, the simulation of a random node failure does not demonstrate any drastic alterations in the topological parameters of the network when 3% of the proteins are removed. The topological parameters show a minor variation and as the plot indicates, it does not affect the innate topology of the system.

4.1.4 Graphical View of the Protein-Protein Interaction network highlighting the hubs

A graphical representation of the protein-protein interaction arrangement should provide for a better interpretation of the system. Figure 4.4 shows a graphical view of the original network constructed using Pajek network analysis tool [80].

The most connected proteins are represented by the nodes in red indicating the lethal nature of those proteins. The clustered environment is due to the large number of proteins (4591) that are represented on the image. The connecting lines denote the interaction (edge) between any two proteins. The removal of proteins causes the network to shrink but the mode of contraction shows the lethality of the hubs. The

sequential removal of hubs causes the network to be break into fragments of smaller clusters. On the contrary, the random removal does not lead to the fracturing of the network and the overall topology remains relatively intact.

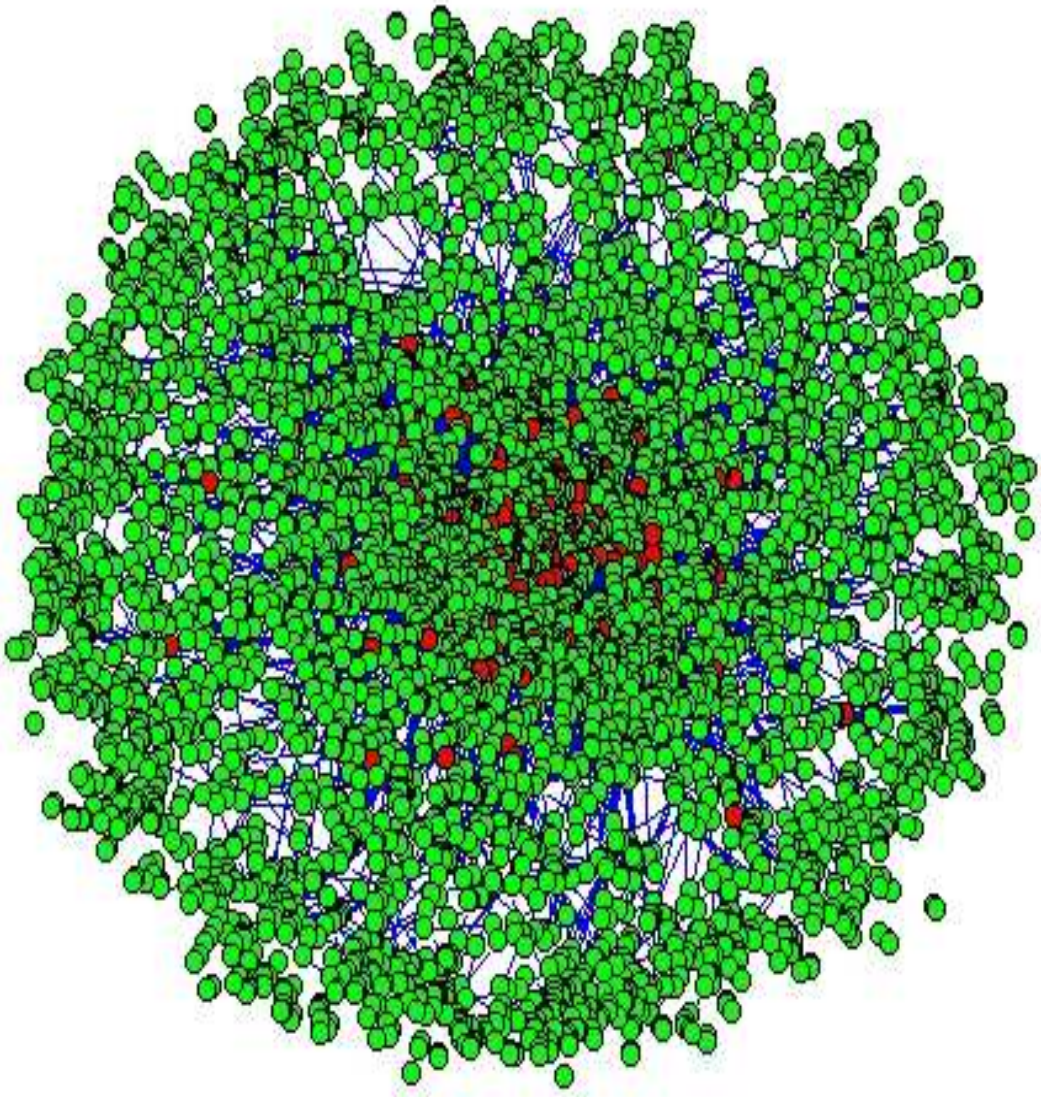

Figure 4.4: Graphical view of the original protein-protein interaction network

4.1.5 Study of the Molecular Functions of the Hubs

The top 3% (~138) of the highly connected proteins are examined to study their functions and their respective roles in biological pathways. Information retrieved from FlyBase provides the molecular function of these proteins [81]. It is found that 62 of these proteins have not been annotated yet. The study of the remaining 76 proteins with an annotation using KEGG Orthology [73, 74] reveals that eight proteins cannot be classified to a specific pathway. Analysis shows that 21 proteins play a part in metabolism, 10 proteins are involved in cellular processes, 17 proteins aid in environmental information processing and 20 proteins are responsible for genetic information processing. Among the proteins involved in metabolism, 13 proteins are components of the central metabolism pathways. Hence, it can be seen that the molecular functions of the annotated hubs are distributed over a wide array of processes essential for the functioning and sustenance of the organism.

4.2 Metabolic Network Analysis

The pathway-based resource is compiled using the information recovered from the KEGG database. The result is a list of the metabolic pathways with each pathway data enumerating the enzyme catalyzing the reaction (represented by EC number), gene that codes for the enzyme (denoted by gene name) and the reaction (designated by compound indices) that it is involved. It is found that there are 90 metabolic pathways involved in the metabolism of *Drosophila melanogaster* (*please contact the author for additional information*).

4.2.1 Determination of the Network Architecture

The plot of the architecture of the metabolic network strongly resembles that of the protein-protein interaction network. The probability distribution of the compounds, acting as substrate (out degree) and as product (in degree) (Figure 4.5a & 4.5b) and of all compounds (Figure 4.5c), follows a power law distribution.

