COMPUTATIONAL IDENTIFICATION OF SIGNALLING AND METABOLIC PATHWAYS OF *PLASMODIUM falciparum*

BY

OYELADE, Olanrewaju Jelili
B.Sc. (Computer Science with Mathematics), Obafemi Awolowo University, 1998
M.Sc. (Computer Science), Obafemi Awolowo University, 2004

A THESIS SUBMITTED TO THE DEPARTMENT OF COMPUTER AND INFORMATION SCIENCES, SCHOOL OF NATURAL AND APPLIED SCIENCES, COLLEGE OF SCIENCE AND TECHNOLOGY, COVENANT UNIVERSITY, OTA, NIGERIA

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF THE DEGREE OF DOCTOR OF PHILOSOPHY IN COMPUTER SCIENCE

2012

CERTIFICATION

It is hereby certified that this thesis, written by Oyelade, Olanrewaju Jelili was supervised by us and submitted to the Department of Computer and Information Sciences, College of Science and Technology, Covenant University, Ota.

1. Prof. Ezekiel Adebiyi — Supervisor

   Signature:…… Date:……

2. Dr. Benedict Brors — Co-Supervisor

   Signature:…… Date:……

3. Prof. C. K. Ayo — Head of Department

   Signature:…… Date:……

DECLARATION

It is hereby declared that, this research was undertaken by Oyelade, Olanrewaju Jelili. The thesis is based on his original study in the Department of Computer and Information Sciences, College of Science and Technology, Covenant University, Ota, under the supervision of Prof. Ezekiel Adebiyi and Dr. Benedict Brors. Ideas and views of this research work are products of the original research undertaken by Oyelade, Olanrewaju Jelili and the views of other researchers have been duly expressed and acknowledged.

Prof. Ezekiel Adebiyi
Supervisor

Signature…….. Date……..

Dr. Benedict Brors
Co- Supervisor

Signature…….. Date……..

DEDICATION

To GOD Almighty, who gave me the inspiration towards the achievement of this work. To Ayodele, my wife, who put up with the clutter and kept me focused and my children, Peculiar and Treasure. Also to my parents, who encouraged my curiosity and the desire to study.

ACKNOWLEDGEMENT

I would sincerely like to thank my Supervisors Prof. E. F. Adebiyi and Dr. Benedict Brors, they have been one of the most patient and helpful guides one can ask for. Working on this thesis has been a great learning experience for me and I am grateful to them for giving me this opportunity. I would also like to thank Miss I. M. Ewejobi for her valuable inputs, and numerous ideas that considerably improved this thesis. My profound gratitude also goes to my Head of Department, Prof. C. K. Ayo for the understanding and cooperation throughout this research work. I am also thankful to my parents, sister and family for all the love and support.

TABLE OF CONTENTS

CHAPTER TWO: LITERATURE REVIEW

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

ADP	Adenosine Diphosphate
ATP	Adenosine Triphosphate
cAMP	cyclic Adenosine Monophosphate
CDK	Cyclin-Dependent Kinase
cGMP	cyclic Guanosine Monophosphate
DAG	Directed Acyclic Graph
DNA	Deoxyribonucleic Acid
FBA	Flux Balance Approach
GO	Gene Ontology
GSK	Glycogen Synthase
HFB	High Flux Backbone
IDC	Intraerythrocytic Development Cycle
KEGG	Kyoto Encyclopedia of Genes and Genomes
MAP	Mitogen Activated Protein
MAPK	Mitogen Activated Protein Kinase
MPMP	Malaria Parasite Metabolic Pathways
MEK	MAPK Kinase
MEKK	MAPK Kinase Kinase
mRNA	messanger Ribonucleic Acid
NADPH	Nicotinamide Adenine Dinucleotide Phosphase
PPI	Protein-Protein Interaction
PV	Parasitophorous Vacuole
PLC	Phospholipase C

PfPKB	*Plasmodium falciparum* Protein Kinase B
RBC	Red Blood Cell
TAP	Transcription Associated Proteins
TCA	Tricarboxylic Acid Cycle

ABSTRACT

Malaria is one of the world's most common and serious diseases causing death up to about three million people each year. Its most severe occurrence is caused by the protozoan *Plasmodium falciparum*. Reports have shown that the resistance of the parasite to existing drugs is increasing. Therefore, there is a huge and urgent need to discover and validate new drug or vaccine targets to enable the development of new treatments for malaria. The ability to discover these drug or vaccine targets can only be enhanced from our deep understanding of the detailed biology of the parasite, for example, how cells function and how proteins organize into modules such as metabolic, regulatory and signal transduction pathways. The formally effective and popular anti-malaria drug chloroquine inhibits multiple sites in metabolic pathways, leading to neutrophil superoxide release. It has therefore been noted that the knowledge of metabolic pathways and recently signalling transduction pathways in *Plasmodium* are fundamental to aid the design of new strategies against malaria. In the first part of this work, a linear-time algorithm for finding paths in a protein-protein interactions network under modified biologically motivated constraints was used. Several important signalling transduction pathways in *Plasmodium falciparum* were predicted. A viable signalling pathway characterized in terms of the genes responsible that may be the PfPKB pathway recently elucidated in *Plasmodium falciparum* was predicted. We obtained from the FIKK family, a signal transduction pathway that ends upon a chloroquine resistance marker protein, which indicates that interference with FIKK proteins might reverse *Plasmodium falciparum* from resistant to sensitive phenotype. We also propose a hypothesis that showed the FIKK proteins in this pathway as enabling the resistance parasite to have a mechanism for releasing chloroquine(via an efflux process). Furthermore, a signalling pathway that may have been responsible for signalling the start of the invasion process of Red Blood Cell(RBC) by the merozoites was also predicted. It has been noted that the understanding of this pathway will give insight into the parasite virulence and will facilitate rational vaccine design against merozoites invasion. And we have a host of other predicted pathways, some of which have been used in this work to predict the functionality of some proteins. In another work, we adapted and extended a method (used in the first work for extracting signalling pathways) to extract linear metabolic pathways from the malaria parasite, *Plasmodium falciparum* metabolic weighted graphs (networks). The weights are calculated using the metabolite degrees. Adopting the representation of the biochemical metabolic network as we have in Koenig *et al*., 2006, we are able to make our algorithm tenable to accept metabolic network from other source apart from KEGG. This gives us opportunity for the first time, to compare the metabolic pathways extracted from different metabolic networks. We run our algorithm (for four selected pathways: Pyruvate, Glutamate, Glycolysis and Mitochondrial TCA) on graph from KEGG and compare our results with the results obtained from recent algorithms: ReTrace and

atommetanet. Our results compare favourably with these two algorithms. Considering the results with genes classified into these pathways from Plasmodb, resulted into a lot of false positiveness. Furthermore, we compared the runs of our algorithm on graphs from KEGG and PlasmoCyc (from BioCyc). The results are remarkably different and the results from PlasmoCyc produced less false positiveness when compared to the results from Plasmodb. We identify 2, 1, 2, 4 gene(s) in addition to belong to these pathways respectively. Some of the genes have not been classified earlier to any known metabolic pathways.

CHAPTER ONE

INTRODUCTION

1.1 Background Information

The most fatal and prevalent form of malaria is caused by the blood borne pathogen *Plasmodium falciparum*. Annually, approximately up to three million people die of malaria. Also, hundreds of millions of people in a year become clinically ill (Bozdech *et al.*, 2003a). The negative influence of these results is huge and its socioeconomic impact is beyond measure. This influence is particularly prominent in the African continent, where an estimated US$12 billion is being lost yearly (Breman *et al.*, 2004, Gallup and Sachs, 2001). Reports have shown that the resistance of the parasite to existing drugs is increasing. The formerly popular anti-malaria drug chloroquine, which inhibits multiple sites in metabolic pathways leading to neurophil superoxide release is largely now ineffective and the currently popular one, artemisinin's biologically mode of action is controversial. Therefore, there is a huge and urgent need to discover and validate new drug or vaccine targets to enable the development of new treatments for malaria (Ben *et al*., 2001). The ability to discover these drug or vaccine targets can only be enhanced from our deep understanding of the detailed biology of the parasite, for example, how cells function and how proteins organize into modules such as metabolic, regulatory and signal transduction pathways. Biologically, a signal transduction pathway is the chain of processes by which a cell converts an extra cellular signal into a response, while metabolic pathways are processes (series of chemical reactions) by which the parasite produces the energy and component it needs to function. In most unicellular organisms, the number of signal transduction and series of chemical reactions influences the number of ways the cell can react and respond to the environment. It has been noted that the

knowledge of signalling transduction and metabolic pathways in *Plasmodium* are fundamental to aid the design of new strategies against malaria (Doering, 1997; Koyama *et al*., 2009).

For the malaria parasites, one of the most commonly used computational method for analyzing microarray gene expression data is clustering. This has been used by LeRoch *et al* (LeRoch *et al.,* 2003) and Bozdech *et al* (Bozdech *et al.*, 2003a). The results obtained have been used to classify and support genes classification into functional modules, namely metabolisms and metabolic pathways. The results obtained have left us with many putative functional genes. The Malaria Parasite Metabolic Pathways (http://sites.huji.ac.il/malaria, 2011), also accessible from plasmoDB (http://www.plasmodb.org, 2011), provides limited information about this. Recent works like Gangman *et al* (Gangman *et al.*, 2007) and Zhou *et al* (Zhou *et al.*, 2005) introduce the use of Gene Ontology [GO] but the results are also still very limited in their application to *P. falciparum* (Oyelade *et al.*, 2008). This is because only a minority of *P. falciparum* proteins is annotated by GO terms.

An extensive analysis of the available protein-protein interaction (LaCount *et al.*, 2005) for *P. falciparum* is not available. And none of the main chains of signal transduction pathways in *P. falciparum* is presently known. The available knowledge about protein interactions and gene co-regulation in a single specie can be represented as a weighted graph of protein interactions, whose vertices represent proteins and whose edges represent interactions; each edge is assigned a weight from available experimental data (using the transcriptomic and protein interaction data), indicating the strength of evidence for the existence of the corresponding interaction. A class of protein signalling cascades (or signal transduction pathways) can be described as chains of interacting proteins, in which protein interactions enable each protein in the path to modify its successor so as

to transmit biological information. Such structures correspond to simple paths in the protein interaction weighted graph (Scott *et al.*, 2006). Identifying biologically meaningful simple paths corresponding to signalling pathways is very straightforward since in most of the signalling cascades the proteins would transmit the signal from the membrane, where the signal is initiated, towards the nucleus by activation of transcription factors, which in turn lead to transcription of the final effectors.

For the first time, to mine out chains of signal transduction pathways in *P. falciparum*, in this research work, we implemented and applied the techniques developed and deployed to the yeast protein network by Scott *et al* (Scott *et al.*, 2006). We consider a new modified biologically motivated extension of the basic path-finding problem. This is essential for application to organisms where not many experimentally validated protein interactions are known, such as *P. falciparum*. Recent work by Bebek G. *et al* (Bebek and Yang, 2007) presented alternative techniques to solving the path-finding problem, but all of these methods suffer from the problem of sparse availability of data. In *P. falciparum*, 60% of its proteins lack resemblance to any existing annotated organism (Gardner *et al.*, 2002), which illustrates the dimension of this problem.

There have been several attempts to automate the reconstruction of metabolic pathways (Kanehisa, 2002; Ron *et al.*, 2008; Green and Karp, 2007; Pinney e*t al.*, 2005 and Pinney e*t al*., 2007). A recent review (Health e*t al*., 2010) and comparison with a manual reconstruction of the pathways (Ginsburg, 2006) shows how in-appropriate the present tools are and also how very limited the manually curated database is. To improve the tools, the review (Health e*t al*., 2010), recommend that the available information on the biochemistry of the parasite should always be considered when attempting to reconstruct its metabolic pathways or when filtering out erroneous pathways

generated by these tools. In line with these suggestions, a number of computation approaches have been developed to find paths in a metabolic network. An overview on these approaches can be found in Health *et al*. (Health e*t al*., 2010) , Pitkaenen *et al*. (Pitkaenen *et al*.,2009) and Planes and Beasley (Planes and Beasley, 2009). Methods developed so far can be classified into two classes. One that view paths from a source reaction to a target reaction, denoted as R-R case and the second view paths from a source compound to a target compound, denoted also as the C-C case. It is interesting to note that the first work (Croes e*t al*., 2006; Croes e*t al.*, 2005) that introduced the R-R concept has been the most effective of all paths finding approaches presented to date in literature (Planes and Beasley, 2009). Furthermore, metabolic pathways can either be linear or nonlinear (such as pentose phosphate pathway). Only two methods presently can extract non-linear pathways (Health e*t al*., 2010; Pitkaenen *et al*.,2009) and none of these methods (either for extracting linear or non-linear pathway) has been applied to the analysis of *P. falciparum* biochemical network as available in MPMP (Ginsburg, 2006), KEGG (Kanehisa, 2002) and BioCyc (Karp e*t al.*, 2005). It is important to note that all the methods developed so far are KEGG based and these methods are not flexible for usage with another biochemical network.

Also in this research work, we adapted the algorithm we developed for mining chains of signal transduction pathways in the first work to *P. falciparum* metabolic weighted graphs (networks). The weights are calculated using the metabolite degrees (Croes e*t al*., 2006). Such graphs were built from BioCyc (easily updated with MPMP) and KEGG. Adopting the representation of the biochemical metabolic network as we have in Koenig *et al*. (Koenig *et al*., 2006), we are able to make our algorithm tenable to accept metabolic network from other sources apart from KEGG.

This gives us opportunity for the first time to compare the metabolic pathways extracted from different metabolic networks.

1.2 Statement of the problem

A major challenge of post-genomic biology is to understand the complex networks of interacting genes, proteins and small molecules that give rise to biological form and function. Protein-protein interactions (for example, integrated with transcriptional data) and biochemical metabolic networks (overlaid also with transcription data) but processed in alignment over an organism reference metabolic maps, resulting into composite graphs, are crucial to the assembly of protein machinery and the formation of protein signalling cascades. Hence, the dissection of protein interaction and biochemical metabolic networks has great potential to improve the understanding of cellular machinery and to assist in deciphering protein function.

1.3 Aim and Objectives of the study

The overall aim of this research work is to investigate the use of graph-based techniques to integrate information, express relationships and make inferences or predictions on biological processes, motivated by data generation in genomics, transcriptomics, metabolomics and proteomics. This aim will be realized through the following objectives:

- To construct graph-based network structures for the protein-protein interactions and biochemical metabolic networks.
- To use composite, graph-based biologically motivated network structure for the prediction of genes into functional modules – Signalling and Metabolic pathways.

1.4 Methodology

In view of the background philosophy and the scope of the research, the following methods would be used to accomplish the stated objectives. Essentially, signaling and metabolic interaction networks using graph-based techniques are established.

Protein-protein interaction data is obtained from the work of LaCount *et al* (LaCount *et al.*, 2005). In their result, we have 2846 interactions between 1309 proteins. In addition to the protein-protein interaction data, the transcriptional data from LeRoch *et al.* (LeRoch *et al.,* 2003) and Bozdech *et al.* (Bozdech *et al.*, 2003a) was also used to measure the interaction reliabilities, depicted by the edges. To mine signaling pathways from this resulting network, the techniques developed and deployed to the yeast protein network by Scott *et al* (Scott *et al.*, 2006) was applied and then consider a new modified biologically motivated extension of the basic path-finding problem to be able to deal with organisms of sparsely populated experimentally verified protein interactions such as the malaria parasite.

The metabolic (network) graph representation in Koenig *et al*. (Koenig *et al*., 2006) was used. In this work, a graph was established by defining neighbours of metabolites. Two metabolites are neighbours if and only if an enzymatic reaction exists that needs one of the metabolites as input (needed substrate) and produce the other as output (product). We downloaded PlasmoCyc version 14.6 for *P. falciparum* 3D7 from Biocyc.org on the 28th July, 2010 and biochemical metabolic files for the *P. falciparum* 3D7 last updated 22nd December, 2010 from KEGG. Based on the graph representation depicted in figure 3.4 of chapter three, from PlasmoCyc, we have 608 compounds and 824 reactions. And from KEGG, we have 3011 compounds and 3524 reactions. We found that not all compounds used in the reactions listed for *P. falciparum* 3D7 in KEGG database are listed in the compounds list for *P. falciparum* 3D7 in KEGG database. Therefore, the

file containing all compounds was used and found 6516 compounds and 4126 reactions. Note that the reactions that were ignored due to the fact we could not find all compounds listed in their definitions are now accounted for.

Note that the graph formation above is bipartite, having two type of nodes, namely compounds and reactions. We transformed this graph formation into one with a single type of node, namely reaction. This representation leads to Reaction-Reaction (R-R) case way of viewing pathway, that is, from a source reaction to a target reaction. Converting the bipartite graphs from PlasmoCyc and KEGG (the one with 6516 compounds and 4126 reactions) to our R-R representation, we have a dense graph of 824 reaction nodes with 40299 edges and another heavily dense graph of 4126 reaction nodes with 780560 edges. We assign weights to the edges on our two graphs using metabolite degrees (Croes e*t al*., 2006).

We adapted the algorithm developed for mining signalling pathways to the resulting *P. falciparum* metabolic weighted graphs (networks) above to extract metabolic pathways.

1.5 Significance of the study

Large-scale interaction detection methods have resulted in a large amount of protein-protein interaction data. And biochemical research has elucidated an increasingly complete image of the metabolic architecture. Studying the resulting networks can help biologists to understand principles of cellular organization and biochemical phenomenon. Functional modules as a critical level of biological hierarchy and relatively independent units play a special role in biological networks. Since network modules do not occur by chance, identification of modules is likely to capture the biologically meaningful interaction. Naturally, revealing modular structures in

biological networks is a preliminary step for understanding how cells function and how proteins organize into a system.

Due to growing resistance of the malaria parasite to existing chemotherapy, there is a huge and urgent need to discover and validate new drug or vaccine targets to enable the development of new treatments. The ability to discover these drug or vaccine targets can only be enhanced from the deep understanding of the detailed biology of the parasite, that is, how cells function and how proteins organize into modules such as metabolic and signal transduction pathways.