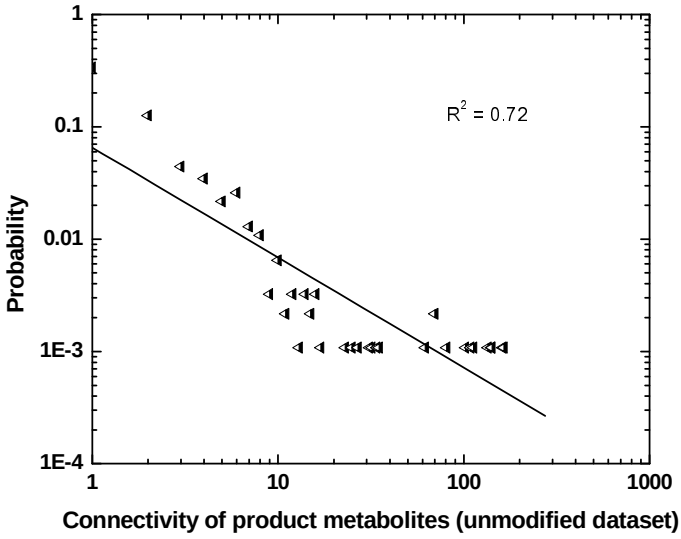

Figure 4.5a: Probability distribution of connectivity of product metabolites

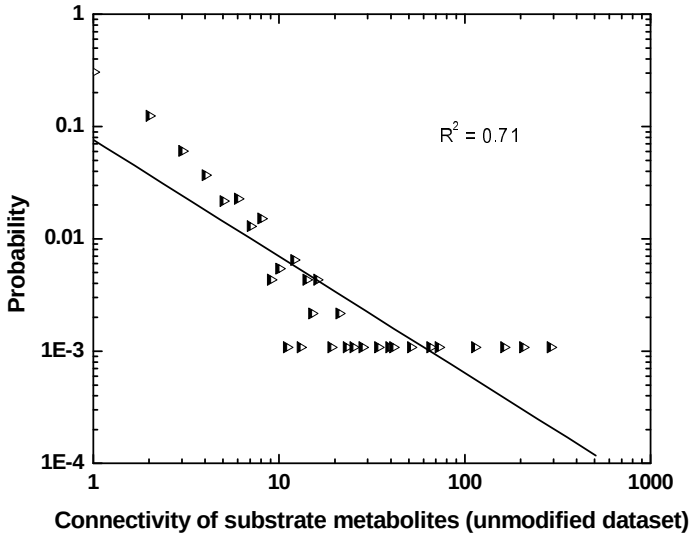

Figure 4.5b: Probability distribution of connectivity of substrate metabolites

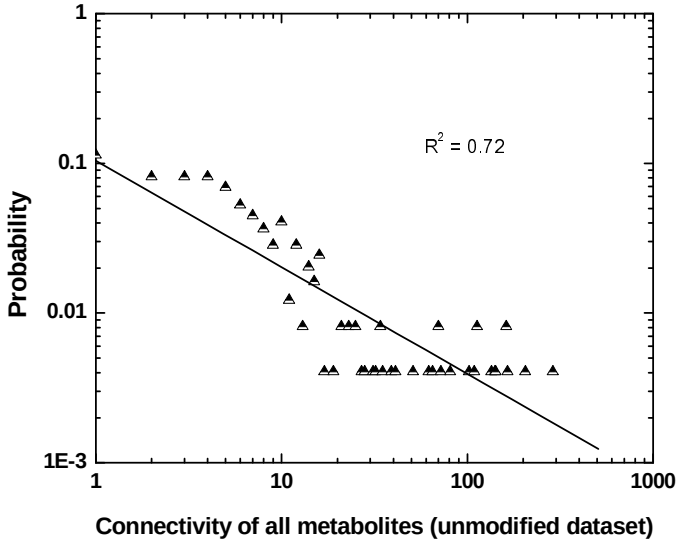

Figure 4.5c: Probability distribution of connectivity of all metabolites

It can be noticed that all the three plots follow a linear correlation on a log-log scale indicating that they follow a scale-free architecture. This signifies that, for all the three cases, there are some highly connected metabolites that link to other metabolites, which participate in only a few reactions as substrate, product or both. The majority of metabolites are present only in a few reactions, i.e., more than 54% metabolites play a role in one or two reactions. At a higher value of the number of connections, the probability decays to a small value and remains approximately constant, as it is most likely that only a single compound would be involved in that many reactions. This is because there are only 12 metabolites that have 100 or more occurrences in the reaction mechanism as substrate, product or both. These hub metabolites have a central responsibility to connect a number of compounds and enable the efficient conversion of one metabolite to another. From the computations, it is found that there are 597 compounds that act as substrate and 611 compounds that occur as product. Of this mix, 282 compounds play a part as both substrate and product. This tally shows that there are 926 compounds involved in the metabolic network of *Drosophila melanogaster*.

The frequency distribution (Figure 4.6) of the compounds that participate in the reaction mechanism, classified based on their links, further reinforces the fact that the metabolic network of *Drosophila* does indeed follow a scale-free topology. The plot shows that there are a high number of substrates and products taking part in five or fewer reactions. As the number of links increases, the compounds involved in that many reactions begins to diminish steadily.

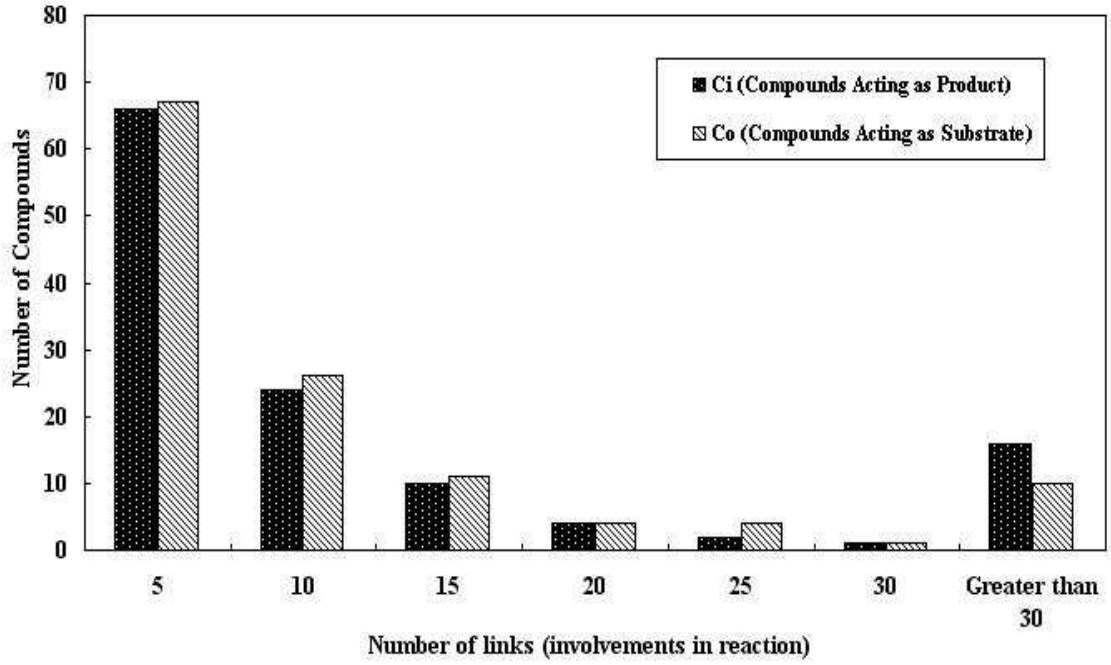

Figure 4.6: Frequency of interaction of compounds involved in the metabolic pathways

To make the graphs clear, the number of links greater than 30 is included in one bar. When the plot is stretched out for all values of links, the number of metabolites involved decline progressively, reach a value of one and remain constant. This emphasizes the fact that at higher degrees, there is only a solitary metabolite contributing to that many numbers of reactions.

4.2.2 Determination of the Topology of the Metabolic Network without the Current Metabolites

The new dataset, devoid of cofactor compounds, is used to plot the probability-connectivity data of the metabolites. As revealed in Figure 4.7, the distribution still decays as power law, irrespective of the absence of the current metabolites.