1.6 Contribution to knowledge

- The study introduced the use of graph-based linear-time algorithm for finding paths in a network of Protein-Protein Interactions in *Plasmodium falciparum*. From this, we predicted several important signalling transduction pathways in *Plasmodium falciparum*. We have predicted a viable signalling pathway characterized in terms of the genes responsible that may be the PfPKB pathway recently elucidated in *Plasmodium falciparum*. We obtained from the FIKK family, a signal transduction pathway that ends upon a chloroquine resistance marker protein, which indicates that interference with FIKK proteins might reverse *Plasmodium falciparum* from resistant to sensitive phenotype. We also proposed a hypothesis that showed the FIKK proteins in this pathway as enabling the resistance parasite to have a mechanism for releasing chloroquine(via an efflux process). Furthermore, we also predicted a signalling pathway that may have been responsible for signalling the start of the invasion process of Red Blood Cell(RBC) by the merozoites. It has been noted that the understanding of this pathway will give insight into the parasite virulence and will facilitate rational vaccine design against merozoites invasion. And we

have a host of other predicted pathways, some of which have been used in this work to predict the functionality of some proteins.

- And lastly, our graph-based linear time algorithm for finding paths was extended to find application in biochemical metabolic network to extract metabolic pathways. Doing this, this gives us opportunity for the first time to compare the metabolic pathways extracted from different metabolic networks. We run our algorithm (for four selected pathways: Pyruvate, Glutamate, Glycolysis and Mitochondrial Tricarboxylic Acid Cycle (TCA)) on graph from KEGG and compare our results with the results obtained from recent algorithms: ReTrace and atommetanet. Our results compare favourably with these two algorithms. Considering the results with genes classified into these pathways from Plasmodb, resulted into a lot of false positiveness. Furthermore, we compare the runs of our algorithm on graphs from KEGG and PlasmoCyc (from BioCyc). The results are remarkably different and the results from PlasmoCyc produce less false positiveness when compared to the results from Plasmodb. We identify 2, 1, 2, 4 gene(s) in addition to belong to these four pathways respectively. Some of the genes have not been classified earlier to any known metabolic pathways.

1.7 Limitation of the scope of the study

Our present work has given lead to several future studies. To further address the problem of data scarcity (in particular with regard to the protein–protein interaction information available for the malaria parasite), we need to develop techniques to deal with missing edges, i.e. protein–protein interaction that have never been observed but exist in reality. One way to do this is to integrate transcription factors into the derived network, resulting into what has been called an integrated

cellular weighted network of transcription–regulation and protein–protein interaction (Yerger-Lotem *et al*., 2004). For the malaria parasite *P. falciparum*, only about a third of the number of transcription-associated proteins (TAPs) usually found in the genome of a free-living eukaryote is presently known (Coulson *et al*., 2004).

Presently (at the time of this work), from plasmodb, for *P. falciparum*, we have 137 metabolic pathways covering 2521 genes. This is just about half of the annotated genes of *P. falciparum*. Therefore, there is a need to deploy our techniques at a large scale for all known pathways. This, we know from findings, will help to both reconfirm existing classifications and classify genes of unknown functions into functional modules - metabolic pathways. We also need to find paths in attempts to engineer the finding of unknown metabolic pathways in *P. falciparum.*

1.8 Arrangement of the Thesis

The rest of the thesis is organized as follows; Chapter 2 describes the literature review; the overview of the malaria parasite, review of the clustering tools, the system biology and network modelling on the properties of biological networks. Chapter 3 describes the research methodology and chapter 4 explains the various results generated from both signalling and metabolic networks. Chapter 5 gives the conclusion of the study and discusses future research directions in this area.

CHAPTER TWO

LITERATURE REVIEW

2.1. Overview of the Malaria Parasites

Approximately 300 million people worldwide are affected by malaria and between 1 and 1.5 million people die from it every year. Previously extremely widespread, the malaria is now mainly dominant in Africa, Asia and Latin America. The problems of controlling malaria in these countries are aggravated by inadequate health structures and poor socio economic conditions. The situation has become even more complex over the last few years with the increase in resistance to the drugs normally used to combat the parasite that causes the disease.

Malaria is caused by protozoan parasites of the genus ***Plasmodium*.** Four species of *Plasmodium* can produce the disease in its various forms:

- *Plasmodium falciparum*
- *Plasmodium vivax*
- *Plasmodium ovale*
- *Plasmodium malaria*

Currently, *Plasmodium vivax* and *Plasmodium falciparum* are "the most commonly encountered malarial parasites" (Carter and Kamini, 2002). *P. vivax* is found in nearly all areas where malaria is endemic and is the only one of the four species whose range expands into the temperate regions (John and William, 2006). *P. falciparum*, on the other hand, is found only in the tropic and subtropic regions, though its prevalence in the tropics is high (Carter and Kamini, 2002). *P. malariae* is seen less frequently than either *P. vivax* or *P. falciparum*, but it is found in the same regions where *P. vivax* or *P. falciparum* are found. Lastly, *P. ovale* is prominent throughout tropical Africa and on the West African coast; however, its distribution is the most limited of the

four human malarial parasites (John and William, 2006; Carter and Kamini, 2002). *P. vivax*, *P. malariae*, and *P.ovale* are associated with a low risk of death, whereas *P. falciparum* carries a high risk of fatality.

P. falciparum is the most widespread and dangerous of the four: untreated it can lead to fatal cerebral malaria. Malaria parasites are transmitted from one person to another by the female anopheline mosquito. The males do not transmit the disease as they feed only on plant juices. There are about 380 species of anopheline mosquito, but only 60 or so are able to transmit the parasite. Like all other mosquitos, the anophelines breed in water, each species having its preferred breeding grounds, feeding patterns and resting place. Their sensitivity to insecticides is also highly variable. Plasmodium develops in the gut of the mosquito and is passed on in the saliva of an infected insect each time it takes a new blood meal. The parasites are then carried by the blood into the victim's liver where they invade the cells and multiply. After 9-16 days they return to the blood and penetrate the red cells, where they multiply again, progressively breaking down the red cells. This induces bouts of fever and anaemia in the infected individual. In cerebral malaria, the infected red cells obstruct the blood vessels in the brain. Other vital organs can also be damaged, leading rapidly to the death of the patient.

2.1.1 *Plasmodium falciparum* Life Cycle

The life cycle of *Plasmodium* (John and William, 2006; Carter and Kamini, 2002) can be divided into two distinct phases: **the asexual cycle in humans** and **the sexual cycle in mosquitoes**. To begin the asexual cycle in humans, an infected female *Anopheles* mosquito injects sporozoites into the new human host during a blood meal. Sporozoites injected into the bloodstream leave the blood vascular system within 30 to 40 minutes and enter the liver. This begins the exo-erythrocytic

stage of the life cycle during which asexual multiplication occurs. Within hepatocytes, the sporozoites undergo many nuclear divisions to become schizonts. This occurs over a period of 6 to 15 days, after which the schizonts burst and release thousands of merozoites into the circulation. This marks the end of the exoerythrocytic cycle.

Upon release, the merozoites invade the red blood cells where they undergo another asexual cycle called erythrocytic schizogony. This is also known as the erythrocytic cycle. During this stage the merozoites develop to form immature or ring stage trophozoites which then progress to mature trophozoites. The mature trophozoites develop into schizonts. The erythrocytic cycle results in the formation of 4 to 36 new parasites in each infected cell within a 44 to 72 hour period. At the end of the cycle, the infected red blood cells burst, releasing the merozoites. At this stage, merozoites can either infect new red blood cells to begin the erythrocytic cycle again, or, through the action of some unknown factor, the merozoites can develop into gametocytes. It is of note that blood stage parasites are responsible for the clinical symptoms of malaria. For example, lysis of the red blood cells is an important cause of malaria-associated anemia. In addition, if a significant number of infected cells rupture simultaneously, the resulting material in the bloodstream is thought to induce a malarial paroxysm.

In the case of the sexual cycle in mosquitoes, when a female *Anopheles* mosquito takes a blood meal from an infected person, both male (microgametocytes) and female (macrogametocytes) may be ingested. The microgametocytes and macrogametocytes mature to become microgametes and macrogametes, respectively. In the midgut of the mosquito, the microgametes fertilize the macrogametes, forming a zygote. The zygote becomes elongated and motile, and is then called an ookinete. The ookinetes invade the midgut wall of the mosquito where they develop into oocytes.

The oocytes grow and develop and finally rupture to release sporozoites as depicted in figure2.1a. The sporozoites make their way to the salivary glands of the mosquito so that they can be inoculated in to the new human host during the mosquito's next blood meal, thus perpetuating the *Plasmodium* life cycle as shown in figure 2.1b.

In summary, malaria parasites undergo three distinct asexual replicative stages (exoerythrocytic schizogony, blood stage schizogony, and sporogony) resulting in the production of invasive forms (merozoites and sporozoites). A sexual reproduction occurs with the switch from vertebrate to invertebrate host and leads to the formation of the invasive ookinete. All invasive stages are characterized by the apical organelles typical of apicomplexan species.

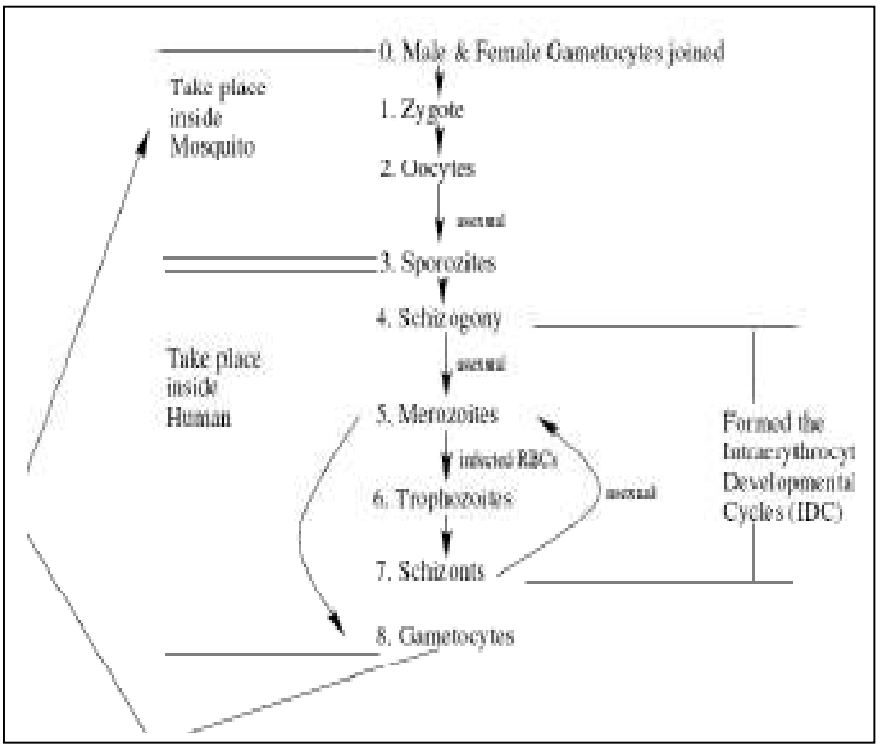


Fig. 2.1a: The asexual cycle in humans and **the sexual cycle in mosquitoes (Adebiyi, 2006)**

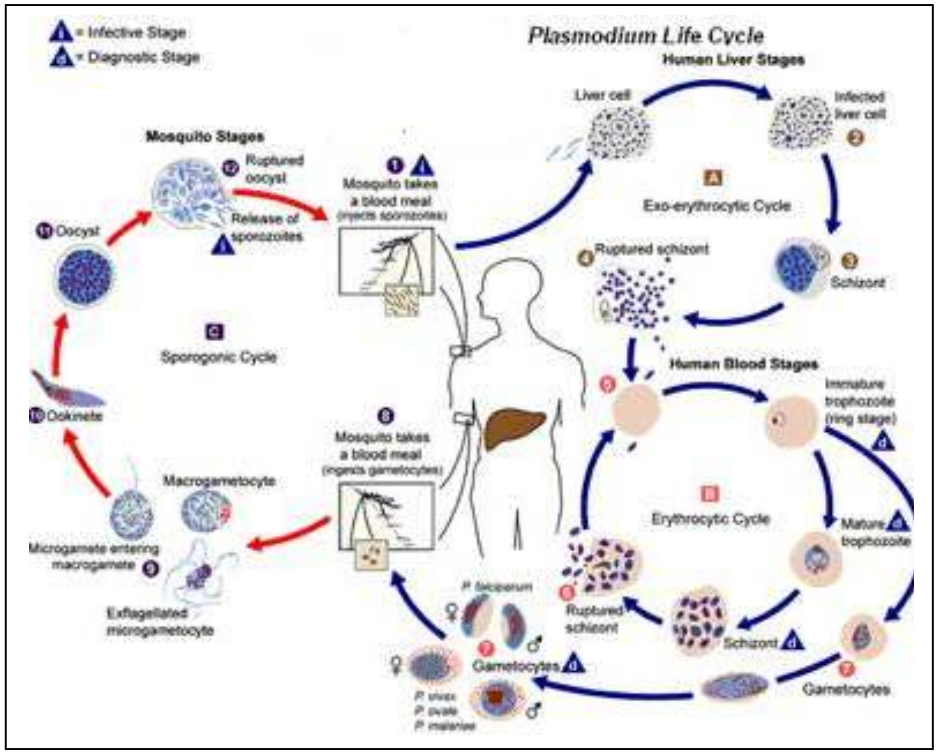


Fig. 2.1b: Life cycle of the parasite *P. falciparum* (http://www.dpd.ede.gov, 2011)

2.1.2 Malaria Parasite-Host Protein-Protein Interactions

Protein-protein interactions (PPIs) are of interest when applying molecular methods to combat *P. falciparum* infection, due to the fact that these connections are utilised by the malaria parasite during vital periods of its life cycle. It has been predicted that 516 PPIs occur between *Homo sapiens* and *P. falciparum* (Dyer *et al.*, 2007). If some of these fundamental interactions could be disrupted, the parasite would not be able to complete specific stages of development and would perish, thereby alleviating malaria infections in the human population. PPIs within the erythrocytic stage of the *P. falciparum* life cycle are vital, as all the pathogenesis associated with malaria occurs during this time. The invasion of and growth within host red blood cells is therefore an important factor in the disease which, if disrupted, could prevent infection.

2.1.2.1 Protein-Protein Interactions during invasion

The erythrocytic stage is initiated when merozoites come into contact with the red blood cell membrane and specialised secretory organelles – micronemes and rhoptries – release parasitic proteins that facilitate invasion of the host cells. These two types of organelles are located in the apical end of invasive merozoites as shown in figure 2.2.

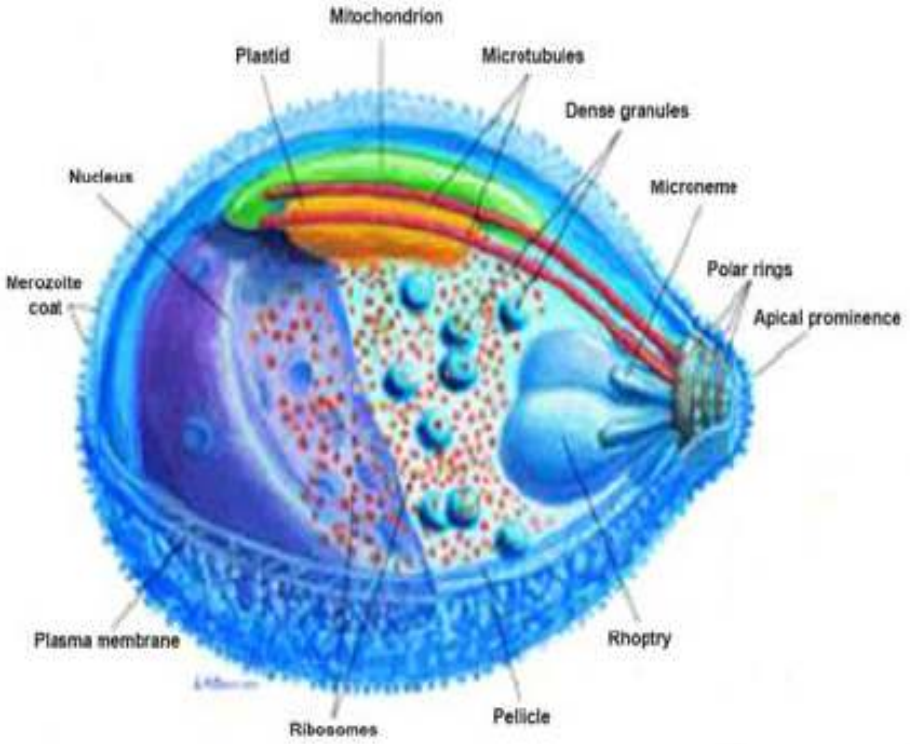


Fig. 2.2: 3D Structure of the Merozoite from the asexual cycle of *P. falciparum* (Bannister *et al*., 2000).

2.1.2.2 Protein Kinases (PKs)

PKs are enzymes that catalyse the phosphorylation of proteins within eukaryotic cells. This regulates protein function – activating, modulating or deactivating the molecules – and thereby controls cell behaviour. Of the proteins expressed in an average eukaryotic cell, almost 33 percent contain covalently bound phosphate molecules (Hubbard and Cohen, 1993). Approximately 3 percent of all eukaryotic genes code for PKs and these are classified according to structural similarity, as well as parallel substrate specificity and mode of regulation (Hanks and Hunter, 1995).

The reaction catalysed by PKs is:

$$Protein + ATP \overset{PK}{\rightarrow} Phosphoprotein + ADP$$

Eukaryotic PKs are divided into seven established groups, with the two main sub-divisions being the protein-serine/threonine kinases and the protein-tyrosine kinases (Hanks and Hunter, 1995). The entire complement of PKs encoded in a genome is termed the kinome.

2.1.2.3 *Plasmodium falciparum* protein kinases

Phosphorylation and dephosphorylation processes play an important role in the life cycle of the malaria parasite (Suetterlin *et al.*, 1991). This is especially true for the intraerythrocytic stage which is accompanied by a distorted phosphorylation pattern of the host RBC membrane (Chishi *et al.*, 1994). This vital stage of the parasite lifecycle is prevented by PK inhibitors (Ward *et al.*, 2004; Anamika *et al.*, 2005).

According to Ward *et al* (Ward *et al.*, 2004) – who identified 65 malaria PK sequences – and Anamika *et al* (Anamika *et al.*, 2005), who identified 99 PKs in the *P. falciparum* genome using various amino acid sequence profile matching algorithms, several of the parasite sequences did not cluster within any of the known eukaryotic PK groups. Furthermore, the highest number of malarial sequences were those involved in the control of cell proliferation, namely the cyclin-

dependent- (CDK), mitogen-activated- (MAPK), glycogen-synthase- (GSK) and CDK-like kinases, along with a kinase family that includes PKs A, G and C (AGC). Interestingly, no malarial PK clustered with the tyrosine kinase group; homologues of MEK (MAPK kinase), MEKK (MAPK kinase kinase) and PKC-like kinases were also lacking in the *P. falciparum* genome. A splinter group of 20 PK-related sequences formed a novel family called FIKK, which seems to be restricted to the *Apicomplexa* (Ward *et al.*, 2004). According to Schneider and Mercereau-Puijalon (Schneider and Mercereau-Puijalon, 2005), even though kinase activity has not been demonstrated in this group, the presence of most of the amino acids necessary for phosphor transfer indicates an enzymatic role. Nunes *et al* (Nunes *et al.,* 2007) provided experimental evidence of kinase activity and transport of some FIKKs to the erythrocyte.

The large divergence in the kinome of *P. falciparum* compared to that of humans (Nunes *et al.,* 2007) suggests that exclusive targeting of parasite enzymes is possible. This is promising as PKs play crucial roles in most cellular processes and thus their targeted inhibition could incapacitate the parasite and prevent disease progression.

2.2 The Clustering Tools

For the malaria parasites, one of the most commonly used computational method for analyzing microarray gene expression data is clustering. This has been used by Bozdech *et al* (Bozdech *et al.*, 2003a) and LeRoch *et al* (LeRoch *et al.,*2003). The results obtained have been used to classify and support genes classification into functional modules, namely metabolisms and metabolic pathways. A gene expression data set from microarray experiment can be represented by a real-valued expression matrix $M = \{w_{ij} \mid 1 \leq i \leq n, 1 \leq j \leq m\}$, where the rows ($G = \{\vec{g_1},...,\vec{g_n}\}$) form the expression patterns of genes, the columns ($T = \{\vec{t_1},...,\vec{t_m}\}$) represent the

expression profiles of timepoints, and each cell w_{ij} is the measured expression level of gene $\boldsymbol{i}$ in timepont $\boldsymbol{j}$.

2.2.1 Challenges of gene clustering

Due to the special characteristics of gene expression data, and the particular requirements from the biological domain, gene-based clustering presents several new challenges and is still an open problem.

- Cluster analysis is typically the first step in data mining and knowledge discovery. The purpose of clustering gene expression data is to reveal the natural data structures and obtain some initial insights regarding data distribution. Therefore, a good clustering algorithm should depend as little as possible on prior knowledge, which is usually not available before cluster analysis.
- Second, due to the complex procedures of microarray experiments, gene expression data often contain a huge amount of noise. Therefore, clustering algorithms for gene expression data should be capable of extracting useful information from a high level of background noise.
- Third, an empirical study has demonstrated that gene expression data are often "highly connected" (Jiang *et al.*, 2003a), and clusters may be highly embedded in one another (Jiang *et al.*, 2003a). Therefore, algorithms for gene-based clustering should be able to effectively handle this situation.

Finally, users of microarray data may not only be interested in the clusters of genes, but also be interested in the relationship between the clusters and the relationship between the genes within the same cluster. A clustering algorithm, which can partition the data set but also provide some graphical representations of the cluster structure (intra- and inter- relationship wise) would be more favoured by the biologists.

2.2.2 Review of existing works

There are several approaches proposed in literature, dealing with identification of clusters of functionally related genes in genomes (Bozdeck *et al*., 2003a; LeRoch *et al.,*2003). According to these approaches, gene-based clustering technique, using microarray data where rows represent the various genes and columns represent the various timepoints employed in the experiments, were able to capture gene expressions that are correlated into the same cluster. However, they could not partition genes of the same biological pathways into the same cluster.

Also, in another work by Gangman *et al.* (Gangman *et al*., 2007), C-Hunter clustering algorithm was applied to the genomes of *Escherichia coli* and *S. cerevisiae* of eukaryotic species. In this work, they showed that clusters identified with this algorithm corresponded to well-documented metabolic pathway clusters. But this algorithm is very ineffective and could not classify genes of *plasmodium species* into their various metabolic (biological) pathways.

Different methods based on modelling with a graph have been developed for analyzing the network structures of PPI networks. Hierarchical clustering methods have been proven to be a good strategy for metabolic and PPI networks. Ravasz *et al*. (Ravasz *et al.*, 2002) analyzed the hierarchical organization of modularity in metabolic networks, and authors of (Brun *et al.*, 2003; Rives A. and Galiitski, 2003 and Lu *et al.*, 2004) applied three different clustering methods respectively, based on different metrics induced by shortest distance, graphical distances, and probabilistic functions, to analyze the module structure of the yeast protein interaction networks on a clustering tree. Several papers (Spirin and Mirny, 2003; Bader and Hogue, 2003 and Bu *et al.*, 2003) have also shown that network modules which are densely connected within themselves but sparsely connected with the rest of network generally correspond to meaningful biological units

such as protein complexes and functional modules. Bu *et al* (Bu *et al.*, 2003) found 48 functional modules in budding yeast by applying a spectral analysis method. Prediction methods of protein complexes which generally correspond to dense subgraphs in the network have been proposed by (Spirin and Mirny, 2003; Bader and Hogue, 2003 and Bu *et al.*, 2003). Several approaches to network clustering that have been used for the analysis of PPI networks, including edge-betweenness clustering (Dun *et al.*, 2005), identification of k-scores (Bader and Hogue, 2003), restricted neighborhood search clustering (RNSC) (King *et al*., 2004) and Markov clustering algorithm (MCL) (Pereira-Leal *et al.*, 2004). Spirin and Mirny (Spirin and Mirny, 2003) detected about 50 network modules by using a combination of three methods and most of which have been proven to be protein complexes or functional modules.

Recently, a novel network clustering method (clique Percolation Method CPM) based on clique percolation has been developed (Palla *et al.*, 2005). It can reveal overlapping module structure of complex networks. But a distinct shortcoming of its application in PPI networks lies in that the method may be restrictive since the basal element of the method is a 3-clique structure. For example, the spoken-like module can not be detected and when the method is applied to large sparse PPI networks, only a few modules can be detected. Stuffen *et al* (Stuffen *et al.*, 2002) studied the problem of identifying pathways in a protein network. They applied an exhaustive search procedure to an unweighted interaction graph, considering all interactions equally reliable. The approach was successful in detecting known signalling pathways in yeast. Also Scott *et al* (Scott *et al.*, 2006) extended the work of Stuffen *et al* by applying color coding technique to an interaction graph of yeast protein network.

2.3 System Biology (Network Biology)

Systems Biology can be defined as an approach to biology where organisms and **biological processes** should be analyzed and described in terms of their **components and their interactions** in a framework of mathematical models (Per, 2003).

In functional genomics, one often uses statements such as 'gene or protein X performs function Y', for example 'the leptin protein regulates the amount of body fat'. But when one looks at this statement, it is clear that it is fundamentally misleading. In the given example, it is clear that the leptin protein is not a machine in itself that computes and performs the regulatory action. Rather, the leptin molecule is a component in a larger system, and it is that system that performs the regulatory function.

Systems Biology begins in the insight that biological processes must be understood in terms of the components that participate in the processes, and that the complexity of biological systems make it difficult, if not impossible, to understand the workings of the system by simple qualitative arguments. **Mathematically strict models** must be formulated. This is required both in order to be able to capture the actual behaviour of the system with acceptable precision, but also to be able to analyze the fundamental behaviour of the system. The mathematical models may be very simple (Boolean on/off), or very complex (including detailed descriptions of interactions at a molecular level). The important issue is that it should be possible to analyze the model, either by some mathematical approach, or to simulate it, in order to evaluate its correspondence with the observed facts (Per, 2003).

Paradoxically, the complexity of biology is the basis for the development of Systems Biology, at the same time as it is the main reason why computational approaches to biological processes have

not been particularly successful in the past. However, the appearance of bioinformatics and functional genomics, and their results (complete genomes, microarray expression analysis, etc) has had a great impact. It now appears possible to obtain data that can be used to build sensible models, and to test them. This is probably the main reason why Systems Biology has become so popular in the last few years.

So far, only very limited results have been obtained. There are only few, well-studied systems on which any deep analysis has been done. However, there are already some insights that may prove to be generally true. For instance, it seems clear that **robustness** is a very important factor in biological systems. This is the property that allows a system to absorb fairly large perturbations, and still function reasonably well. The functionally important behaviour of a system has a certain degree of resilience to damage. Some studies have pointed to different ways in which evolution have favoured systems that are robust in different ways (Per, 2003).

One important goal of Systems Biology is to understand life processes in sufficient detail to make predictions about their behaviour. If we want to make a particular system behave in a certain way, how should we change the system, or what type of perturbation should we apply? If we want to make a bacterium produce propanol instead of ethanol, then how should we change the metabolic network of the bacterium? Or, if we want to produce a pharmaceutical drug that can help with deficiencies of the insulin regulatory system that is the basis for diabetes type II (obesity-related diabetes), what components should we focus on? Which are the best drug targets? (Per, 2003)

Network biology is a general term for an emerging field that concerns the study of interactions between biological elements (Alm and Arkin, 2003). The term *molecular interaction networks* may designate several types of networks depending on the kind of molecules involved. Classically,

one distinguishes between gene regulatory networks, signal transduction networks and metabolic networks. Protein-protein interaction networks represent yet another type of network. One of the declared objectives of network biology (or systems biology in general) is whole cell simulation (Kitano, 2002). However, studying the dynamics of a network requires knowledge on reaction mechanisms.

Besides the fact that such knowledge is often unavailable or unreliable, the study of the static set of reactions that constitute a biochemical network is equally important, both as a first step towards introducing dynamics, and in itself. Indeed, such static set represents not what is happening at a given time in a given cell but instead the capabilities of the cell, including capabilities the cell does not use. A careful analysis of this set of reactions for a given organism, alone or in comparison with the set of other organisms, may also help to arrive at a better understanding of how metabolism evolves. More precisely, the term "metabolism" should be understood as the static set of reactions involved in the synthesis and degradation of small molecules. A major issue concerning the study of biochemical networks is the problem of their organization. Several attempts have been made to decompose complex networks into parts. These "parts" have been called modules or motifs, but no definition of such terms seems to be completely satisfying. Modules have first been mentioned by Hartwell *et al.* (Hartwell *et al.*, 1999) who outline the general features a module should have but provide no clear definition for it. In the context of metabolic networks, a natural definition of modules could be based on the decomposition of a metabolic network into the metabolic pathways one can find in databases: modules would thus be the pathways as those have been established. The advantage of this definition of a module is that it reflects the way metabolism has been discovered experimentally (starting from key metabolites and studying the ability of an organism to synthesize or degrade them). The drawback is that it is

not based on objective criteria and therefore is not universal (indeed, the number of metabolic pathways and the frontiers between them vary from one database to the other). Several attempts to give systematic and practical definitions have been made using graph formalisms (Guiner and Nunes, 2005; Ma *et al.*, 2004; and Schuster *et al*., 2002) and constraint-based approaches (Papin *et al.*, 2004). Graph based methods ranges from a simple study of the local connectivity of metabolites in the network (Schuster *et al*., 2002) to the maximization of a criterion expressing modularity (number of links within modules) (Guiner and Nunes, 2005). The only information used in these methods is the topology of the network. In the case of constraint-based approaches, the idea is quite different. First, a decomposition of the network into functional sets of reactions is performed by analysis of the stoichiometric matrix (Papin *et al.*, 2004) and then modules are defined from the analysis of these functional states. The result is not a partition in the sense that a single reaction might belong to several modules. Unlike the definition of module, the notion of motif has not been studied in the context of metabolic networks. In general, depending on what definition is adopted for modules and motifs, there is no clear limit between the two notions besides the difference in size. In the context of regulatory networks, motifs have been defined as small, repeated and perhaps evolutionary conserved subnetworks.

In contrast with modules, motifs do not function in isolation. Furthermore, they may be nested and overlapping (Wolf and Arkin, 2003). This definition refers to general features that regulatory motifs are believed to share but it provides no practical way to find them. A more practical definition has been proposed, still in the context of gene regulatory networks (and other types of non-biological networks such as the web or social networks). These are “network motifs” and represent patterns of interconnections that recur in many different parts of a network at frequencies

much higher than those found in randomized networks (Shen-Orr *et al.*, 2002). This definition is purely topological and disregards the nature of the components in a motif. It assumes that the local topology of the network is sufficient to model function (which is understood here as the dynamic behaviour of the motif). This assumption seems acceptable when studying the topology of the internet and may also hold when analyzing gene regulatory networks, but it appears not adapted to metabolic networks.

The complexity of biological systems, and the vast amount of information now available at the level of genes, proteins, cells, tissues and organs, requires the development of mathematical models that can define the relationship between structure and function at all levels of biological organization (Hunter and Borg, 2003).

Modules encourage hierarchical thinking with regard to networks and facilitate the analysis genotype-phenotype relationships. The genotype of an organism can be defined by sequencing methods. The individual genes are transcribed to mRNA and then translated to generate a set of protein products whose individual functions can be characterized. Small-scale modules can be mathematically described from the interaction of these protein components. Large-scale modules might arise from the interaction of several small-scale modules as depicted in the figure 2.3 below.

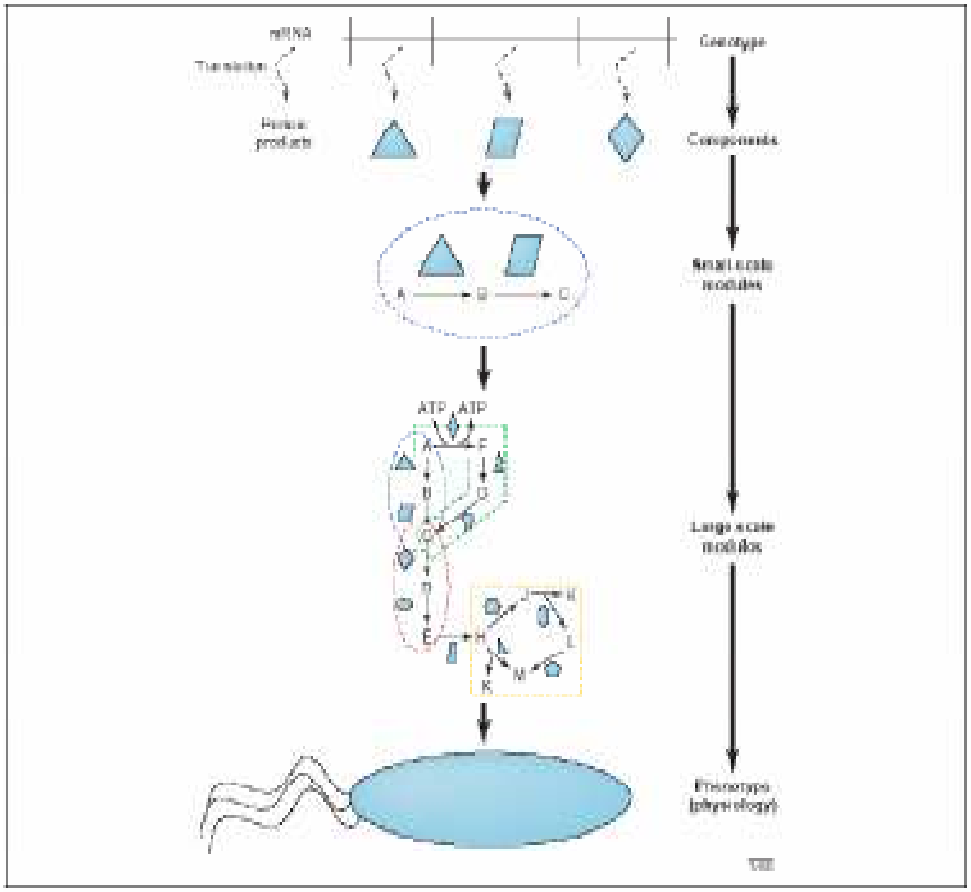


Fig 2.3: Hierarchy in Biological Network. The individual genes are transcribed to mRNA and then to generate a set of protein products whose individual functions can be characterized. Small-scale modules can be mathematically described from the interaction of these protein components. Large-scale modules might arise from the interaction of several small-scale modules. The interaction of such large-scale modules leads to the physiology (phenotype) of an organism (Kitano, 2002).

2.4 Network modelling

Traditionally, the study of complex networks has been the territory of mathematics, especially the graph theory. Initially the graph theory focused on regular graphs, with no apparent design principles were described as random graphs, proposed as the simplest and most straightforward realisation of a complex network. The pioneer of the theory was Leonhard Euler, who studied first regular graphs in 18^{th} century. In the 20th century the theory became much more statistically and algorithmically oriented (Porekar, 2002).

Later in 1950's, graph theory was used to describe large networks, with no particular distributions of nodes and link, whose organization principles were not easily definable. These networks were first studied by Paul Erdös and Alfred Rényi and were called "random graphs", due to their generating method: we start with N nodes and connect every pair of them with probability p. The resulting graph has on average $p\ (N(N-1))/2$ edges distributed randomly. The degree distribution of such graph is Poisson with peak at $P(k)$. This model has guided our thinking for decades after it has been presented (Porekar, 2002).

The topology of real large networks (i.e. Internet, WWW, telephone networks, ecological networks) substantially differs from the topology of random graphs produced by the simple Erdös-Rényi (ER) model, therefore new methods tools and models needed to be developed.
In past years, we witnessed dramatic advances in this direction. The computerization of data acquisition has led to the emergence of large databases on the topology of various real networks. Wide availability of computer power allows to investigate networks containing millions of nodes, exploring questions that could not be answered before as well as the slow but noticeable

breakdown between different science disciplines allows scientists to access different databases, allowing to uncover the generic properties of large networks. Networks found in nature show degree distribution that greatly differs from the Poisson degree distribution of random graphs. Because of existence of a few vertices with high degree, the distribution of real networks has a power-law tail $P(k).k$, which indicates scale free properties (Porekar, 2002).

2.4.1 Random network

The Erdös–Rényi (ER) model of a random network (figure2.4-**A**) starts with N nodes and connects each pair of nodes with probability p, which creates a graph with approximately p $(N(N- 1))/2$ randomly placed links (figure 2.4-**Aa**) (Barabasil and Oltvai, 2004). The node degrees follow a Poisson distribution (figure 2.4- **Ab**), which indicates that most nodes have approximately the same number of links (close to the average degree $<k>$). The tail (high k region) of the degree distribution $P(k)$ decreases exponentially, which indicates that nodes that significantly deviate from the average are extremely rare. The clustering coefficient is independent of a node's degree, so $C(k)$ appears as a horizontal line if plotted as a function of k (figure 2.4-**Ac**). The mean path length l is proportional to the logarithm of the network size, $l \approx \log N$, which indicates that, it is characterized by the small-world property (Barabasil and Oltvai, 2004).