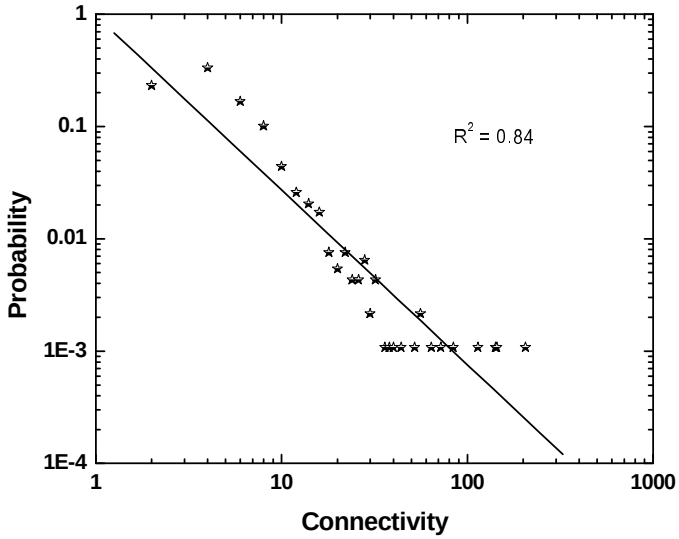

Figure 4.7: The probability-connections distribution plot for the metabolic interaction network

The results indicate that over 56% of the metabolites are involved in just one or two interactions while only about 2% metabolites participate in 30 or more interactions, a classic characteristic of scale-freeness. The exponent γ in the power-law distribution ($y = 0.9755x^{-1.5545}$) is found to be 1.55. This value is lesser than those obtained for other networks like the World Wide Web and the Internet that follow a scale-free architecture [22]. Nevertheless, the intrinsic topology of the network remains unaltered. The mean degree of the network is established to be 3.59. This signifies that on average, every metabolite that is part of the arrangement interacts with approximately three or four other metabolites in the network.

The average path length for this network is calculated as 5.29. Thus, on average, each metabolite can be converted to any other metabolite (that can be reached or converted) in five steps [49]. This shows that the metabolic network is highly compact as indicated by the small number of steps required to get from one metabolite to

another. This small world nature supports the fact that the metabolic interaction network of *Drosophila* follows a scale-free architecture. The diameter of the network is computed to be 18.

4.2.3 Effect of Connectivity on the Metabolic Interaction Network

The simulation of coordinated and random attacks shows that as the most connected metabolites are taken away from the network, the average path length and the diameter of the network are severely enhanced altering the inherent architecture of the network (Figures 4.8 and 4.9).

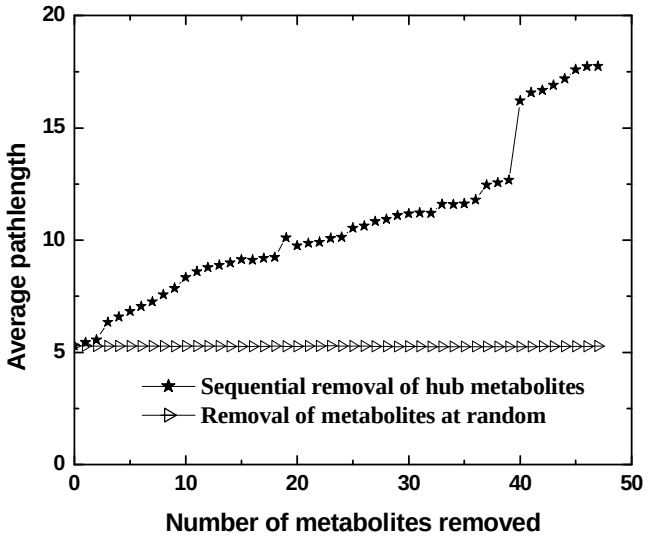

Figure 4.8: Effect of sequential and random exclusion of metabolites on the average path length of the metabolic network

After the removal of 5% of the hubs from the network, the average path length increases to 17.75.

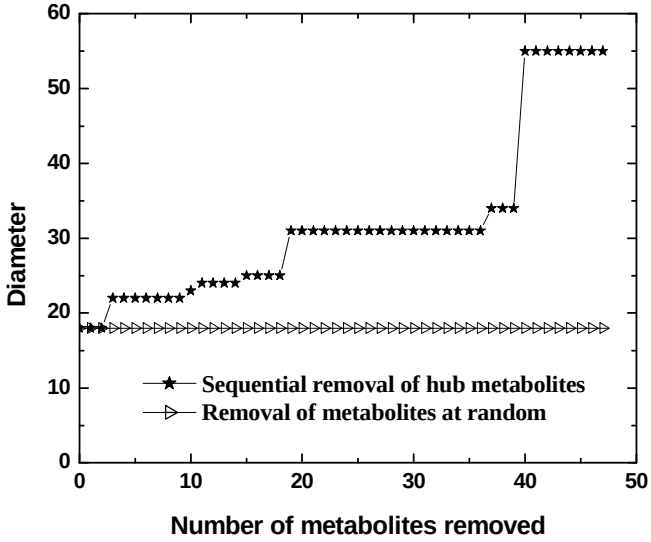

Figure 4.9: Effect of sequential and random exclusion of metabolites on the diameter of the metabolic network

The diameter increases to 55 after the elimination of 5% of the hubs from the network. It can be seen that the average path length and diameter suffer a more than three-fold augmentation after the 5% removal of the hub metabolites. The abnormal increase in the diameter from 34 to 55 on the elimination of L-Serine can be attributed to the fact that the organism loses its competent shortest pathway and a new set of pathways is organized. The longest pathway among the shortest pathways involved in the network, after the removal of L-Serine, is 55. This produces a radical change in the network design affecting the ability of the organism to produce a particular metabolite in a relatively small number of reactions.

Conversely, when the metabolites are eradicated in a random fashion, there scarcely is any change in the topological parameters. The elimination of a compound result in the search for a novel pathway, as the original pathway cannot be traversed. For a random compound, the organism is able to choose an alternate pathway without affecting the topology of the network, thereby providing development and survival stability. The novel pathway would lead to the target metabolite in more or less the same number of steps as the original pathway as proven by the steadfast nature of the topological parameters. Therefore, the critical nature of the hubs can be witnessed by the sweeping alterations that occur due to their simulated elimination.

4.2.4 Visualization of the Metabolic Interaction Network

Figure 4.10 shows the hubs (in red) that are essential to the network, the removal of which causes the collapse of its natural topology. The dense nature of the metabolites is when it is in its intrinsic state. The elimination of the hub metabolites causes this compact design to isolate into small groups, thereby disrupting the ability of the compounds to convert to one another in a small number of steps. The effective linking mechanism is lost and the system may be rendered meaningless.