2.4.2 Scale free network

Scale-free networks (figure 2.4-**B**) are characterized by a power-law degree distribution; the probability that a node has k links follows $P(k) \approx k^{-\gamma}$, where γ is the degree exponent (Barabasil and Oltvai, 2004). The probability that a node is highly connected is statistically more significant than in a random graph, the network's properties often being determined by a relatively small number of highly connected nodes that are known as hubs (figure 2.4-**Ba**; blue nodes). In the

Barabási–Albert model of a scale-free network, at each time point a node with M links is added to the network, which connects to an already existing node I with probability $\pi_I = \frac{k_I}{\sigma_j k_j}$ where k_I is the degree of node I (figure 2.4-**Ba**) and J is the index denoting the sum over network nodes. The network that is generated by this growth process has a power-law degree distribution that is characterized by the degree exponent $\gamma = 3$. Such distributions are seen as a straight line on a $log-log$ plot (figure 2.4-**Bb**). The network that is created by the Barabási–Albert model does not have an inherent modularity, so $C(k)$ is independent of k (figure 2.4-**Bc**). Scale-free networks with degree exponents $2<\gamma<3$, a range that is observed in most biological and non-biological networks, are ultra-small, with the average path length following $\ell \approx log\, log\, N$, which is significantly shorter than $log\, N$ that characterizes random small-world networks (Barabasil and Oltvai, 2004).

The origin of the scale-free topology in complex networks can be reduced to two basic mechanisms: **Growth** and **Preferential attachment**. Growth means that the network emerges through the subsequent addition of new nodes, such as the new red node that is added to the network that is shown in part **a**. Preferential attachment means that new nodes prefer to link to more connected nodes. For example, the probability that the red node will connect to node 1 is twice as large as connecting to node 2, as the degree of node 1 (k_1=4) is twice the degree of node 2 (k_2=2). Growth and preferential attachment generate hubs through a 'rich-gets-richer' mechanism: the more connected a node is, the more likely it is that new nodes will link to it, which allows the highly connected nodes to acquire new links faster than their less connected peers. In protein interaction networks, scale-free topology seems to have its origin in gene duplication. Part **b** shows a small protein interaction network (blue) and the genes that encode the proteins (green). When

cells divide, occasionally one or several genes are copied twice into the offspring's genome (illustrated by the green and red circles). This induces growth in the protein interaction network because now we have an extra gene that encodes a new protein (red circle). The new protein has the same structure as the old one, so they both interact with the same proteins. Ultimately, the proteins that interacted with the original duplicated protein will each gain a new interaction to the new protein. Therefore proteins with a large number of interactions tend to gain links more often, as it is more likely that they interact with the protein that has been duplicated. This is a mechanism that generates preferential attachment in cellular networks. Indeed, in the example that is shown in part **b** it does not matter which gene is duplicated, the most connected central protein (hub) gains one interaction. In contrast, the square, which has only one link, gains a new link only if the hub is duplicated (Barabasil and Oltvai, 2004).

2.4.3 Hierarchical network

To account for the coexistence of modularity, local clustering and scale-free topology in many real systems it has to be assumed that clusters combine in an iterative manner, generating a hierarchical network (figure 2.4-**C**). The starting point of this construction is a small cluster of four densely linked nodes (see the four central nodes in figure 2.4-**Ca**). Next, three replicas of this module are generated and the three external nodes of the replicated clusters connected to the central node of the old cluster, which produces a large 16-node module. Three replicas of this 16-node module are then generated and the 16 peripheral nodes connected to the central node of the old module, which produces a new module of 64 nodes. The hierarchical network model seamlessly integrates a scale-free topology with an inherent modular structure by generating a network that has a power-law degree distribution with degree exponent = $1 + \ell n4/\ell n3 = 2.26$ (figure 2.4-**Cb**) and a large, system-size independent average clustering coefficient $\langle C \rangle \approx 0.6$. The most important signature of

hierarchical modularity is the scaling of the clustering coefficient, which follows $C(k) \approx k^{-1}$ a straight line of slope -1 on a $log-log$ plot (figure2.4-**Cc**). A hierarchical architecture implies that sparsely connected nodes are part of highly clustered areas, with communication between the different highly clustered neighbourhoods being maintained by a few hubs (figure2.4-**Ca**) (Barabasil and Oltvai, 2004).

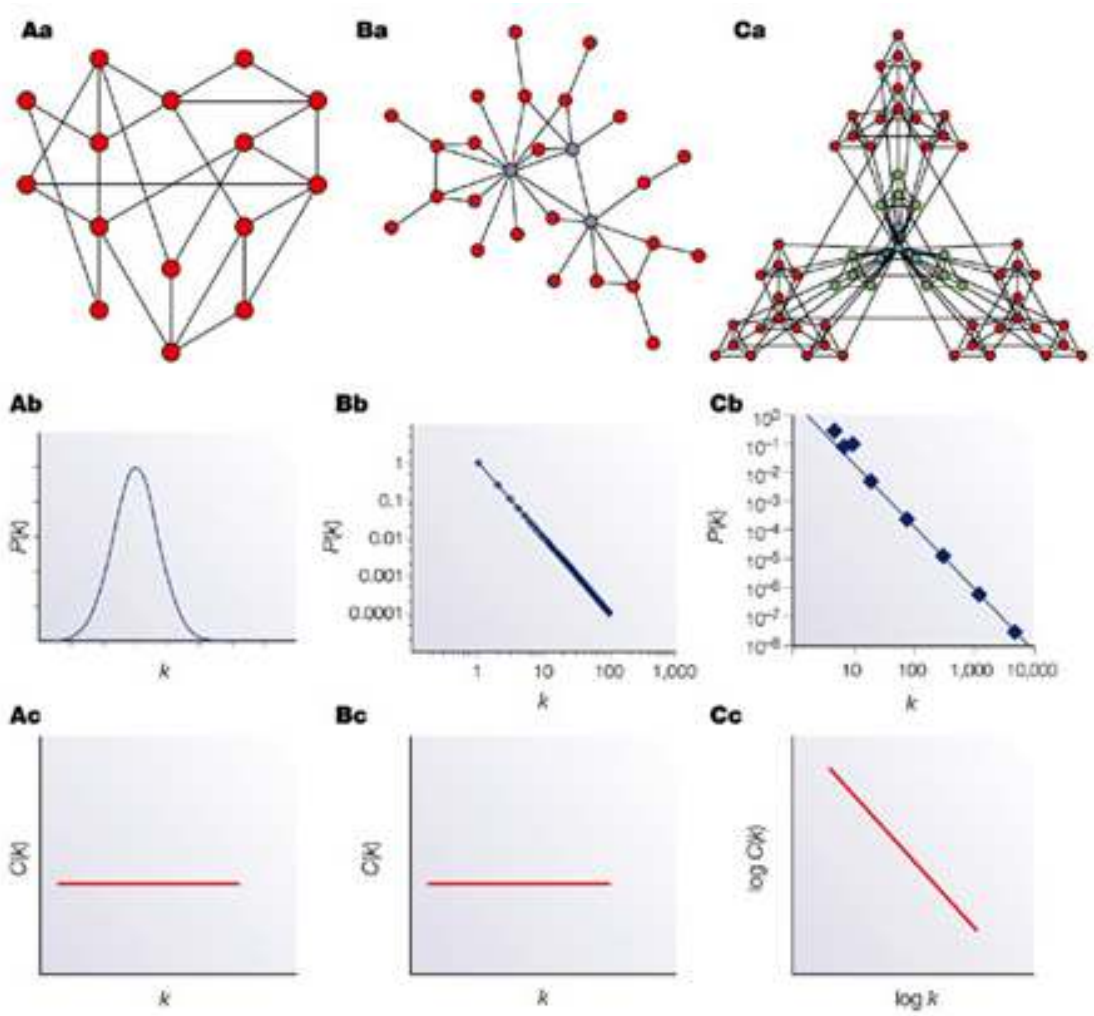


Fig. 2.4.: (A) Random network. In Ab above, the nodes degree follows a Poisson distribution which indicates that most nodes are well represented by the average node with degree <k> and also the C(k) as depicted in Ac appears as a horizontal line if plotted as a function of k. This shows that the clustering coefficient is independent of a node's degree. (B) Scale free network. Ba characterizes a typical nodes degree distribution, the resulting network generated has a power-law distribution as seen as straight line on log-log plot in Bb above. The C(k) is also independent of k as depicted in Bc. (C) Hierarchical network. The hierarchical network model integrates a scale-free topology with an inherent modular structure by generating a network that has a power-law degree distribution in Cb. The most important signature of hierarchical modularity is the scaling of the clustering coefficient, which follows $C(k) \approx k^{-1}$ a straight line of slope -1 on a $log-log$ plot as depicted in fig. Cc (Barabasil and Oltvai, 2004).

2.4.4 Graph Generation Models

A biological network such as a protein-protein interaction or a synthetic lethal interaction network can be represented as an undirected graph $G = (V, E)$, where V is the set of vertices and E is the set of edges. Vertices in such a graph represent proteins while edges represent interactions between the proteins. Let G have n vertices and m edges. Let N_v be the set of neighbours of a vertex v. Let $d_v = |N_v|$ be the degree of vertex v. Researchers have observed that networks such as the Internet, citation patterns in scientific papers and some biological networks are scale free (Barabasi and Albert, 1999). A network is *scale-free*, if its degree distribution follows a power law i.e., the probability $P(k)$ that a vertex in the network has degree k is proportional to $k^{-\gamma}$ for some $\gamma > 0$. A graph generation model is an algorithm or certain steps to follow, so that we may generate graphs with certain properties.

2.4.4.1 Erdös–Rényi Model

Although the Erdös–Rényi model was not initially proposed to explain the evolution or structure of biological networks, we include it since it is a well studied model for the generation of random graphs. An Erdös–Rényi graph $G(n, p)$ is a graph with n vertices such that the probability of having an edge (u, v) in G is p for any vertices u and v in G (Erdos and Renyi, 1960).

Although the Erdös–Rényi model is a well studied model, it does not fully capture the features exhibited by biological networks. The presence of many highly connected hubs is a feature that is observed in biological networks. An Erdös–Rényi graph is unlikely to have such a property, since the probability of occurrence of every edge is the same.

2.4.4.2 Barabási Albert Model

Barabási and Albert (Barabasi and Albert, 1999) conjecture that the scale free property of complex networks arises because:

1. the network grows with the addition of more vertices and
2. a new vertex preferentially attaches itself to vertices with high degree.

The proposed of a growth based model (Barabasi and Albert, 1999) in which a new vertex is created at each time step and the newly arrived vertex preferentially attaches itself to existing vertices with higher degree. Therefore in this case vertices with higher degree have a higher probability of connecting to the new vertex. The probability p_v of creating an edge between an existing vertex v and the newly added vertex is

$$p_v = \frac{(d_v + 1)}{(|E| + |V|)} \qquad 2.1$$

where $|E|$and $|V|$ are, respectively, the number of edges and vertices currently in the network (counting neither the new vertex nor the other edges that it is incident on).

Due to preferential attachment, a vertex with a higher degree will continue to increase its connectivity at a higher rate; this does explain the presence of hubs in such networks.

2.4.4.3 Watts Strogatz Small World Model

The small world model is a graph generating model proposed by Watts and Strogatz (Watts and Strogatz, 1998). Graphs which have the small world property have low characteristic path lengths i.e., the average distance between any two vertices in the graph is small and also high clustering coefficient. The algorithm to generate a graph takes as input a regular graph with n vertices with k edges incident on each vertex and a probability p. The algorithm chooses an edge at random with

probability p, and then one of the end points of the edge is changed to another vertex, again chosen at random.

2.4.4.4 Eppstein Wang Model

Eppstein and Wang (Eppstein and Wang, 2002) proposed a steady state method for generating scale-free networks of web graphs. A steady state model is not a growth based model i.e., the model does not involve the addition of new vertices or edges. The input to the algorithm is the number of edges m, the number of vertices n and a model parameter r. The model starts by generating a graph with n vertices and m edges, by randomly adding edges between the vertices until there are m edges. The algorithm then modifies the initial graph by executing the following sequence of steps r times:

1. Pick a vertex v at random. Repeat this step until $d_v > 0$.
2. Pick an edge $(u, v) \in G$ at random.
3. Pick a vertex x at random.
4. Pick a vertex y proportional to degree of y.
5. If (x, y) is not an edge and if x is not y, then add edge (x, y) to G and remove edge (u, v) from G.

This is a simple model for generating scale-free networks, because it produces a power-law graph without the addition of extra vertices and edges, by evolving the existing graph while maintaining the same number of edges and vertices. Eppstein and Wang simulated the model on graphs with different sizes and different densities, where $density = m/n$. Each simulation was performed five times and the model parameter r was chosen to be 10^7. The degree distribution was observed

to converge to a power law distribution as the value of r increased, for many sizes and densities of the graph.

2.4.5 Types of Networks

The following definitions taken from (Daaron and Asu, 2009) are useful for our discussion.

Social and economic networks: A set of people or groups of people with some pattern of contacts or interactions between them. For examples, Facebook, friendship networks, business relations between companies, intermarriages between families, labor markets.

Information networks: Connections of "information" objects. For examples, Network of citations between academic papers, World Wide Web (network of Web pages containing information with links from one page to other), semantic (how words or concepts link to each other)

Technological networks: Designed typically for distribution of a commodity or service.

Infrastructure networks: e.g., Internet (connections of routers or administrative domains), power grid, transportation networks (road, rail, airline, mail)

Temporary networks: e.g., ad hoc communication networks, sensor networks, autonomous vehicles.

Biological networks: A number of biological systems can also be represented as networks. For examples, Food web, protein interaction network, network of metabolic pathways.

We represent a network by a graph (N, g), which consists of a set of nodes $N = \{1, \ldots, n\}$ and an nxn matrix $g = [g_{ij}]$ $i, j \in N$ (referred to as an adjacency matrix), where $g_{ij} \in \{0,1\}$ represents the availability of an edge from node i to node j.

The edge weight $g_{ij} > 0$ can also take on non-binary values, representing the intensity of the interaction, in which case we refer to (N, g) as a weighted graph.

We refer to a graph as a directed graph (or digraph) if $g_{ij} \neq g_{ji}$ and an undirected graph if $g_{ij} = g_{ji}$ for all $i, j \in N$ as shown in figure 2.5a.

We consider "sequences of edges" to capture indirect interactions. For an undirected graph (N, g), we have the following definitions:

- A **walk** is a sequence of edges $\{i_1, i_2\}, \{i_2, i_3\}, \ldots, \{i_{2k-1}, i_{2k}\}$.
- A **path** between nodes i and j is a sequence of edges $\{i_1, i_2\}, \{i_2, i_3\}, \ldots, \{i_{2k-1}, i_{2k}\}$ such that $i_1 = i$ and $i_k = j$, and each node in the sequence $i_1, \ldots, i_k$ is distinct.
- A **cycle** is a path with a final edge to the initial node.
- A **geodesic** between nodes i and j is a "shortest path" (i.e., with minimum number of edges) between these nodes. A path is a walk where there are no repeated nodes. The length of a walk (or a path) is the number of edges on that walk (or path).

The diagrams that illustrate these are depicted in figure 2.5b (the red colour display the interested paths).

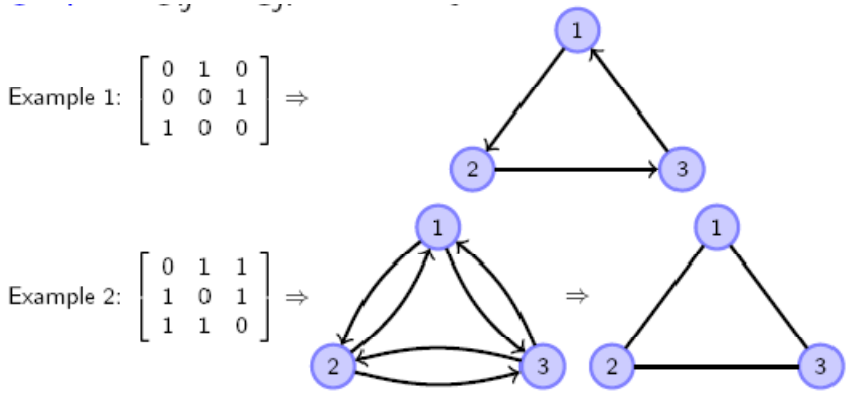


Fig. 2.5a: A Simple directed and undirected graphs (Daaron. and Asu., 2009)

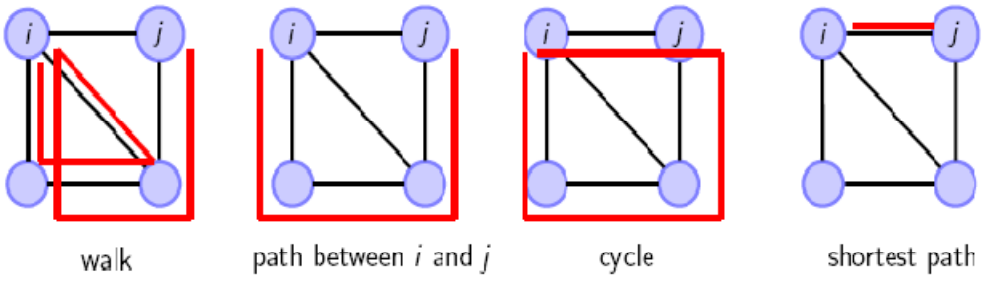


Fig. 2.5b: A graph that denotes walk, path, cycle and shortest path (Daaron and Asu, 2009)

2.4.6 Neighborhood and Degree of a Node

The **neighborhood** of node i is the set of nodes that i is connected to. While the **degree** of a node in a network (sometimes referred to as the connectivity) is the number of connections or edges the node has to other nodes. If a network is directed, meaning that edges point in one direction from one node to another node, then nodes have two different degrees, the in-degree, which is the number of incoming edges, and the out-degree, which is the number of outgoing edges (Daaron. and Asu, 2009).

The **degree distribution** $\boldsymbol{P(k)}$ of a network is then defined to be the fraction of nodes in the network with degree k. Thus if there are n nodes in total in a network and n_k of them have degree k, we have $P(k) = n_k/n$.

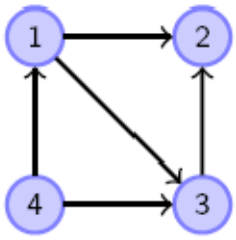
For undirected graphs:

The **degree** of node i is the number of edges that involve i (i.e., cardinality of his neighborhood).

For directed graphs:

Node $i's$ in-degree is $\sum_j g_{ji}$.

Node $i's$ out-degree is $\sum_j g_{ij}$.



For example, in the graph above, the node 1 has in-degree 1 and out-degree 2.

2.4.7 Some properties of networks

- **Small world effect:** A network has the small-world effect if most pairs of vertices are connected by a short path through the network. It is the small average shortest path between two nodes scaling logarithmically with network size. If the number of vertices within a distance r of a typical central vertex grows exponentially with r and this is true of many networks, including the random graph, then the value of l will increase as $log\, n$ increases.
- **Transitivity/clustering:** If vertex A is connected to vertex B and vertex B to vertex C, then likely vertex A will also be connected to vertex C.
- **Scale free effect:** This is based on connectivity distribution $P(k)$.
- **Network resilience and robustness:** *network robustness or resilience* is a measure of the network's response to perturbations or challenges (such as failures or external attacks) imposed on the network. A computable measure for network robustness allows us to (a) compare different networks and (b) improve a network to achieve a desirable level of robustness.

- **Betweenness centrality of vertices: Betweenness** is a centrality measure of a vertex within a graph. Vertices that occur on many shortest paths between other vertices have higher betweenness than those that do not.

For a graph $G = (V, E)$ with n vertices, the betweenness $C_B(v)$ for vertex v is computed as follows:

1. For each pair of vertices (s, t), compute all shortest paths between them.
2. For each pair of vertices (s, t), determine the fraction of shortest paths that passes through the vertex in question (here, vertex v).

3. Sum this fraction over all pairs of vertices (s, t).

This can be express mathematically as (Shivaram, 2005):

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}} \qquad 2.2$$

where σ_{st} is the number of shortest paths from *s* to *t*, and $\sigma_{st}(v)$ is the number of shortest paths from s to t that pass through a vertex v. This may be normalised by dividing through the number of pairs of vertices not including *v*, which is $(n-1)(n-2)/2$ for directed graphs and $(n-1)(n-2)$ for undirected graphs. For example, in an undirected star graph, the center vertex (which is contained in every possible shortest path) would have a betweenness of $(n-1)(n-2)$ while the leaves (which are contained in no shortest paths) would have a betweenness of 0.

- **Closeness centrality of vertices**

In graph theory (http://en.wikipedia.org., 2011), **closeness** is a centrality measure of a vertex within a graph. Vertices that are 'shallow' to other vertices (that is, those that tend to have short geodesic distances to other vertices within the graph) have higher closeness. Closeness is preferred in network analysis to mean shortest-path length, as it gives higher values to more central vertices, and so is usually positively associated with other measures such as degree.

In the network theory, **closeness** is a sophisticated measure of centrality. It is defined as the mean geodesic distance (i.e., the shortest path) between a vertex *v* and all other vertices reachable from it:

$$\frac{\sum_{t \in V \setminus v} d_G(v,t)}{n-1} \qquad 2.3$$

where $n \geq 2$ is the size of the network's 'connectivity component' V reachable from v.

Closeness can be regarded as a measure of how long it will take information to spread from a given vertex to other reachable vertices in the network.

The closeness $C_c(v)$ for a vertex v can also be defined to be the reciprocal of the sum of geodesic distances to all other vertices of V (Sabidussi, 1966).

$$C_c(v) = \frac{1}{\sum_{t \in V \setminus v} d_G(v,t)} \qquad 2.4$$

2.5 Biochemical networks

Biochemical networks are the central processing units of life. They can perform a variety of computational tasks. In a biochemical network, computations are performed by molecules that chemically and physically interact with each other (Othmer, 2006).

The three types of networks treated here are defined as follows.

- **Signal transduction networks:** The pathways and the molecular components, such as kinases, G-proteins, second messengers, etc., involved in transducing a signal from one location to another. It is frequently used in the context of transduction of extracellular into intracellular signals.
- **Metabolic networks:** The pathways and the molecular components (metabolites, enzymes, control factors) involved in the biosynthesis of new components, the conversion of molecular 'foodstuffs' into energy, etc. One of the most important examples is the glycolytic pathway, which converts sugars into energy-storing molecules such as ATP.
- **Protein-Protein Interaction networks:** Protein–Protein Interactions occur when two or more proteins bind together, often to carry out their biological function. Many of the most important molecular processes in the cell such as DNA replication are carried out by large molecular machines that are built from a large number of protein components organised by their protein–protein interactions (Othmer, 2006).

2.5.1 Protein-Protein Interaction networks

Protein–Protein Interactions (PPIs) are when two or more proteins bind together, often to carry out their biological function. Many of the most important molecular processes in the cell such as

DNA replication are carried out by large molecular machines that are built from a large number of protein components organized by their protein-protein interactions. Protein interactions have been studied from the perspectives of biochemistry, quantum chemistry, molecular dynamics, signal transduction and other metabolic or genetic/epigenetic networks. Indeed, protein–protein interactions are at the core of the entire interactomics system of any living cell (http://en.wikipedia.org, 2011).

The interactions between proteins are important for the majority of biological functions. For example, signals from the exterior of a cell are mediated to the inside of that cell by protein–protein interactions of the signaling molecules. This process, called signal transduction, plays a fundamental role in many biological processes and in many diseases (e.g. cancers). Proteins might interact for a long time to form part of a protein complex, a protein may be carrying another protein (for example, from cytoplasm to nucleus or vice versa in the case of the nuclear pore importins), or a protein may interact briefly with another protein just to modify it (for example, a protein kinase will add a phosphate to a target protein). This modification of proteins can itself change protein–protein interactions. For example, some proteins with SH2 domains only bind to other proteins when they are phosphorylated on the amino acid tyrosine while bromodomains specifically recognize acetylated lysines. In conclusion, protein–protein interactions are of central importance for virtually every process in a living cell. Information about these interactions improves our understanding of diseases and can provide the basis for new therapeutic approaches (http://en.wikipedia.org, 2011).

The main structures of PPI are that, it relate network structure to biological function and also have common properties of biological networks.

2.5.2 Signalling transduction networks

Signal transduction has historically been viewed in terms of linear signaling pathways that lead from a specific input to a particular outcome (Othmer, 2006). However, experiments show that while individual cells may receive multiple simultaneous inputs, they are able to rapidly integrate these signals so as to produce the appropriate response; thus the information conveyed by the signal transduction machinery is often distributed among numerous pathways, and the same stimulus can generate different responses depending on the setting. Binding of a ligand to a signaling receptor can initiate transcription of many genes, and the same signaling molecule may trigger very different responses, depending on the cell type, its internal state, and the state of its local environment (e.g. neighboring cells). Therefore, a signal transduction pathway is the chain of processes by which a cell converts an extra cellular signal into a response and the molecular components, such as kinases, G-proteins, second messengers, etc are key molecules involved in transducing a signal from one location to another. It is frequently used in the context of transduction of extracellular into intracellular signals (Othmer, 2006).

Since most organisms maintain a clear distinction between inside and outside, many primary environmental signals do not penetrate very far into the organism. Instead there are mechanisms for transducing an external signal into an internal signal, and where appropriate, an internal response. For example, at the cellular level extracellular hydrophilic first messenger signals elicit a response through a transduction system in the cell membrane that translates the signal into an intracellular second messenger signal. Similarly, in the sensory systems of higher organisms, light or mechanical stimuli are transduced by a multi-step cascade into an electrical signal that is processed at a higher level.

Therefore, when we speak of signal transduction we invariably refer to the molecular network involved in transducing extracellular signals into intracellular signals. Lipid-soluble molecules can pass through the cell membrane, but most signals are proteins or peptides and these require specialized machinery. Figure 2.6(a) shows a canonical type of signal transduction system, while Figure 2.6(b) shows an abstract version of the steps involved.

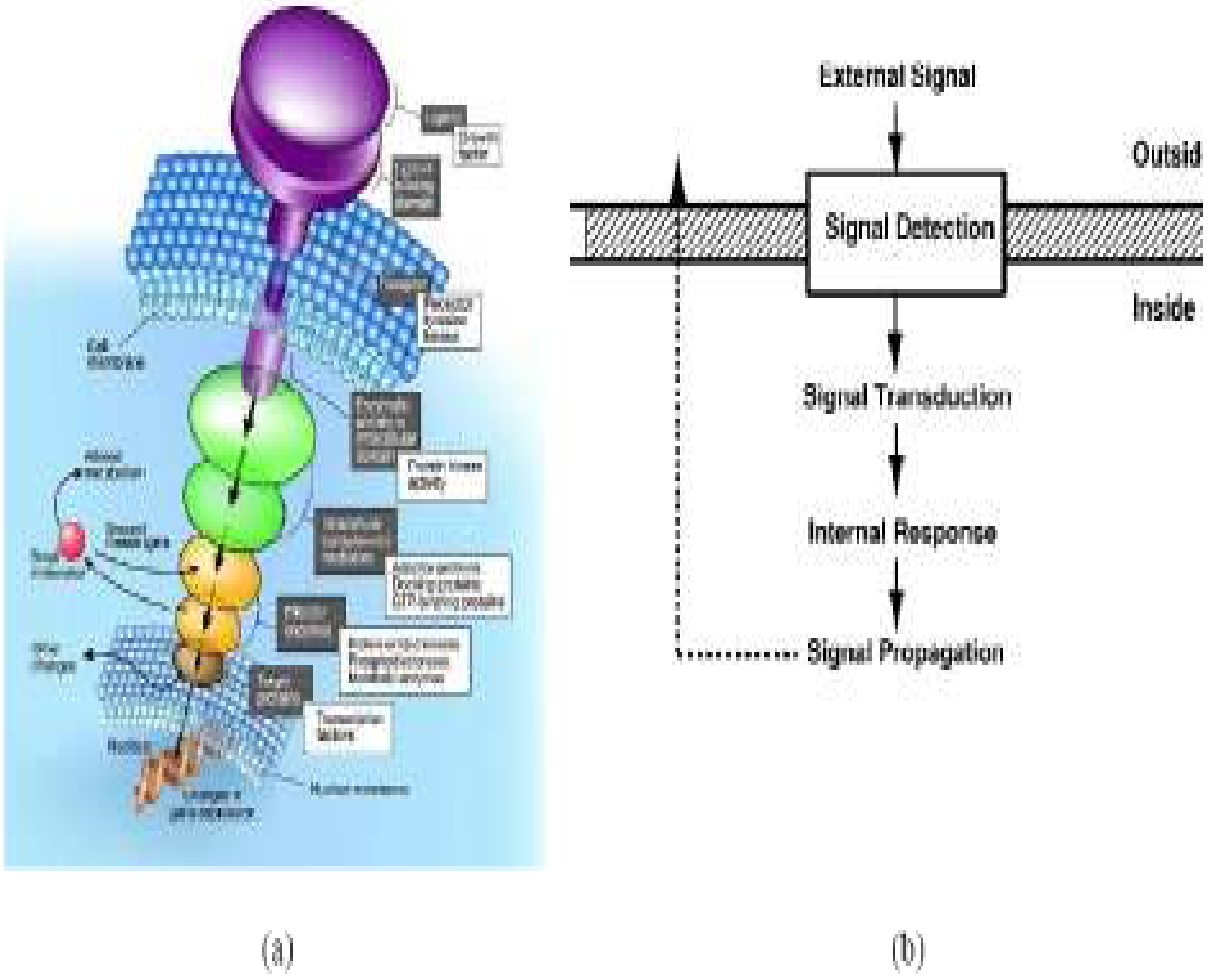


Fig. 2.6: (a) A signal transduction pathway and (b) an abstract rendition of it(Othmer, 2006)

2.5.3 Metabolic networks

Metabolism is the cellular process by which organic molecules are synthesized or degraded, usually via enzyme-catalyzed reactions, and the interconnected components and reactions form a network, called the metabolic network (Othmer, 2006). A highly schematized form that shows the major subdivisions of glycolysis, the Kreb's or citric acid cycle, and electron transport, is given in Figure 2.7. An intermediate level of detail is shown in Figure 2.8. Metabolic reactions are characterized as either catabolic or anabolic, depending on their function. Catabolic reactions are used for breakdown foodstuffs for the production of energy in the form of ATP, production of reducing power in the form of NADPH, and regeneration of small molecules for anabolism. Anabolic reactions typically involve production of small molecules and building blocks that are not sufficiently available in the food, and synthesis of macromolecules such as proteins and nucleic acids. All cells process glucose initially by glycolysis, in which each molecule of glucose is broken down into two molecules of pyruvate and yields a net of two ATP molecules. Glycolysis by itself is anaerobic, i. e., it doesn't required oxygen. Cells that are capable of aerobic metabolism pass the product of glycolysis into the aerobic pathway. Glycolysis takes place in the cytoplasm of cells.

There are two ways in which one can think of the set of reactions in Figure 2.8, firstly as a set of individual enzyme-catalyzed steps, as in Figure 2.9(a), or as a network of connected steps, as in Figure 2.9(b). Experimentalist who studies individual steps would adopt the viewpoint in (a), but to understand how the connected set of reactions functions, it is obviously necessary to adopt the viewpoint in (b). In the approach developed later, the individual molecules or combinations of

them will be represented by the nodes in a graph, and the reactions between them will be represented by the edges in that graph. This then suggests that the topology of the graph by itself may play a role, and also leads to an easier understanding of how changes in various parts of the pathway affect the fluxes between a chosen pair of nodes.

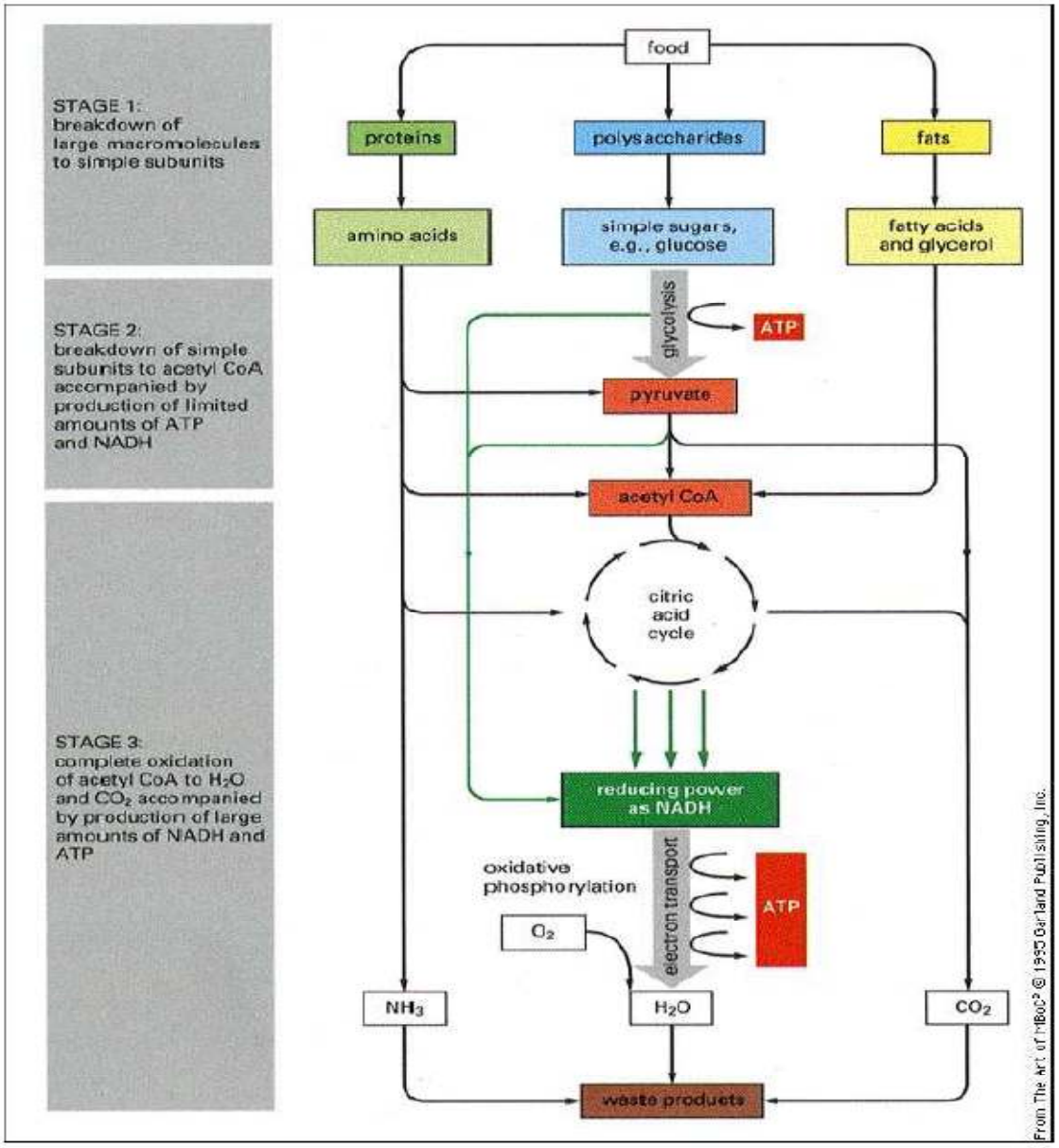


Figure 2.7. A schematic of the metabolic pathways for cellular respiration (Albert *et al.*, 1994).

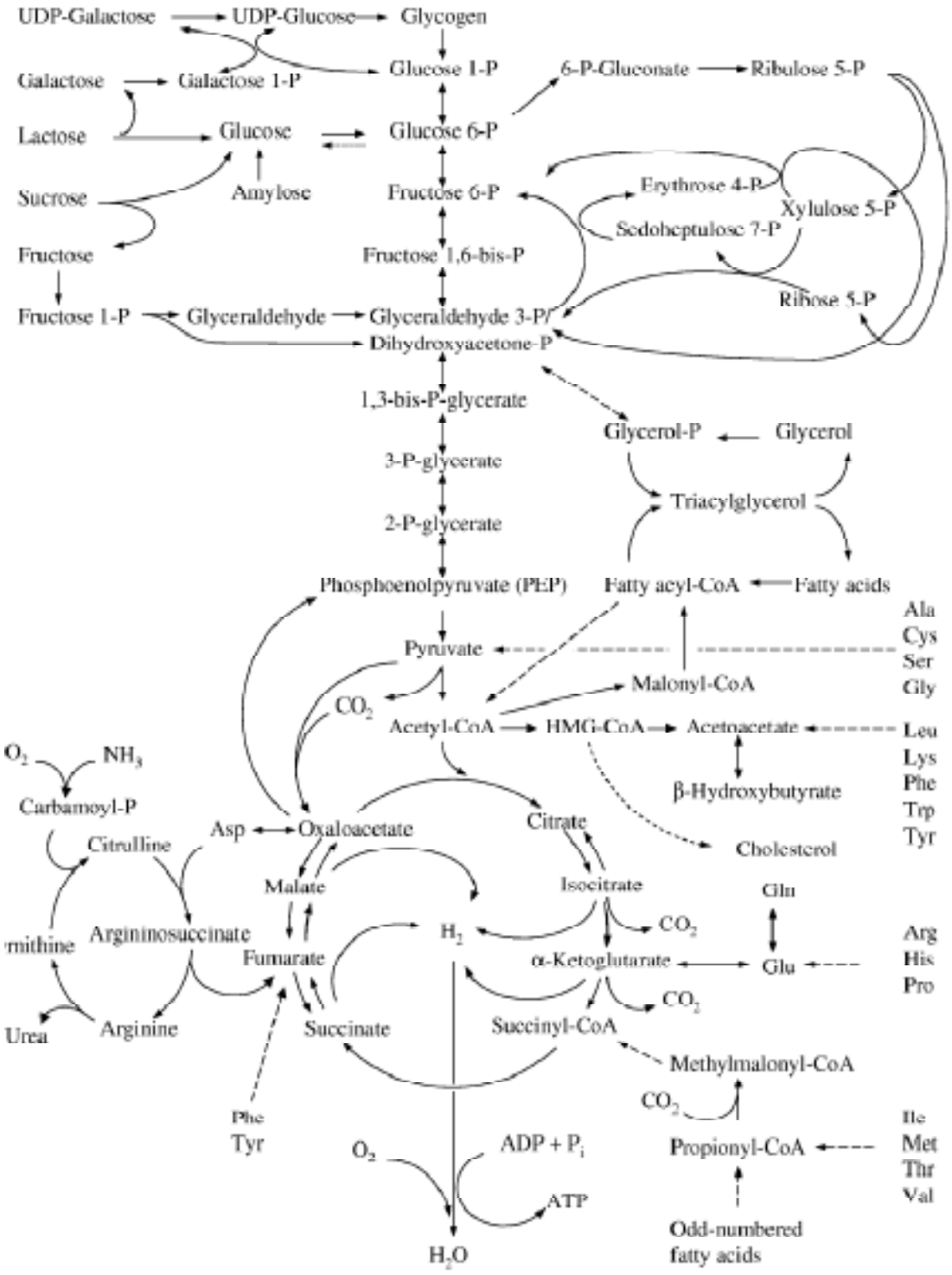


Figure 2.8: A stripped-down version of the glycolytic pathway and the Kreb's cycle (Palmer, 2006).