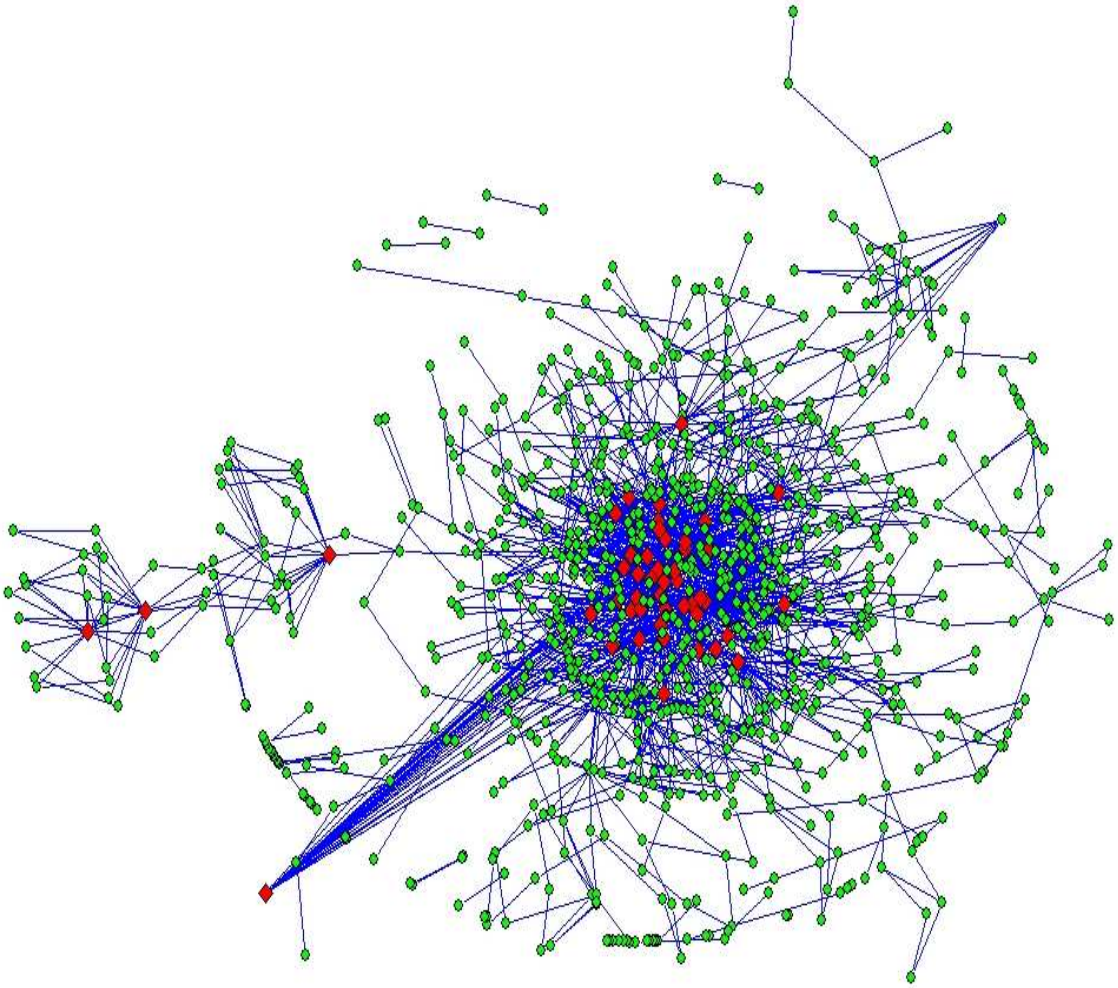

Figure 4.10: The metabolic interaction network of *Drosophila melanogaster* after the elimination of current metabolites

4.2.5 Distance of Metabolites to the Central Metabolism compounds

Using the breadth first search algorithm [47], the compounds are sorted based on their distance to the central metabolism and the number of connections as shown in Figure 4.11.

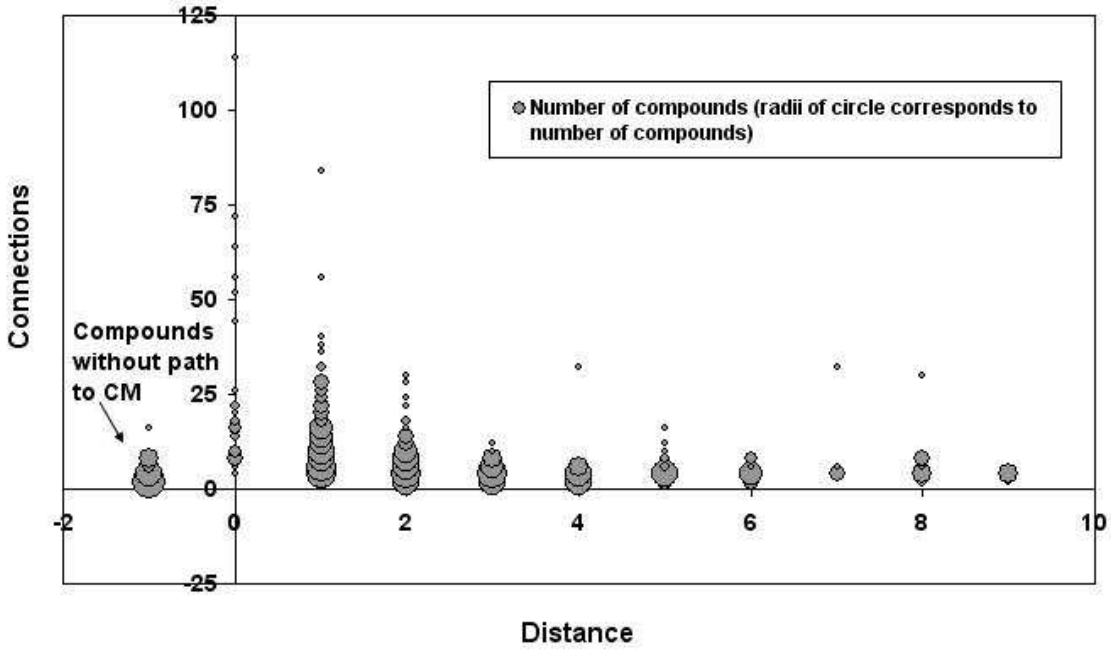

Figure 4.11: Plot illustrating the correlation between the distance to the central metabolism and the number of connections for each compound

The computation of the number of steps required to convert a specific compound to another compound that is involved in the central metabolism offers an insight into the biodegradable nature of that compound [79]. From the calculations and the subsequent plot, concrete observations can be made.

Thirty metabolites are involved in the central metabolism pathways. There are only a few compounds having the same specific number of connections in the central metabolism and hence, the smaller radii of circles. There are a large number of compounds having ten or less number of connections with short distances to the central metabolism, i.e., one or two steps, as indicated by the spheres of larger radii. It is found that about 53% of the compounds can be converted to central metabolism compounds in one or two steps. This proves that the majority of the compounds have a high degree of biodegradability. About 16% metabolites cannot be linked to any compound of the

central metabolism pathways. The rest of the compounds are sparsely distributed over the higher number of steps of conversion with the longest link being nine steps. The plot can be viewed as a power law distribution since the number of compounds decrease with increase in the distance to the central metabolism pathway compounds.

The distance of the hub metabolites to the central metabolism compounds is also analyzed. It is determined that among the top 5% of the hub metabolites used for the earlier study, 85% of the metabolites are either part of the central metabolism pathways or can be changed into one of them in a single step. This shows that the hub metabolites are closely linked to central metabolism. Consequently, this is unambiguous evidence that the central metabolism pathways are central to not only the function and maintenance of the metabolic processes but also the topological structure of the organism.

CHAPTER 5
CONCLUSIONS AND FUTURE DIRECTION

Protein-protein interaction network

The high confidence protein-protein interaction dataset is used to find the network architecture and to determine other topological considerations. The dataset reveals an inherent scale-free architecture, thereby underlining the fact that the majority of complex biological networks have a scale-free topology. The shortest path length of all the accessible proteins and the average path length and diameter of the network are also determined. The far-reaching alteration in the network topology (two-fold increase of the topological parameter values) due to the exclusion of hubs of the network confirms their essence to the network. The analysis of the molecular function of the well-connected metabolites proves that an extensive range of biochemical processes is managed by the hubs.

The protein-protein interaction network of *Drosophila melanogaster* renders a few moot points. A striking feature that deserves investigation is the ability of the organism to transfer to an alternate connection mechanism most times, when a non-hub protein is removed, to form new links and entail the smooth progress of the formation of biochemical products. Despite the fact that the highly connected proteins are removed, the organism is able to sustain its routine tasks, as it locates another protein that has a similar function to the protein that is being eradicated. This resilient nature of the organism needs further consideration. A physical denotation involving the strength of each interaction would provide a more comprehensive understanding of the network.

Metabolic interaction network

The study of the intrinsic doctrines that characterize the metabolic network of *Drososphila melanogaster* provides a crucial understanding of the construction blocks of the organism. The evaluation of the metabolic network of *Drosophila* identifies that 926 compounds are involved in the metabolic reactions as substrate, as product or as both. The graphical plots show that the probability distribution of these metabolites follows a scale-free design. The frequency distribution of these metabolites also indicates that only a few hubs exist in the network that participates in 30 or more reactions.