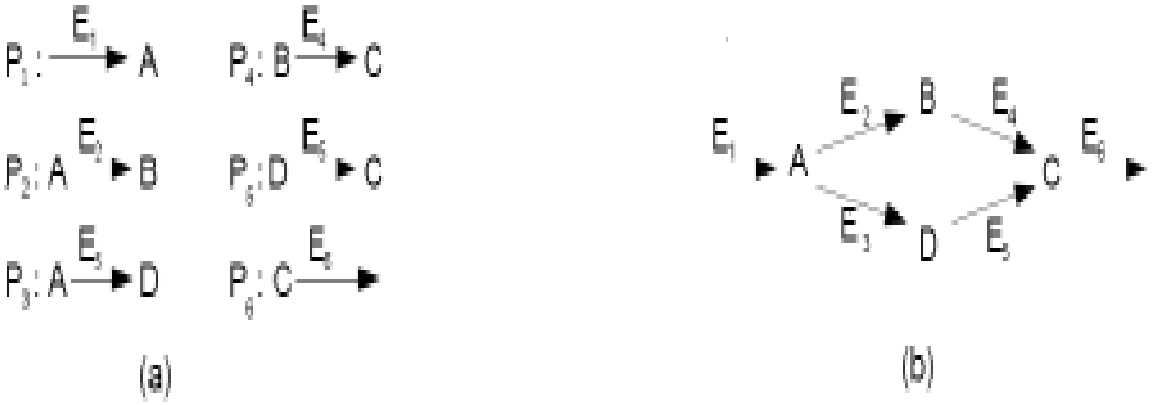


Figure 2.9: (a) A set of reactions viewed as individual steps, or (b) as a connected set (Othmer, 2006).

2.5.4 Transcription Regulatory Networks

Recent progress in molecular biology has led to a complete map of the genome of a number of organisms, and the challenge now is to discover the regulatory networks that govern the interactions between genes, the messenger RNAs and proteins they encode, and other cellular components (Othmer G., 2006) (Figure 2.10(a)). Signal transduction and gene regulatory networks control transduction of extracellular signals into patterns of gene expression via a few common modes

(Figure 2.10(b)), and understanding how these networks integrate different signals in the presence of fluctuations in the amounts of signaling molecules is a major problem that will undoubtedly require new mathematical approaches for its resolution. The crucial role of inter-regulation amongst genes is especially evident during the development of a multicellular adult from a unicellular egg, for what a cell becomes depends on where it is in a developing aggregate of cells. Pattern formation in development refers to the spatially- and temporally-organized expression of genes in a multicellular array, and this is controlled by the inputs and outputs of the gene control networks.

Gene regulatory networks can be viewed as directed graphs, in which nodes represent transcription factors, and in which edges represent the regulatory interactions.

2.5.5 Network Utilization

Despite their impressive successes, purely topologic approaches have important intrinsic limitations. For example, the activity of the various metabolic reactions or regulatory interactions differs widely, some being highly active under most growth conditions while others are switched

on only for some rare environmental circumstances. Therefore, an ultimate description of cellular networks requires us to consider the intensity (i.e., strength), the direction (when applicable) and the temporal aspects of the interactions. While we know little about the temporal aspects of the various cellular interactions, recent results have shed light on how the strength of the interactions is organized in metabolic and genetic regulatory networks (Almaas, *et al.*, 2004; Kutznetsov *et al.*, 2002; and Farkas, *et al.*, 2002) and how the local network structure is correlated with these link strengths.

2.5.6 Flux Utilization

In metabolic networks, the flux of a given metabolic reaction, representing the amount of substrate being converted to a product within unit time, offers the best measure of interaction strength. Recent advances in metabolic flux-balance approaches (FBA) (Edwards J. and Palson, 2000; Edwards, *et al.*, 2001; Ibara *et al.*, 2002; Segre, *et al.*, 2002; and Emmerling *et al.*, 2002) allow us to calculate the flux for each reaction, and they have significantly improved the ability to generate quantitative predictions on the relative importance of the various reactions, thus leading to experimentally testable hypotheses. The FBA approaches can be described as follows: Starting from a stoichiometric matrix model of an organism, for example. one for *E. coli* contains 537 metabolites and 739 reactions (Edwards and Palson, 2000; Edwards, *et al.*, 2001; and Ibara *et al.*, 2002), the steady state concentrations of all metabolites must satisfy

$$\frac{d}{dt}(A_i) = \sum_j S_{ij} V_j = 0 \qquad 2.5$$

where S_{ij} is the stoichiometric coefficient of metabolite A_i in reaction j and V_j is the flux of reaction j. We use the convention that if metabolite A_i is a substrate (product) in reaction j, $S_{ij} < 0$

($S_{ij} > 0$), and we constrain all fluxes to be positive by dividing each reversible reaction into two "forward" reactions with positive fluxes. Any vector of positive fluxes $\{V_j\}$ which satisfies Eq. (2.1) corresponds to a state of the metabolic network, and hence, a potential state of operation of the cell. Assuming that the cellular metabolism is in a steady state and optimized for the maximal growth rate (Edwards, *et al.*, 2001; and Ibara *et al.*, 2002), FBA allows us to calculate the flux for each reaction using linear optimization, providing a measure of each reaction's relative activity (Almaas *et al.*, 2004). A striking feature of the resulting flux distribution from such modeling of both *H. pylori*, *E. coli* and *S. cerevisiae* is its overall in homogeneity: reactions with fluxes spanning several orders of magnitude coexist under the same conditions (Fig. 5a). This is captured by the flux distribution for *E. coli,* which follows a power law where the probability that a reaction has flux V is given by $P(V) \approx (V + V_0)^{-\alpha}$. This flux exponent is predicted to be α *=1.5* by FBA methods (Almaas *et al.*, 2004). In a recent experiment (Emmerling *et al.*, 2002) the strength of the various fluxes of the *E. coli* central metabolism was measured, revealing (Almaas *et al.*, 2004) the power-law flux dependence $P(V) \approx V^{-\alpha}$ with $\alpha \cong 1$. This power law behavior indicates that the vast majority of reactions have quite small fluxes, while coexisting with a few reactions with extremely large flux values.

The observed flux distribution is compatible with two quite different potential *local* flux structures (Almaas *et al.*, 2004). A homogeneous local organization would imply that all reactions producing (consuming) a given metabolite has comparable fluxes. On the other hand, a more delocalized "hot backbone" is expected if the local flux organization is heterogeneous, such that each metabolite has a dominant source (consuming) reaction.

To distinguish between these two scenarios for each metabolite *I* produced (consumed) by *k* reactions. A measure $Y(k,i)$ (Barthelemy, *et al.*, 2003; and Derrida and Flyvbjerg, 1987) has been defined as

$$Y(k,i) = \sum_{j=1}^{k} \left(\frac{\bar{v}_{ij}}{\sum_{l=1}^{k} \bar{v}_{il}}\right)^2 \qquad 2.6$$

where $\bar{v}_{ij}$ is the mass carried by reaction *j* which produces (consumes) metabolite *i*. If all reactions producing (consuming) metabolite *i* have comparable $\bar{v}_{ij}$ values, $Y(k,i)$ scales as $1/k$. If, however, a single reaction's activity dominates Eq. (2.2), we expect $Y(k,i) \sim 1$, i.e., $Y(k,i)$ is independent of *k*. For the *E. coli* metabolism optimized for succinate and glucose uptake we find that both the *in* and *out* degrees follow the power law $Y(k,i) \sim k^{0.27}$, representing an intermediate behavior between the two extreme cases (Almaas *et al.*, 2004). This suggests that the large-scale inhomogeneity observed in the overall flux distribution is increasingly valid at the level of the individual metabolites as well: for most metabolites, a single reaction carries the majority of the flux. Hence, the majority of the metabolic flux is carried along linear pathways – the metabolic high flux backbone (HFB) (Almaas *et al.*, 2004).

2.5.7. Gene Interactions

One can also investigate the strength of the various genetic regulatory interactions provided by microarray datasets. Assigning each pair of genes a correlation coefficient which captures the degree to which they are co-expressed, one finds that the distribution of these pair-wise correlation coefficients follows a power law (Kutznetsov *et al.*, 2002; and Farkas *et al.*, 2003). That is, while the majority of gene pairs have only weak correlations, a few gene pairs display a significant correlation coefficient. These highly correlated pairs likely correspond to direct regulatory and

protein interactions. This hypothesis is supported by the finding that the correlations are larger along the links of the protein interaction network and between proteins occurring in the same complex than for pairs of proteins that are not known to interact directly (Dezso, *et al*., 2003; Grogoriev, 2001; Jansen *et al.*, 2002; and Ge *et al.*, 2001).

Taken together, these results indicate that the biochemical activity in both the metabolic and genetic networks is dominated by several 'hot links' that represent a few high activity interactions embedded into a web of less active interactions. This attribute does not seem to be a unique feature of biological systems: hot links appear in a wide range of non-biological networks where the activity of the links follows a wide distribution (Goh *et al.*, 2002; deMenezes and Barabasi, 2004). The origin of this seemingly universal property is, again, likely rooted in the network topology. Indeed, it seems that the metabolic fluxes and the weights of the links in some non-biological system (Goh *et al.*, 2002; deMenezes and Barabasi, 2004) are uniquely determined by the scale-free nature of the network. A more general principle that could explain the correlation distribution data as well is currently lacking.

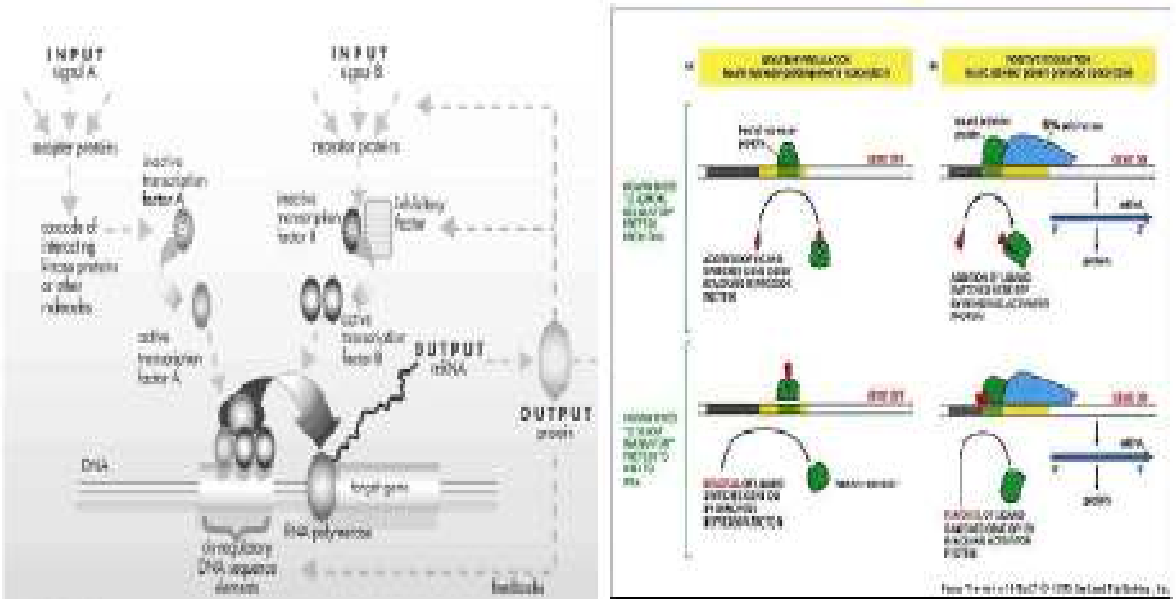


Figure 2.10: (a) A schematic of gene regulation showing the steps in transducing an extracellular signal to a change in gene expression. (b) The common modes of gene regulation (Othmer, 2006).

2.6. Graph-based methods for interaction network in system biology

Protein-Protein Interactions (PPIs) networks are modelled by undirected graphs, where the nodes are proteins and two nodes are connected by an undirected edge if the corresponding proteins physically bind. While the transcriptional regulatory networks can be modelled as directed weighted graphs, where the weights of directed edges capture the degree of the regulatory effect of the transcription factors (i.e. source nodes) to their regulated genes (sink nodes). Metabolic networks generally require more complex representations, such as hypergraphs, as reactions in metabolic networks generally convert multiple reaction inputs into multiple outputs with the help of other components. An alternative, reduced representation for a metabolic network, is a weighted bipartite graph, where two types of nodes are used to represent reactions and compounds, respectively, and the edges connect nodes of different types, representing either substrate or product relationships (Aittokallio and Benno, 2006).

The representation of complex cellular networks as graphs has made it possible to systematically investigate the topology and function of these networks using well-understood graph-theoretical concepts that can be used to predict the structural and dynamical properties of the underlying network. Such prediction can suggest new biological hypotheses regarding the unexplored new interactions of the global network or the function of individual cellular components that are testable subsequent experimentation. Mathematical modelling also enables an iterative process of network reconstruction, where model simulations and predictions are closely coupled with new experiments chosen systematically to maximize their information content for subsequent model adjustments, providing increasingly more accurate descriptions of the network properties (Papin *et al*., 2005). The topological relations underlying graph-based methods can also convey structure of putative pathways. This helps avoiding approaches that test many known sets of molecules without

causal interactions (Curtisa *et al.*, 2005). Furthermore, graph formalisms may provide powerful tools for omics data integration to address fundamental biological questions at the systems level (Joyce and Palsson, 2006).

A substantial effort has been devoted to develop graph-based methods for a wide range of computational and biological tasks. The selected methods are presented in the broader context of network analysis, summarizing some of the basic concepts and themes such as scale-free networks, pathways and modules as depicted in table 2.1. The order of sections roughly reflects the increasing demands placed for the type and amount of data the methods require and their applicability to address more targeted problems in cell biology.

2.6.1 The characterization of network topology

The most general level of network analysis comes from global network measures that allow us to characterize and compare the given network topologies (i.e. the configuration of the nodes and their connecting edges). Global measures such as the degree distribution (i.e. the degree of a node is the number of edges it participate in) and the clustering coefficient (the number of edges connecting the neighbours of the node divided by the maximum number of such edges) have recently reviewed in the context of cellular networks (Barabasi and Olttvai, 2004) and in proteomics (Grindrod and Kibble, 2004). Several types of surveyed biological networks, such as PPI, gene regulation and metabolic networks are thought to display scale-free topologies (i.e. most nodes have only a few connections whereas some nodes are highly connected), characterized by a power-law degree distribution that decays slower than exponential. This type of network topology is frequently observed in numerous non-biological networks and it can be generated by simple and

elegant evolutionary models, where new nodes attach preferentially to sites that are already highly connected. Numerous improvements to this generic model include, for instance, iterative network duplication and integration to its original core, leading to hierarchical network topologies, which are characterized by non-constant clustering coefficient distribution (Barabasi and Olttvai, 2004; and Albert, 2005).

It is however observed that, in practice, the architecture of large-scale biological networks is determined with sampling methods, resulting in subnets of the true network, and only these partial networks can be applied to characterize the topology of the underlying, hidden network (Lappe, 2004). It has recently been recognized that it is possible to extrapolate from subnets to the properties of the whole network only if the degree distributions of the whole network and randomly sampled subnets share the same family of probability distributions (Stumpf *et al.*, 2005). While this is the case in specific classes of network graph models, including classical Erdös-Rényi and exponential random graphs, the condition is not satisfied for scale-free degree distributions. The recent studies in interactome networks have revealed that the commonly accepted scale-free model for PPI networks may fail to fit the data (Przulj *et al.*, 2004). Moreover, limited sampling alone may as well give rise to apparent scale-free topologies, irrespective of the original network topology (Han *et al.*, 2005). These results suggest that interpretation of the global properties of the complete network structure based on the current-still-limited-accuracy and coverage of the observed networks should be made with caution. Moreover, while the scale-free and hierarchical graph properties can efficiently characterize some large-scale attributes of networks, the local modularity and network clustering is likely to be the key concepts in understanding most cellular mechanisms and functions.

2.6.2 The Graph Analysis of Interaction Patterns

As an alternative to the study of global graph characteristics, elementary graph algorithms have been used to characterize local interconnectivity and more detailed relationships between nodes. Such graph methods can facilitate addressing fundamental biological concepts, such as essentiality and pathways, especially when additional biological information is incorporated into the analysis in addition to the primary data. For example, while gene expression clustering traditionally makes the assumption that genes with similar expression profiles have similar functions in cells, a more targeted approach could aim at identifying the genes participating in a particular cellular pathway where not every components has a similar transcriptional profile (Zhou *et al*., 2002). Once the network of interest had been represented as a graph, the conventional graph-driven analysis work-flow involves the following two steps:-

- Applying suitable graph algorithms to compute the local graph properties, such as the number and complexity of given subgraph, the shortest path length of indirectly connected nodes or the presence of central nodes of the network and

- Evaluating the sensitivity and specificity of the model predictions using curated databases of known positive examples or random models of synthetic negative examples, respectively.

2.6.3 The Subgraphs and Centrality Statistics

A subgraph represents a subset of nodes with a specific set of edges connecting them. As the number of distinct subgraphs grows exponentially with the number of nodes, efficient and scalable

heuristics have been developed and applied for detecting the given subgraphs and their frequencies in large networks. Under the random graph model, it is also possible to calculate analytically the estimated distribution of different subgraphs with given number of nodes, edges and their specific global properties like degree distribution and clustering coefficient (Vazques *et al.*, 2004).

Centrality is a local quantitative measure of the position of a node relative to the other nodes, and can be used to estimate its relative importance or role in global network organization. Different flavours of centrality are based on the node's connectivity (degree centrality), its shortest paths to other nodes (closeness centrality) or the number of shortest paths going through the node (betweenness centrality). Estrada (Estrada, 2006) recently showed that centrality measures based on graph spectral properties can distinguish essential proteins in PPI network of yeast *Sacharonyces cereviae* (essential genes are those upon which the cell depends for viability). In particular, the best performance in identifying essential proteins was obtained with a novel measure introduced to account for the participation of a given node in all subgraphs of the network (subgraph centrality), which gives more weight to smaller subgraphs. It was proposed that ranking proteins according to their centrality measures could offer a means to selecting possible targets for drug discovery (Estrada, 2006). A similar approach to characterize the importance of individual nodes, based on trees of shortest paths and concepts of bottleneck nodes, demonstrated that 70% of the top 10 most frequent 'bottleneck' proteins were unviable and structural proteins that do not participate in cellular signalling (Przulj *et al.*, 2004). With degree centrality analysis in the metabolic networks of *Escherichia coli*, *S. cerevisae* and *Staphylococcus aureus*, it was demonstrated that most reactions identified as essential turned out to be those involving the production or consumption of low-degree metabolites (Samal *et al.*, 2006).

Table 2.1: Examples of some Graph-based approaches to cellular network analysis(Tero and Benno, 2006)

Network topology	**Interaction patterns**	**Network decomposition**
Global structural properties	Local structural connectivity	Hierarchical functional organization
Scale-free topology	Subgraphs	Modules
Degree distribution	Centrality	Motifs
Clustering coefficient	Pathways	Clusters

2.6.4 Paths and Pathways

In the theory of directed graphs, a path is a chain of distinct nodes, connected by directed edges, without branches or cycles. Such pathways in cellular network graphs can represent, for instance, a transformation path from a nutrient to an end product in a metabolic network, or a chain of post-translational modifications from the sensing of a signal to its intended target in a signal transduction network (Albert, 2005). Pathways redundancy (the presence of multiple paths between the same pair of nodes) is an important local property that is thought to be one of the reasons for the robustness of many cellular networks. Betweenness centrality can be used to measure the effect of node perturbations on pathway redundancy, whereas path lengths characterize the response times under perturbations. With shortest paths and centrality-based predictions in the *S. cerevisiae* PPI and metabolic networks, respectively, the existence of alternate paths that bypass viable proteins can be demonstrated, whereas lethality corresponds to the lack of alternative pathways in the perturbed network (Przulj *et al.*, 2004; and Palumbo *et al.*, 2005). Besides the various commercial software packages for pathway analysis there exist also freely available tools for some specific graph queries, such as finding shortest paths between two specified seed nodes on degree weighted metabolic networks (Croes *et al.*, 2005) or searching for linear paths that are similar to query pathways in terms of their composition and interaction patterns on a given PPI network (Shlomi *et al.*, 2006).

A **linear pathway** has a well-defined source, a chain of intermediary nodes, and a sink (end) node. The clustering coefficient of each node is zero, because there are no edges among first neighbors. Both the maximum and average path length increase linearly with the number of nodes and are

long for pathways that have many nodes. This type of graph has been widely used as a model of an isolated signal transduction pathway (Albert, 2005).

The relatively high degree of noise inherent in the interactions data in current PPI databases can make pathway modelling very challenging. Integration of prior biological knowledge, such as Gene Ontology (GO), can be used to make the process of inferring models more robust by providing complementary information on protein function. GO terms and their relationships are encoded in the form of directed acyclic graph (DAG). Guo *et al*. (Guo *et al*., 2006) recently assessed the capability of both GO graph structure-based and information content-based similarity measures on DAG to evaluate the PPIs involved in human regulatory pathways. They also showed how the functional similarity of proteins within known pathways decays rapidly as their path length increases. While most of the analysis methods designed for PPI networks consider unweighted graphs, where each pairwise interaction is considered equally important, Scott J. *et al*. (Scott *et al.*, 2006) recently presented linear-time algorithms for finding paths and more general graph structures such as trees that can also consider different reliability scores for PPIs. By exploiting a powerful randomized graph-algorithm called color coding, they efficiently recovered several known *S. cerevisiae* signalling pathways such as MAPK, and showed that in general the pathways they detected score higher than those found in randomized networks. In addition to known pathways, they also predicted novel putative pathways in the PPI network that are functionally enriched (i.e. share significant number of common GO annotations) (Scott *et al.*, 2006).

2.6.5 Network decomposition into functional modules

The decomposition of large networks into distinct components, or modules, has come to be regarded as a major approach to deal with the complexity of large cellular networks (Hartwell *et al.*, 1999; Lee *et al.*, 2002; and Milo *et al.*, 2002). In cellular networks, a module refers to a group of physically or functionally connected biomolecules (nodes in graphs) that work together to achieve the desired cellular function (Barabasil and Oltvai, 2004). To investigate the modularity of interaction networks, tools and measures have been developed that can not only identify whether a given network is modular or not, but also detect the modules and their relationships in the network. By subsequently contrasting the found interaction patterns with other large-scale functional genomics data, it is possible to generate concrete hypothesis for the underlying mechanisms governing e.g. the signalling and regulatory pathways in a systematic and integrative fashion. For instance, interaction data together with mRNA expression data can be used to identify active subgraphs, that is, connected regions of the network that show significant changes in expression over particular subnets of experimental conditions (Ideke *et al.*, 2002).

2.6.6 Clustering Coefficient

A measure that gives insight into the local structure of a network is the so-called clustering of a node: the degree to which the neighborhood of a node resembles a complete subgraph.

For a node *i* with degree k_i the clustering is defined as

$$C_i = \frac{2n_i}{k_i(k_i - 1)} \qquad 2.7$$

representing the ratio of the number of actual connections between the neighbors of node *i* to the number of possible connections. For a node which is part of a fully interlinked cluster $C_i = 1$,

while C_i = 0 for a node where none of its neighbors are interconnected. Accordingly, the overall clustering coefficient of a network with *N* nodes is given by $\langle C \rangle = \sum \frac{C_i}{N}$ quantifying a network's potential modularity. By studying the average clustering of nodes with a given degree *k*, information about the actual modular organization of a network can be extracted (Ravasz *et al.*, 2002; Ravasz and Barabasi, 2003; Dorogovtsev *et al.*, 2002; and Vazquez *et al.*, 2002). For all metabolic networks available, the average clustering follows a power-law form as $C(k) \sim k^{-\alpha}$ [45] suggesting the existence of a hierarchy of nodes with different degrees of modularity (as measured by the clustering coefficient) overlapping in an iterative manner (Ravasz *et al.*, 2002). In summary, we have seen strong evidence that biological networks are both scale-free (Jeong *et al.*, 2000; and Jeong *et al.*, 2001) and hierarchical (Ravasz *et al.*, 2002).

2.7 Summary

The large-scale data on biomolecular interactions that is becoming available at an increasing rate enables a glimpse into complex cellular networks. Mathematical graphs are a straightforward way to represent this information, and graph-based models can exploit global and local characteristics of these networks relevant to cell biology. Most current research activities concern the dissection of networks into functional modules, a principal approach attempting to bridge the gap between our very detailed understanding of network components in isolation and the 'emergent' behaviour of the network as a whole, which is frequently the phenotype of interest on a cellular level. Approaches developed for DNA and protein sequence analysis, such as multiple alignment and statistical over-representation of parts, are being carried over to address these problems. Network graphs have the advantage that they are very simple to reason about, and correspond by and large to the information that is globally available today on the network level. However, while binary

relation information does represent a critical aspect of interaction networks, many biological processes appear to require more detailed models. Therefore, we expect that one of the main directions in the development of graph-based methods will be their extension to other types of large-scale data from existing and new experimental technologies. This may eventually prove mathematical models of large-scale data sets valuable in medical problems.

The power laws in system biology are abundant in nature, affecting both the construction and the utilization of real networks. The power-law degree distribution has become the trademark of scale-free networks and can be explained by invoking the principles of network growth and preferential attachment. However, many biological networks are inherently modular, a fact which at first seems to be at odds with the properties of scale-free networks. However, these two concepts can co-exist in hierarchical scale-free networks. In the utilization of complex networks, most links represent disparate connection strengths or transportation thresholds. For the metabolic network of *E. coli*, we can implement a flux-balance approach and calculate the distribution of link weights (fluxes), which (reflecting the scale-free network topology) displays a robust power-law, independent of exocellular perturbations. Furthermore, this global in homogeneity in the link strengths is also present at the local level, resulting in a connected "hot-spot" backbone of the metabolism. Similar features are also observed in the strength of various genetic regulatory interactions. Despite the significant advances witnessed the last few years, network biology is still in its infancy, with future advances most notably expected from the development of theoretical tools, development of new interactive databases and increased insights into the interplay between biological function and topology.

CHAPTER THREE

RESEARCH METHODOLOGY

3.1 Constructing Protein-Protein Interaction Network for Signalling Pathways Extraction

Protein-Protein Interaction data was obtained from the work of LaCount *et al*. (LaCount *et al*., 2005). Their results comprise 2846 interactions between 1309 proteins. We model all protein-protein interaction data of a gene using an interaction graph, where vertices are the gene's interacting proteins, and whose edges represent pairwise interactions between distinct proteins.

3.1.1 Estimation of interaction probabilities

To add weight to the edges, several authors have suggested methods for evaluating the reliabilities of protein interactions (Deng *et al.*, 2003; Bader, *et al.*, 2004; and Von *et al.*, 2002). In this work, a method developed by Bader *et al.* (Bader, *et al.*, 2004) was used and assigned confidence values to protein interaction networks using a logistic regression model that consists of three parameters (X_1, X_2, X_3):.

- The number of times an interaction between the proteins was experimentally observed;
- The Pearson correlation coefficient of expression measurements for the corresponding genes.
- The proteins' small world clustering coefficient (Goldberg and Roth, 2003) which is defined as the hypergeometric function for the overlap in the neighborhoods of two proteins.

We describe these parameters in detailed as follows:

(a) The number of experimental observations was shown by several authors (Deng *et al.*, 2003) to be predictive on the reliability of an interaction. For *Plasmodium falciparum*, we used the experimental study of LaCount *et al.* (LaCount *et al*., 2005). Here, we defined the number of observations as the number of times the interaction was observed in the corresponding study.

(b) Suppose x and y are two level of expression profile for two genes. The Pearson correlation coefficient between the two genes is defined as:

$$\rho = \frac{\frac{1}{m}\sum_{i=1}^{m} x_i y_i - \bar{x}\bar{y}}{\sigma_x \sigma_y}, \qquad 3.1$$

where $\bar{x}\ and\ \bar{y}$ are the sample means and σ_x and σ_y are the standard deviations of x and y respectively. The correlation coefficient quantifies the similarity of expression between two genes and was shown to be correlated to whether the corresponding proteins interact or not (Grigoriev, 2001; and Ge *et al.*, 2001). To estimate this parameter for the edges, we used the following Microarray Datasets; Bozdeck-3D7 data (Bozdech *et al*., 2003a) over 54 conditions; Bozdeck-HB3 data (Bozdech *et al.*, 2003a) over 49 conditions and LeRoch data (LeRoch, 2003) over 16 conditions.

(c) For proteins v and w, we denote the sets of proteins that interact with them by $N(v)$ and $N(w)$, respectively. Let N be the total number of proteins in the network. The small-world clustering coefficient for v and w is given as:

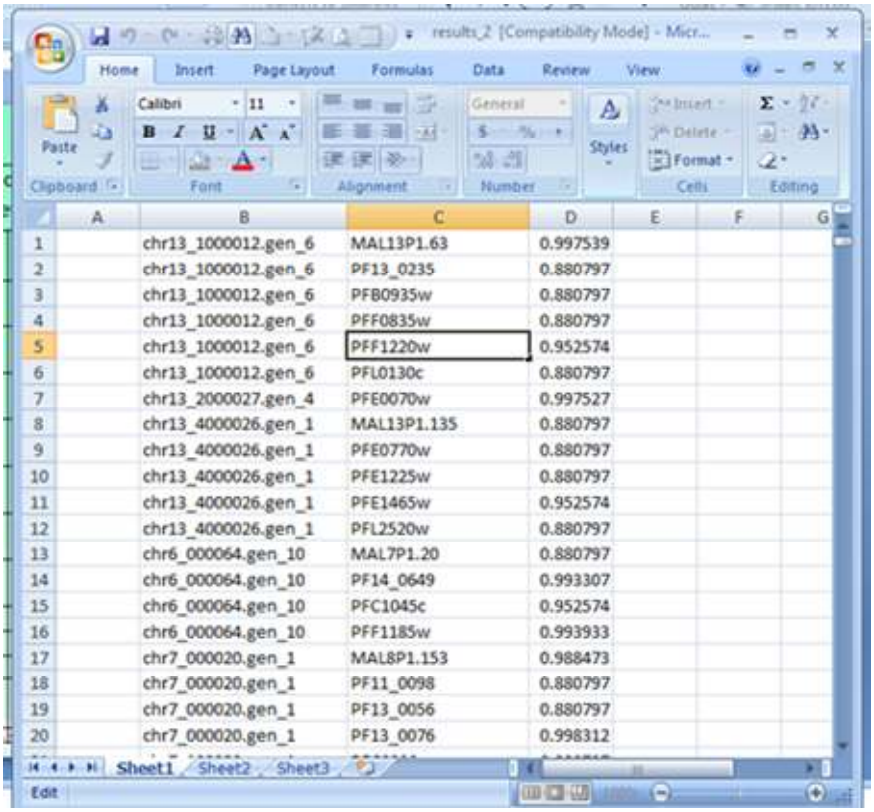
$$C_{vw} = -log \sum_{i=|N(v)\cap N(w)|}^{\min(|N(v)|,|N(w)|)} \frac{\binom{|N(v)|}{i}\binom{N-|N(v)|}{|N(w)|-i}}{\binom{N}{|N(w)|}} \qquad 3.2$$

The clustering coefficient was suggested by Goldberg and Roth (Goldberg and Roth, 2003) to account for similarity in network connections.

According to the logistic distribution, the probability of a true interaction T_{uv} given the three input variables, $X = (X_1, X_2, X_3)$, is:

$$\Pr(T_{uv}|X) = \frac{1}{1 + \exp\left(-\beta_0 - \sum_{i=1}^{3} \beta_i X_i\right)}, \qquad 3.3$$

where $\beta_0 \ldots \beta_3$ are the parameters of the distribution. The snapshot of the resulted edge weights generated from our protein-protein interaction network are shown in figure 3.1.



	A	B	C	D	E	F	G
1		chr13_1000012.gen_6	MAL13P1.63	0.997539			
2		chr13_1000012.gen_6	PF13_0235	0.880797			
3		chr13_1000012.gen_6	PFB0935w	0.880797			
4		chr13_1000012.gen_6	PFF0835w	0.880797			
5		chr13_1000012.gen_6	PFF1220w	0.952574			
6		chr13_1000012.gen_6	PFL0130c	0.880797			
7		chr13_2000027.gen_4	PFE0070w	0.997527			
8		chr13_4000026.gen_1	MAL13P1.135	0.880797			
9		chr13_4000026.gen_1	PFE0770w	0.880797			
10		chr13_4000026.gen_1	PFE1225w	0.880797			
11		chr13_4000026.gen_1	PFE1465w	0.952574			
12		chr13_4000026.gen_1	PFL2520w	0.880797			
13		chr6_000064.gen_10	MAL7P1.20	0.880797			
14		chr6_000064.gen_10	PF14_0649	0.993307			
15		chr6_000064.gen_10	PFC1045c	0.952574			
16		chr6_000064.gen_10	PFF1185w	0.993933			
17		chr7_000020.gen_1	MAL8P1.153	0.988473			
18		chr7_000020.gen_1	PF11_0098	0.880797			
19		chr7_000020.gen_1	PF13_0056	0.880797			
20		chr7_000020.gen_1	PF13_0076	0.998312			

Fig. 3.1: Resulted edges weight generated on *Plasmodium falciparum* PPI network constructed from logistic distribution. For example, in the PPI network above, the interaction of the gene chr13_1000012.gene_6 in column B with the following genes MA13P1.63, PF13_0235, PFB0935w, PFF1220w, PFL0130c and PFE0070w in column C have their corresponding weights(w) 0.997539, 0.880797, 0.880797, 0.880797, 0.952574 and 0.880797 in column D respectfully.

3.1.2 Weighted Graph Representation of the Protein-Protein Interaction Network

The protein–protein interaction network of *Plasmodium falciparum* is represented as a weighted graph $G = (V, E; w)$. The vertices of the graph are the set of unique proteins, and therefore$|V| = 1{,}309$. The edges of the graph are the interactions, and therefore$|E| = 2{,}846$, while w is the weight along the edge between vertices.

Given an undirected weighted graph $G = (V, E, w)$ with n vertices, m edges and a set I of start vertices, we wish to find, for each vertex v, a minimum-weight simple path of length k that starts within I and ends at v. If no such simple path exists, this should be reported.

Signalling networks are modelled as directed weighted graphs, where the weights of directed edges capture the degree of the regulatory effect of the transcription factors (source nodes) to their regulated genes (sink nodes).