The presence of current metabolites generates an artificially short pathway between any two metabolites and proves detrimental to the computation of the topological parameters. The elimination of these cofactor compounds helps to determine the realistic number of steps required for the conversion of one compound to another and generates pragmatic topological parameter values. The exclusion of the hub metabolites demonstrates the harmful effects on the system as it disturbs the existing topology of the network. Their removal is lethal to the overall topology of the system leading to the failure of the efficient innate pathways. Although the organism is able to find alternate pathways to form products, it requires greater number of intermediate steps.

The average path length and the network diameter suffer a three-fold increase resulting from the top 5% exclusion of the hubs. This shows that a hub metabolite cannot be chosen as a target compound for therapeutic research as any interruption to the original pathway may cause the system to be secluded into diminutive groups causing a larger number of steps to be utilized for the production of the same compound. In some cases, the mechanism may not be able to find any avenues for an alternate pathway due to the formation of small clusters of metabolites causing the loss

of that specific product metabolite. A random metabolite with less participation in reactions could have a better efficacy to serve as a target metabolite and resist intrinsic errors.

The investigation of the distance of metabolites to central metabolism pathways indicates that a large number of the metabolites can be converted to central metabolism compounds in a single or couple of steps. This biodegradability enables the organism to convert many metabolites to the central metabolism compounds, for the generation of energy with minimal consumption of nutrients, to supplement its needs.

A weight-based interaction study can generate a better interpretation of the network. An understanding of the importance of each individual reaction can be offered because of such an analysis. The major hurdle in the assignment of weights to the interactions is the determination of the strength of each interaction. The influence of each interaction in the context of the web of interactions must also be considered. The binding forces that govern the interactive mechanism must be evaluated before a physical denotation can be provided to them. These factors require an advanced understanding about the interactive forces.

An alternate view to this problem would be to utilize the hubs of the network. The hubs play a critical role in linking several other nodes, thereby enabling them to have shorter path lengths. The adverse consequence observed due to the exclusion of well-linked nodes underscores their importance to the use of shorter path lengths. A higher weight could be provided to the edges of nodes not interacting with a well-connected node. For such nodes, the absence of any links to a hub causes a longer path length to a target node. On the other hand, the edges of nodes that link to a highly connected node can be provided with a lower weight. This approach will provide path lengths to a target node, based on weights. Nodes that have a link to a hub will produce

shorter weight-based path lengths to its target than when computed in their absence. An enhanced review could be available in the not so distant future.

LIST OF REFERENCES

[1] Beyond Books, Apex Learning Inc, 2005. The Cell: Down to Basics. Accessed May 22, 2005 at *http://www.beyondbooks.com/lif71/4.asp*

[2] Source: *www.encarta.msn.com*; retrieved May 23, 2005.

[3] Crick, F., 1970. Central Dogma of Molecular Biology. *Nature* 227: 561-563.

[4] Source: *http://cats.med.uvm.edu/cats_teachingmod/microbiology/courses/genomics/introduction/1_2_bgd_gen_pro_dog.html*; retrieved June 05, 2005.

[5] Lechtman, M.D., Roohk, B. and Egan, R.J., 1993. The games cells play: basic concepts of cellular metabolism. *Benjamin Cummings Publishing Company*.

[6] Dagley, S. and Nicholson, D.E., 1970. An Introduction to Metabolic Pathways. *Blackwell Scientific Publications*.

[7] Source: *http://www.humboldt.edu/~rap1/C431.F01/PathwayDiagrams*; retrieved June 15, 2005.

[8] Buehler, L.K., 2000. What is life? Accessed June 10, 2005 at *http://www.whatislife.com/*

[9] Lewin, B., 1997. Genes VI. *Oxford University Press (Sd)*.

[10] Source: U.S. Department of Energy Human Genome Program *http://biophysics.asu.edu/workshop/report/html/biomat.html*; retrieved June 29, 2005.

[11] Gomperts, B.D., Kramer, I.M. and Tatham, P.E.R., 2003. Signal Transduction. *Academic Press*.

[12] Source: *http://www.genome.jp/dbgetin/get_pathway?org_name=dme&mapno=04010*; retrieved June 28, 2005.

[13] Skeie, G.O., Romi, F., Aarli, J.A., Bentsen, P.T. and Gilhus, N.E., 2003. Pathogenesis of Myositis and Myasthenia Associated with Titin and Ryanodine Receptor Antibodies. *Annals of the New York Academy of Sciences* 998: 343-350.

[14] Source: *http://web.indstate.edu/thcme/mwking/amino-acids.html*; retrieved June 16, 2005.

[15] Source: *http://www.geneticsolutions.com/PageReq?id=1530:1873#Proteins*; retrieved July 01, 2005.

[16] Uetz, P. and Vollert, C.S., 2003. Protein-Protein Interactions. Accessed July 14, 2005 at *http://itgmv1.fzk.de/www/itg/uetz/publications/Uetz2003-PPI.pdf*

[17] Pandey, A. and Mann, M., 2000. Proteomics to study genes and genomes. *Nature* 405: 837-846.

[18] Duan, X.J., Xenarios, I. and Eisenberg, D., 2002. Describing biological protein interactions in terms of protein states and state transitions: the Live DIP database. *Molecular and Cellular Proteomics* 1: 104-116.

[19] Auerbach, D., Thaminy, S., Hottiger, M.O. and Stagljar, I., 2002. The post-genomic era of interactive proteomics: facts and perspectives. *Proteomics* 2 (6): 611-623.

[20] Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. and Y. Sakaki., 2001. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences* USA 98: 4569-4574.

[21] Uetz, P. and Hughes, R.E., 2000. Systematic and large-scale two-hybrid screens. *Current Opinions in Microbiology* 3: 303-308.

[22] Albert, R. and Barabási, A-L., 2002. Statistical mechanics of complex networks. *Reviews of Modern Physics* 74: 47-97.

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

[24] Source: *http://jaeger.earthsci.unimelb.edu.au/msandifo/Teaching/Mineralogy2/lattice.html*; retrieved June 26, 2005.

[25] Erdős, P. and Rényi, A., 1959. On random graphs. *Publicationes Mathematicae* 6: 290-297.

[26] Erdős, P. and Rényi, A., 1960. On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences* 5: 17-61.

[27] Erdős, P. and Rényi, A., 1961. On the strength of connectedness of a random graph. *Acta Mathematica Scientia Hungary* 12: 261-267.

[28] Bornholdt, S. and Schuster, H.G., (Editors) 2003. Handbook of Graphs and Networks. *Wiley-Vch*.

[29] Barábasi, A-L., 2002. Linked: The New Science of Networks. *Perseus Publishing*.

[30] Watts, D.J. and Strogatz, S.H., 1998. Collective dynamics of 'small-world' networks. *Nature* 393: 440-442.

[31] Strogatz, S.H., 2001. Exploring complex networks. *Nature* 410: 268-276.

[32] Adamic, L.A., Huberman, B.A., Barabási, A-L., Albert, R., Jeong, H. and Bianconi, G., 2000. Power-law distribution of the World Wide Web. *Science* 287: 2115.

[33] Albert, R., Jeong, H. and Barabási, A-L., 1999. Diameter of the World Wide Web. *Nature* 401: 130-131.

[34] Faloutsos, M., Faloutsos, P. and Faloutsos, C., 1999. On power-law relationships of the internet topology. *Computer Communications Review* 29: 251-262.

[35] Amaral, L.A.N., Scala, A., Barthélémy, M. and Stanley, H.E., 2000. Classes of small world networks. *Proceedings of the National Academy of Sciences* USA 97: 11149-11152.

[36] Newman, M.E.J., 2000a. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences* USA 98: 404-409.

[37] Newman, M.E.J., Strogatz, S.H. and Watts, D.J., 2001. Random graphs with arbitrary degree distributions and their applications. *Physical Review E* 64: 026118.

[38] Milgram, S., 1967. The small world problem. *Psychology Today* 2: 60-67.

[39] Travers, J. and Milgram, S., 1969. An experimental study of the small world problem. *Sociometry* 32: 425.

[40] Barabási, A-L. and Albert, R., 1999. Emergence of scaling in random networks. *Science* 286: 509-512.

[41] Barabási, A-L., Albert, R. and Jeong, H., 1999. Mean-field theory for scale-free random networks. *Physica A* 272: 173-187.

[42] Albert, R., Jeong, H. and Barabási, A-L., 2000. Error and attack tolerance of complex networks. *Nature* 406: 378-382.

[43] Watts, D.J., 1999. Small Worlds. *Princeton University Press.*

[44] Barabási, A-L. and Oltvai, Z.N., 2004. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics* 5 (2): 101-113.

[45] Fell, D.A. and Wagner, A., 2000. The small world of metabolism. *Nature Biotechnology* 18: 1121-1122.

[46] Bilke, S. and Peterson, C., 2001. Topological properties of citation and metabolic networks. *Physical Review E* 64: 036106.

[47] Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A. and Wiener, J., 2000. Graph structure in the web. *Computer Networks* 33: 309-320.

[48] Barthélémy, M. and Amaral, L. A. N., 1999. Small-World Networks: Evidence for a Crossover Picture. *Physical Review Letters* 82 (15): 3180-3183.

[49] Ma, H.W. and Zeng, A-P., 2003. Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics* 19 (2): 270-277.

[50] Podani, J., Oltvai, Z.N., Jeong, H., Tombor, B., Barabási, A-L. and Szathmary, E., 2001. Comparable system-level organization of Archaea and Eukaryotes. *Nature Genetics* 29: 54-56.

[51] Ma, H.W. and Zeng, A-P., 2002. The hierarchical structure giant strong component and centrality of metabolic networks. *Bioinformatics* 19 (11): 1423-1430.

[52] Hasty, J. and Collins, J.J., 2001. Unspinning the web. *Nature* 411: 30-31.

[53] Jeong, H., Mason, S.P., Barabási, A-L. and Oltvai, Z.N., 2001. Lethality and centrality in protein networks. *Nature* 411: 41-42.

[54] Schwikowski, B., Uetz, P. and Fields, S., 2000. A network of protein-protein interactions in yeast. *Nature Biotechnology* 18: 1257-1261.

[55] Rain, J.C., Selig, L., De Reuse, H., Battaglia, V., Reverdy, C., Simon, S., Lenzen, G., Petel, F., Wojcik, J., Schachter, V., Chemama, Y., Labigne, A. and Legrain, P., 2001. The protein-protein interaction map of *Helicobacter pylori*. *Nature* 409: 211-215.

[56] Patterson, J. and Stone, W., 1952. Evolution in the Genus *Drosophila*. *Macmillan Publishers*.

[57] Demerec, M., 1950. Biology of *Drosophila*. *John Wiley and Sons, Inc*.

[58] Manning, G., 1999. The Drosophila Virtual Library (On-line). Accessed May 29, 2005 at *http://www.ceolas.org/fly/*

[59] Patterson, J., Wagner, R. and Wharton, L., 1943. The *Drosophilidae* of the Southwest. *The University of Texas Press*.

[60] Source: Department of Biology, Memorial University of Newfoundland; retrieved July 06, 2005.

[61] Giot, L., Bader, J.S., Brouwer, C., Chaudhri, A., Kuang, B., Li, Y., Hao, Y.L., Ooi, C.E., Godwin, B., Vitols, E., Vijayadamodar, G., Pochart, P., Machineni, H., Welsh, M., Kong, Y., Zerhusen, B., Malcolm, R., Varrone, Z., Collis, A., Minto, M., Burgess, S., McDaniel, L., Stimpson, E., Spriggs, F., Williams, J., Neurath, K., Loime, N., Agee, M., Voss, E., Furtak. K., Renzulli, R., Aanensen, N., Carrola, S., Bickelhaupt, E., Lazovatsky, Y., DaSilva, A., Zhong, J., Stanyon, C.A., Finley Jr., R.L., White, K.P., Braverman, M., Jarvie, T., Gold, S., Leach, M., Knight, J., Shimkets, R.A., McKenna, M.P., Chant, J. and Rothberg, J.M., 2003. A protein interaction map of *Drosophila melanogaster*. *Science* 302: 1727-1736.

[62] US EPA, Metabolic Engineering Working Group, 2002. Accessed June 12, 2005 at *http://www.epa.gov/opptintr/metabolic*

[63] Fields, S. and Song, O-K, 1989. A novel genetic system to detect protein-protein interactions. *Nature* 340: 245-246.

[64] Bartel, P.L. and Fields, S., (Editors) 1997. The Yeast Two-Hybrid System. *Oxford University Press*.

[65] Fields, S. and Sternglanz, R., 1994. The two-hybrid system: an assay for protein-protein interactions. *Trends in Genetics* 10: 286-292.

[66] Source: *http://www.bioteach.ubc.ca/MolecularBiology/AYeastTwoHybridAssay*; retrieved July 22, 2005.

[67] Finley, R. and Brent, R., 1994. Interaction mating reveals binary and ternary connections between *Drosophila* cell cycle regulators. *Proceedings of the National Academy of Sciences* USA 91, 12980-12984.

[68] Source: *http://cran.r-project.org/*; accessed October 16, 2004.

[69] Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J,M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. and Sherlock, G., 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics* 25 (1): 25-29.

[70] Source: *http://www.curagen.com/*; retrieved May 15, 2004.

[71] Department of Computer Science, University of Saskatchewan, 1999. Graph Theory: An Introduction. Accessed June 19, 2005 at *http://www.cs.usask.ca/resources/tutorials/csconcepts/1999_8/tutorial/index.html*

[72] Siek, J.G., Lee, L-Q. and Dumsdaine, A., 2001. Boost Graph Library, The: User Guide and Reference Manual. *Addison Wesley Professional*.

[73] Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H. and Kanehisa, M., 1999. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* 27: 29-34.

[74] Kanehisa, M., Goto, S., Kawashima, S. and Nakaya, A., 2002. The KEGG databases at GenomeNet. *Nucleic Acids Research* 30 (1): 42-46.

[75] Neidhardt, F.C., Ingraham, J.L. and Schaechter, M., 1990. Physiology of the Bacterial Cell: a molecular approach. *Sinauer Associates*.

[76] Schuster, S., Pfeiffer, T., Moldenhauer, F., Koch, I. and Dandekar, T., 2002. Exploring the pathway structure of metabolism: decomposition into subnetworks and application to *Mycoplasma pneumoniae*. *Bioinformatics* 18: 351-361.

[77] Schilling, C.H., Letscher, D. and Palsson, B.O., 2000. Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *Journal of Theoretical Biology* 203: 229-248.

[78] Schuster, S., Fell, D.A. and Dandekar, T., 2000. A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nature Biotechnology* 18: 326-332.

[79] Pazos, F., Valencia, A. and De Lorenzo, V., 2003. The organization of the Microbial Biodegradation Network from a Systems-Biology perspective. *EMBO Reports* 4(10): 994-999.

[80] Batagelj, V. and Mrvar, A., 1998. Pajek – program for large network analysis. *Connections* 21: 47-57.

[81] Drysdale, R.A., Crosby, M.A. and The FlyBase Consortium, 2005. FlyBase: genes and gene models. *Nucleic Acids Research* 33: D390-D395. *http://www.flybase.org/*; retrieved Jan 24, 2005.

[82] Oltvai, Z.N. and Barabási, A-L., 2002. Life's complexity pyramid. *Science* 298: 763-764.

APPENDIX A

Datasets for Protein-Protein Interaction Network

Table 4.1: Data for the probability - degree plot of proteins

Degree '*k*'	Number of proteins with '*k*' interactions	Probability, *p*(*k*)
1	2549	0.5555
2	955	0.2081
3	484	0.1055
4	252	0.0549
5	137	0.0299
6	72	0.0157
7	54	0.0118
8	22	0.0048
9	15	0.0033
10	16	0.0035
11	15	0.0033
12	7	0.0015
13	2	0.0004
14	2	0.0004
15	1	0.0002
16	1	0.0002
17	1	0.0002
18	2	0.0004
20	1	0.0002
42	1	0.0002

Table 4.2: Data for the effect of removal of most connected and random proteins

Proteins removed	Effect due to removal of most connected proteins		Effect due to removal of proteins at random	
	Ave. path length	Diameter	Ave. path length	Diameter
0	9.4210	27	9.4210	27
1	9.7130	27	9.4275	27
2	9.8014	27	9.4275	27
3	9.8106	27	9.4267	27
4	9.9013	27	9.4239	27
5	9.9798	27	9.4326	27
6	10.0325	27	9.4326	27
7	10.0670	27	9.4327	27
8	10.1429	27	9.4337	27
9	10.1530	27	9.4358	27
10	10.2149	27	9.4350	27
11	10.2320	27	9.4107	26
12	10.2693	27	9.4108	26
13	10.3009	27	9.4084	26
14	10.3043	27	9.4091	26
15	10.3373	27	9.4093	26
16	10.4461	27	9.4098	26
17	10.4912	27	9.4098	26
18	10.6458	27	9.4098	26
19	10.6406	27	9.4098	26
20	10.6863	27	9.4095	26
21	10.7232	27	9.4095	26
22	10.7461	27	9.4095	26
23	10.7749	28	9.4110	26

24	10.7980	28	9.4110	26
25	10.8398	28	9.4221	26
26	10.9118	28	9.4255	26
27	10.9226	28	9.4249	26
28	10.9665	28	9.4250	26
29	11.0039	28	9.3967	26
30	11.1073	28	9.3967	26
31	11.1430	29	9.3956	26
32	11.1978	29	9.3956	26
33	11.2014	29	9.3947	26
34	11.2029	29	9.3886	26
35	11.3233	29	9.3886	26
36	11.3224	29	9.3886	26
37	11.4097	29	9.3889	26
38	11.4982	29	9.3897	26
39	11.5519	29	9.3897	26
40	11.5630	29	9.3897	26
41	11.5725	29	9.3894	26
42	11.6140	29	9.3879	26
43	11.7990	30	9.3879	26
44	11.8509	30	9.3879	26
45	11.8708	30	9.3877	26
46	11.8970	30	9.3877	26
47	11.8753	30	9.3877	26
48	11.9617	30	9.3881	26
49	12.0045	30	9.3876	26
50	12.0418	30	9.3865	26
51	12.1578	33	9.3868	26
52	12.2964	33	9.3918	26
53	12.3133	33	9.3919	26
54	12.3480	33	9.3922	26

55	12.3553	33	9.3922	26
56	12.4987	33	9.3923	26
57	12.5243	33	9.3923	26
58	12.5634	33	9.3915	26
59	12.6680	33	9.4008	26
60	12.7972	33	9.4032	26
61	12.9928	33	9.4031	26
62	13.0759	33	9.4018	26
63	13.0821	33	9.4018	26
64	13.0462	33	9.4056	26
65	13.2314	36	9.4058	26
66	13.3839	36	9.4058	26
67	13.5508	36	9.4054	26
68	13.5801	36	9.4054	26
69	13.6341	36	9.4054	26
70	13.7852	36	9.4114	26
71	13.8500	38	9.4109	26
72	13.8802	38	9.4252	26
73	13.9410	38	9.4252	26
74	13.9763	38	9.4253	26
75	14.0936	38	9.4255	26
76	14.0967	38	9.4235	26
77	14.1760	38	9.4243	26
78	14.2230	38	9.4238	26
79	14.2566	38	9.4442	26
80	14.2678	38	9.4570	26
81	14.2716	38	9.4431	26
82	14.2834	38	9.4431	26
83	14.2911	38	9.4428	26
84	14.5738	38	9.4428	26
85	14.9188	38	9.4388	26

86	15.3576	41	9.4388	26
87	15.6914	45	9.4388	26
88	15.8573	46	9.4399	26
89	15.8336	46	9.4399	26
90	16.1163	47	9.4400	26
91	16.1166	47	9.4400	26
92	16.3057	49	9.4400	26
93	15.6318	41	9.4426	26
94	15.6349	41	9.4426	26
95	15.6872	41	9.4450	26
96	15.8466	41	9.4459	26
97	16.0872	41	9.4456	26
98	16.0873	41	9.4457	26
99	16.1242	41	9.4484	26
100	16.2894	41	9.4484	26
101	16.4279	41	9.4496	26
102	16.4205	41	9.4496	26
103	16.5212	41	9.4483	26
104	16.5385	41	9.4441	26
105	16.6393	41	9.4441	26
106	16.5493	41	9.4428	26
107	16.5848	41	9.4491	26
108	16.6906	41	9.4346	26
109	16.8295	41	9.4346	26
110	16.8990	41	9.4362	26
111	16.9005	41	9.4362	26
112	16.8994	41	9.4387	26
113	17.0092	41	9.4263	26
114	17.2628	41	9.4255	26
115	17.1620	41	9.4032	26
116	17.3658	48	9.4034	26

117	17.7162	48	9.4034	26
118	17.7163	48	9.4042	26
119	18.1350	51	9.4042	26
120	18.1253	51	9.4014	26
121	18.1804	51	9.4062	26
122	18.0899	51	9.4056	26
123	18.1894	51	9.4144	26
124	18.2038	51	9.4144	26
125	18.3722	51	9.4634	27
126	18.6617	51	9.4610	27
127	18.7171	51	9.4586	27
128	18.6890	51	9.4592	27
129	19.0626	51	9.4591	27
130	19.0636	51	9.4524	27
131	19.1477	51	9.4524	27
132	19.2083	51	9.4524	27
133	19.2765	51	9.4523	27
134	20.0621	51	9.4543	27
135	20.0625	51	9.4543	27
136	20.6910	54	9.4544	27
137	20.7146	54	9.4539	27
138	20.9273	54	9.4539	27

Datasets for Metabolic Interaction Network

Table 4.3: Data for the frequency distribution plot

Number of links	C_i	C_o
5	66	67
10	24	26
15	10	11
20	4	4
25	2	4
30	1	1
Greater than 30	16	10

Table 4.4: Data for the probability - connectivity plot of metabolites

Connectivity '*k*'	Number of metabolites with '*k*' interactions	Probability, *p*(*k*)
2	215	0.2322
4	309	0.3337
6	155	0.1674
8	94	0.1015
10	41	0.0443
12	24	0.0259
14	19	0.0205
16	16	0.0173

18	7	0.0076
20	5	0.0054
22	7	0.0076
24	4	0.0043
26	4	0.0043
28	6	0.0065
30	2	0.0022
32	4	0.0043
36	1	0.0011
38	1	0.0011
40	1	0.0011
44	1	0.0011
52	1	0.0011
56	2	0.0022
64	1	0.0011
72	1	0.0011
84	1	0.0011
114	1	0.0011
142	1	0.0011
144	1	0.0011
206	1	0.0011

Table 4.5: Data for the effect of elimination of most connected and random metabolites

Metabolites Removed	**Effect due to elimination of most connected metabolites**		**Effect due to elimination of metabolites at random**	
	Ave. path length	**Diameter**	**Ave. path length**	**Diameter**
0	5.2879	18	5.2879	18
1	5.4373	18	5.2890	18

2	5.5511	18	5.2845	18
3	6.3393	22	5.2849	18
4	6.5751	22	5.2850	18
5	6.8310	22	5.2841	18
6	7.0441	22	5.2787	18
7	7.2498	22	5.2787	18
8	7.5698	22	5.2790	18
9	7.8522	22	5.2710	18
10	8.3312	23	5.2747	18
11	8.6030	24	5.2778	18
12	8.7825	24	5.2778	18
13	8.8813	24	5.2769	18
14	8.9842	24	5.2633	18
15	9.1287	25	5.2647	18
16	9.1158	25	5.2664	18
17	9.1913	25	5.2722	18
18	9.2411	25	5.2755	18
19	10.1006	31	5.2724	18
20	9.7518	31	5.2695	18
21	9.8591	31	5.2742	18
22	9.9081	31	5.2773	18
23	10.0777	31	5.2859	18
24	10.1254	31	5.2818	18
25	10.5334	31	5.2765	18
26	10.6257	31	5.2776	18
27	10.8346	31	5.2655	18
28	10.9240	31	5.2656	18
29	11.0945	31	5.2605	18
30	11.1710	31	5.2632	18
31	11.2050	31	5.2666	18
32	11.1982	31	5.2606	18

33	11.5987	31	5.2521	18
34	11.5863	31	5.2545	18
35	11.6177	31	5.2545	18
36	11.7872	31	5.2571	18
37	12.4558	34	5.2614	18
38	12.5586	34	5.2553	18
39	12.6639	34	5.2587	18
40	16.1976	55	5.2571	18
41	16.5625	55	5.2603	18
42	16.6692	55	5.2688	18
43	16.8993	55	5.2638	18
44	17.1851	55	5.2669	18
45	17.5962	55	5.2701	18
46	17.7263	55	5.2723	18
47	17.7456	55	5.2739	18

Table 4.6: Correlation between the distance to central metabolism and the number of connections

Distance	Number of connections	Number of compounds
0	4	1
0	6	1
0	8	5
0	10	3
0	14	2
0	16	3
0	18	2
0	20	1
0	22	2

0	26	1
0	44	1
0	52	1
0	56	1
0	64	1
0	72	1
0	114	1
0	142	1
0	144	1
0	206	1
1	2	5
1	4	55
1	6	84
1	8	38
1	10	24
1	12	17
1	14	13
1	16	10
1	18	3
1	20	4
1	22	4
1	24	3
1	26	3
1	28	5
1	32	2
1	36	1
1	38	1
1	40	1
1	56	1
1	84	1
2	2	29

2	4	78
2	6	38
2	8	27
2	10	12
2	12	5
2	14	4
2	16	1
2	18	2
2	22	1
2	24	1
2	28	1
2	30	1
3	2	37
3	4	60
3	6	18
3	8	7
3	10	1
3	12	1
4	2	23
4	4	31
4	6	6
4	32	1
5	2	5
5	4	22
5	6	2
5	8	2
5	10	1
5	12	1
5	16	1
6	2	4
6	4	12

6	6	1
6	8	3
7	4	5
7	6	1
7	32	1
8	2	2
8	4	8
8	6	1
8	8	4
8	30	1
9	2	1
9	4	8
-1	2	109
-1	4	29
-1	6	3
-1	8	8
-1	16	1

APPENDIX B

Glossary

Chapter 1

1.1 Phenomena used to determine the existence of an organism

Activity: A movement is the foremost indicator of the very existence of an organism. Although plants may not move about physically as other entities, they do have internal movement.

Consumption: Exchanging matter with the outer world is a signature of life. Organisms exhibit breathing and ingest matter for sustenance. They show signs of metabolism to feed them and excretion to purge waste matter after utilizing the components required for their nourishment.

Development: An increase, as in size, value or strength leading to the evolution from a simpler to a more complex form, is witnessed in all micro and macro living beings. All plants, animals, fungi and humans demonstrate evidence of growth.

Reproduction: Life has the inexorable property to multiply itself. It is a means to pass life onto posterity.

Stimulus Response: Organisms show some form of reaction to an external stimulus. Any change in the surrounding environment is dealt accordingly by the mechanism of the organism [82].

1.2 Organic Components of the Cell

Carbohydrate sugars are the main source of cellular energy and act as the structural components of cells.

Lipids are bipolar molecules, the configuration of which is liable for many properties of the biological membrane. Some hormones are also derived from lipids.

Nucleic acids are highly essential macromolecules that refer to a group of multi-faceted compounds found in all living cells and viruses. They are composed of purines, pyrimidines, carbohydrates and phosphoric acid. The nucleic acids that are present in the form of DNA and RNA direct the cellular function and heredity factors of the organism.

Proteins are large, organized molecules composed of one or more amino acid chains. The order of the amino acids is determined by the base sequence of nucleotides (base pairs) in the gene that codes for a particular protein. Proteins are vital for the structure, function, and regulation of cells, tissues, and organs. Hormones, enzymes and antibodies are some of the proteins found in an organism.

Chapter 2

2.1 Network Theory

Characteristic or Average Path length: The average path length is computed as the mean of the shortest path length for all pairs of nodes accessible within the network or graph.

Clustering Coefficient: A network is clustered if the probability of two nodes being connected by an edge is higher when the nodes have a common neighbor (that is,

another node in the network to which both are attached). The clustering coefficient is a measure of the interrelatedness of the neighbors of an entity.

Degree: The number of edges connected to a node.

Degree Distribution: The probability that a selected node has exactly ‘k’ links.

Diameter: The diameter is the length of the longest geodesic path between any two nodes among the calculated shortest paths.

Directed Graph: A directed graph is one in which the edges have a definite direction or in other words, go only in one way.

Edge (Link): In communications systems and network topologies, a route between any two points or the connection between two nodes.

Geodesic Path: The shortest path required by one node to reach another node in the network.

Graph: A graph is the symbolic representation of a network. It implies an abstraction of the reality, so it can be simplified as a set of linked nodes.

In-Degree: The in-degree is the number of incoming edges or edges pointing towards the node.

Node (Vertex): In network topology, a terminal of any branch of a network or an interconnection common to two or more branches of a network.

Out-Degree: The out-degree is the number of outgoing edges or edges going away from the node.

Path length (Pathway): The path length or pathway is a measure of the number of links needed for a specific node to get to any other node in the network.

Undirected Graph: In an undirected graph there is no specific direction for the edges. The edges may go in both ways.

Un-weighted Graph: All the edges of the graph are equal.

Weighted Graph: The edges of a weighted graph are not equal. Each edge has a specific strength allotted to it depending on the network that is being studied.

APPENDIX C

4.1 & 4.2 Probability Distribution - sample calculation

Consider a network of ten nodes namely, A, B, C, D, E, F, G, H, I and J each with 7, 1, 2, 3, 2, 4, 7, 2, 6 and 11 links respectively.

N: the total number of nodes in the network = 10

k: the number of links = 1, 2, 3….

p(*k*): probability of a node having '*k*' links = (number of nodes having *k* links)/*N*

Hence, for *k* =2, $p(k_2)$ = 3 (i.e., nodes C, E and H)/10 = **0.3**.