In general, this problem is NP-hard, as the travelling-salesman problem is reducible to it. A standard dynamic programming algorithm exists for this problem, which runs in $O(kn^k)$ and requires also $O(kn^k)$ memory (Scott *et al.*, 2006). In an attempt to reduce the time and space complexity, the color coding idea was introduced by Scott *et al.* (Scott *et al.*, 2006). "The idea of color coding is to assign each vertex a random color between $\mathbf{1}$ and $\boldsymbol{k}$ and, instead of searching for paths with distinct vertices, search for paths with distinct colors". The introduction of this, greatly reduced the complexity of the dynamic programming algorithm, and the paths extracted are necessarily simple. However, a path fails to be discovered if any two of its vertices receive the same color, so many random colorings need to be tried to ensure that the desired paths are not missed. The running time of the color coding algorithm is exponential in $\boldsymbol{k}$ and linear time in $\boldsymbol{m}$,

and the storage requirement is exponential in k and linear time in n. This method is much more cost-effective when n is much larger than k, as is the case in our application. In Scott *et al.* (Scott *et al.*, 2006), the color-coding solutions were extended to several biologically motivated extensions of the basic path-finding problem. These include: (1) constraining the set of proteins occurring in a path; (2) constraining the order of occurrence of the proteins in a path; and (3) finding pathway structures that are more general than simple paths. Due to the scarcely experimental populated protein interaction network, like the one we are considering, and due to the fact that little is known about the order of proteins in any signalling pathway in *P. falciparum*, we will consider only solving the first but modified version, namely, given a set of proteins, constrain the maximal number of proteins occurring in a path. Note that this is slightly different to the first problem solved by Scott *et al.* (Scott *et al.*, 2006). They searched in their first biologically motivated problem as indicated above, pathway with a set of proteins, but in *P. falciparum*, known set of proteins that probably formulate signalling pathways are poorly known for a number of reasons (Doering, 1997; Koyama *et al.*, 2009; and Ward *et al.*, 2004). The challenge to extract well defined sets using other eukaryotes is further complicated by the fact that about 60% of the *P. falciparum* proteins are hypothetical and share little or no sequence similarity with other eukaryotes (Koyama *et al.*, 2009; Gardner *et al.*, 2002). Therefore, for each given set that we formulated as shown in chapter four; we sought for pathways that contain a maximal number of proteins in that set. That means other proteins not in that set may be found in our predicted signalling pathways.

3.1.2.1 Constructing the Minimum Paths Algorithm

Let $G = (V, E; w)$ be a digraph with edge cost function $w : E \rightarrow \mathbb{R}$. Let extend the cost function w to the cost matrix $w' : V^2 \rightarrow \mathbb{R} \cup \{\infty\}$ where

$$w'(u,v) = \begin{cases} w(u,v) & if (u,v) \in E \\ 0 & if\ u = v \\ \infty & else \end{cases} \qquad 3.4$$

Normally, the simplest cost function is unit cost where $w(e) = 1$ for all $e \in E$; this can be generalize to both positive cost function where $w(e) > 0$ and negative cost function where $w(e) < 0$. The size parameters for complexity considerations are, as usual, $\mathbf{n} = |\mathbf{V}|$ and $\mathbf{m} = |\mathbf{E}|$. We usually let $\mathbf{V} = (\mathbf{1}, \ldots, \mathbf{n})$.

3.1.2.1.1 Minimum cost (weight) paths: Let $G = (V, E; w)$ be a weighted directed graph. If $e = (u, v)$, we write $w(u,v)$ for $w(e)$. The cost of a path $p = (v_0, v_1, \ldots, v_k)$ is

$$cost(p) = \sum_{i=1}^{k} w(v_{i-1}, v_i) \qquad 3.5$$

The distance from u to v, denoted by $\delta(u,v)$ is the cost of the minimum cost path from u to v. If there is no path from u to v, then $\delta(u,v) = \infty$. The single-source shortest path problem is that, given a graph G and a source vertex s, determine the shortest path from s to v and hence $\delta(s,v)$ for each vertex v.

There are three basic versions:

1. **Single-pair minimum paths:** Given an edge-weighted diagraph $G = (V, E; w, s, t)$ with source and sink $s, t \in V$, find the minimum path from s to t.

2. **Single-source minimum paths:** Given an edge- weighted diagraph $G = (V, E; w, s, t)$ with source $s \in V$, find minimum paths from s to each $t \in V$.
3. **All-pairs minimum paths:** Given an edge- weighted diagraph $G = (V, E, w)$, find the minimum paths between s to t for all $s, t \in V$.

3.1.2.1.2 Path Length and Link Distance: If w is the unit weight (cost) then $w(p) = k$ is just the length of the path $p = (v_0, \ldots, v_k)$, the minimum length (that is, the shortest) of a path from i to j may be called the link distance from i to j. Say j is reachable from i if the links distance from i to j is finite.

3.1.2.1.3 Link-bounded minimum paths: Let k be a $non-negative$ integer. We define a path to be the exact k-link minimum path if it has minimum cost among all k-link paths from its source to its terminus. Let $\delta^{(=k)}(i,j)$ denote the cost of an exact k-link minimum path from i to j and we again have the exact k-link minimum cost matrix $w'^{(=k)}$. We can also consider $at\ most\ k$ links: the corresponding matrix is given by

$$\delta^{(k)}(i,j) = \min_{0 \le \ell \le k} w'^{(=\ell)}(i,j) \qquad 3.6$$

We call $\delta^{(k)}$ the k-link minimum cost matrix.

3.1.2.1.4 Minimum path tree (T): The single-source path algorithms construct a set of minimum paths that comes from a node reachable from the root appears in tree T. Under unit cost, this tree is

just the breadth first search (BFS) tree. If s can reach a negative cycle, then the minimum path tree rooted at s is not defined. The following lemma is a characterization of minimum path trees.

LEMMA: *(minimum path tree), Suppose that* $T \subseteq E$ *is a tree rooted at* $s \in V$ *and* T *spans the set of nodes reachable from* s. *For any node* i *in the tree, let* $d(i)$ *denote the cost from* s *to* i *along a path of* T. *Then* T *is a minimum path tree iff* $\forall (i,j) \in E, d(j) \leq d(i) + w(i,j)$.

The minimal path algorithm employed for extracting signalling pathways from our PPI network is now summarized in the figure 3.2 below.

Function Minimum Paths(Graph,Source)

1. $S \leftarrow \{s\}$
2. $for\ each\ v \in V\ do$
3. $\quad d[v] \leftarrow C[s,v]$
4. $\quad if\ w(s,v) \in E\ then\ P[v] \leftarrow s\ else\ P[v] \leftarrow 0$
5. $\quad for\ i = 1\ to\ n-1\ do$
6. $\quad\quad choose\ w \in V - S\ with\ smallest\ d[w]$
7. $\quad\quad S \leftarrow S \cup \{w\}$
8. $\quad\quad for\ each\ vertex\ v \in V - S\ do$
9. $\quad\quad\quad if\ d[v] > d[w] + C[w,v] then$
10. $\quad\quad\quad\quad d[v] \leftarrow d[w] + C[w,v]$
11. $\quad\quad\quad\quad P[v] \leftarrow w$
12. $\quad\quad\quad endif$
13. $\quad\quad endfor$
14. $\quad endfor$
15. $endfor$
16. end

Fig. 3.2: An algorithm for Minimum path problem. S is the source node, $C[s,v]$ is the weight between s and v and $P[v]$ is the predecessor of the node s.

3.1.3 Statistical Evaluation and Scoring Functional Enrichment of the PPI Network

In order to detect the functional characteristics of the numerically computed modules, we compared them with known functional classification. The paths computed are evaluated using two measures, namely the weighted p-value and the functional enrichment.

3.1.3.1 Calculating Weighted p-value

The meaning of a p-value is related to hypothetical replications of the experiment performed. By definition, if the null hypothesis is true, no more than a fraction α of the replications of an experiment or analysis will yield a p-value smaller than α. This property of the p-value is the basis of all statistical inference based on it. However, as it is a statement about replications of the experiment or analysis, its meaning and interpretation are closely tied to the sampling scheme implied in the model (Goeman and Peter, 2006).

Therefore, given a path with weight w, its weight p-value is defined as the percent of top-scoring paths in random networks (computed using the same algorithm that is applied to the real network) that have weight w or lower, where random networks are constructed by shuffling the edges and weights of the original network, preserving vertex degrees. The p-value for a module M and functional category F is defined by the hypergeometric distribution as:

$$p = 1 - \sum_{i=0}^{k-1} \frac{\binom{|F|}{i}\binom{N-|F|}{|M|-i}}{\binom{N}{|M|}} \qquad 3.7$$

Where module M contains k proteins in F and the PPI networks contains N proteins. The smallest p-value over all functional categories is defined as the p-value of a module which also means that the module is assigned the corresponding function category.

3.1.3.2 Gene Ontology (GO) Annotation

GO is a set of associations from biological phrases to specific genes that are either chosen by trained curators or generated automatically (Ashburner *et al.*, 2000). GO is designed to rigorously encapsulate the known relationships between biological terms and all genes that are instances of these terms. The GO associations allow biologists to make inferences about groups of genes instead of investigating each one individually. With GO, each gene can be automatically assigned its respective attributes.

GO terms are organized hierarchically such that higher level terms are more general and thus are assigned to more genes, and more specific decedent terms are related to parents by either "is a" or "part of" relationships. For example, the nucleus is part of a cell, whereas a neuron is a cell. The relationships form a directed acyclic graph (DAG), where each term can have one or more parents and zero or more children. Users may select the level of generality the terms capture and carry out their analysis accordingly. To identify larger patterns within this group is to seek enrichment - to assess whether some subset of the group shows significant over-representation of some biological characteristic.

Ontology is a structured form of knowledge giving clear definitions of concepts and the relationships among them. Ontology encode, in general, shared consensus among a community of users. Therefore, one advantage of using ontology is the consistency of concepts definition.

Ontologies have become central to biological research, linking literature and biological databases. One of the most used biological ontology is the Gene Ontology (GO), which organizes shared biological knowledge in a structured form, representing a semantic space with clear biological concepts, called GO terms, their definitions and the relations among them (Ashburner *et al.*, 2000).

The three categories of GO

Biological process refers to a biological objective to which the gene or gene product contributes. A process is accomplished via one or more ordered assemblies of molecular functions. Processes often involve a chemical or physical transformation, in the sense that something goes into a process and something different comes out of it. Examples of broad (high level) biological process terms are 'cell growth and maintenance' or 'signal transduction'. Examples of more specific (lower level) process terms are 'translation', 'pyrimidine metabolism' or 'cAMP biosynthesis'.

Molecular function is defined as the biochemical activity (including specific binding to ligands or structures) of a gene product. This definition also applies to the capability that a gene product (or gene product complex) carries as a potential. It describes only what is done without specifying where or when the event actually occurs. Examples of broad functional terms are 'enzyme', 'transporter' or 'ligand'. Examples of narrower functional terms are 'adenylate cyclase' or 'Toll receptor ligand'.

Cellular component refers to the place in the cell where a gene product is active. These terms reflect our understanding of eukaryotic cell structure. As is true for the other ontologies, not all terms are applicable to all organisms; the set of terms is meant to be inclusive. Cellular component includes such terms as 'ribosome' or 'proteasome', specifying where multiple gene products would be found. It also includes terms such as 'nuclear membrane' or 'Golgi apparatus'.

Biological process, molecular function and cellular component are all attributes of genes, gene products or gene-product groups as shown in figures 3(a-c) (Ashburner *et al.*, 2000).

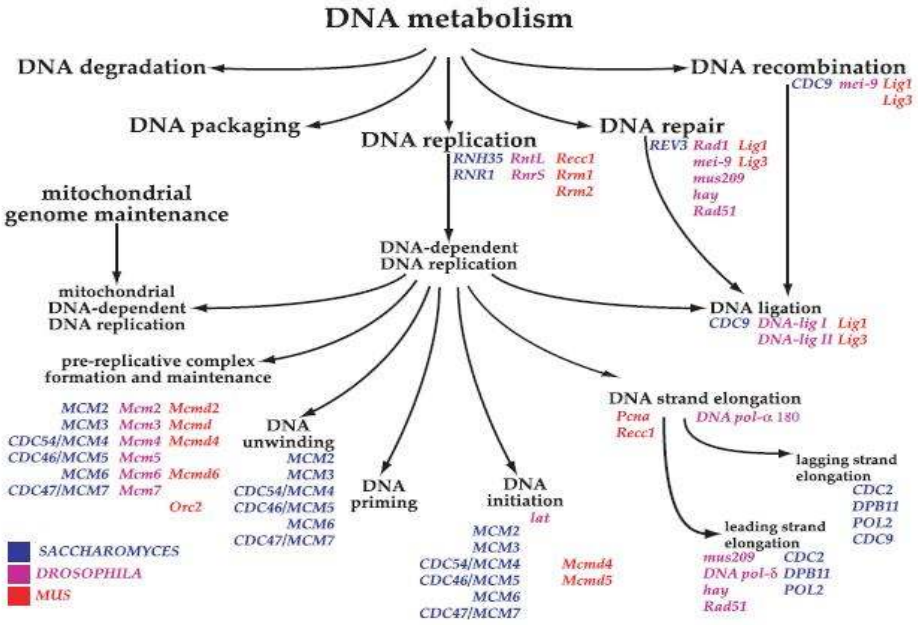


Fig. 3.3a: Biological process ontology. This illustrates a portion of the biological process ontology describing DNA metabolism, a node may have more than one parents, for example, 'DNA ligation' has three parents, 'DNA-dependent DNA replication', 'DNA repair' and 'DNA recombination'(Ashburner *et al.*, 2000)

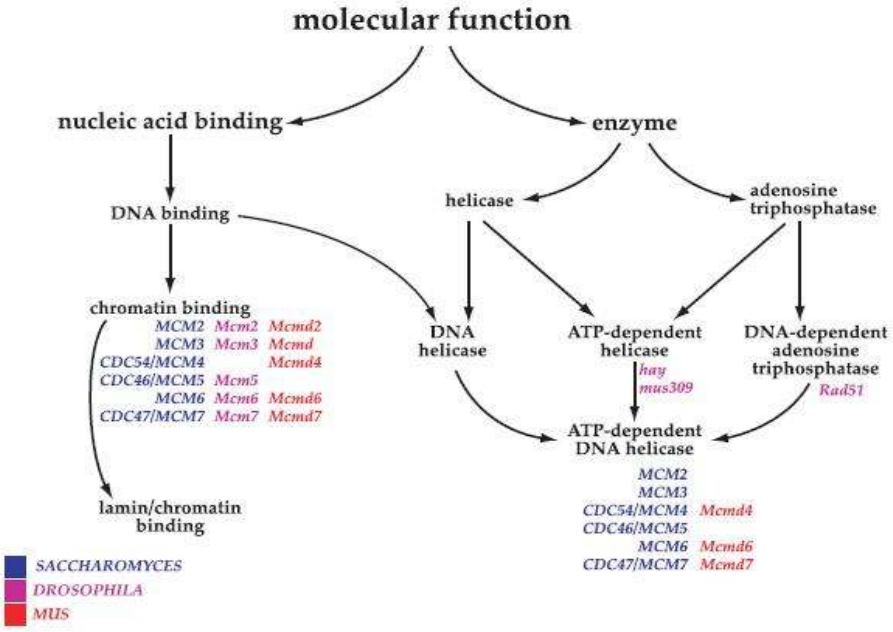


Fig. 3.3b: Molecular function ontology. The ontology is not intended to represent a reaction pathway, but instead reflects conceptual categories of gene product function. A gene product can be associated with more than one node within an ontologies as illustrated by MCM proteins. These proteins have been shown to bind chromatin and to possess ATP-dependent DNA helicase activity, and are annotated to both nodes (Ashburner *et al.*, 2000).

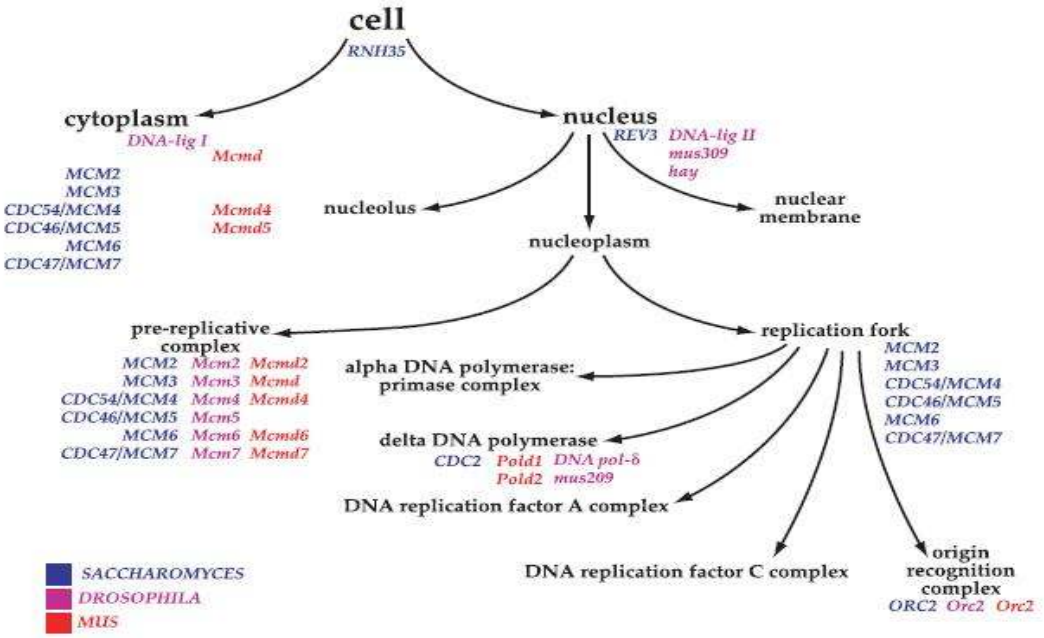


Fig. 3.3c: Cellular component ontology. The ontologies are designed for a genetic eukaryotic cell, and are flexible enough to represent the known differences between diverse organisms (Ashburner *et al.*, 2000).

Ontologies have long been used in an attempt to describe all entities within an area of reality and all relationships between those entities. Ontology comprises a set of well-defined terms with well-defined relationships. The structure itself reflects the current representation of biological knowledge as well as serving as a guide for organizing new data. Data can be annotated to varying levels depending on the amount and completeness of available information. This flexibility also allows users to narrow or widen the focus of queries. Ultimately, ontology can be a vital tool enabling researchers to turn data into knowledge.

The GO project is one of the major efforts in Molecular Biology, for constructing a BioOntology of broad scope and wide applicability. A tremendous effort is being made to annotate all the gene products from many organisms using GO. Along with the Gene Ontology project, many collaborating groups have created organism specific databases, where the organism's gene products are annotated with GO terms. Examples of these dedicated databases include *SGD*, for *Saccharomyces cerevisiae*, with 6463 annotated gene products (Cherry *et al.*, 1998); *FlyBase*, for *Drosophila melanogaster*, with 11312 annotated gene products (Ashburner and Drysdale, 1994); and *WormBase*, for *Caenorhabditis elegans*, with 14698 annotated gene products (Harris *et al.*, 2003). Gene products may have, on each of the GO perspectives, one or more GO terms associated to them. The characterization of gene products through their annotations facilitates their comparison. The annotation of gene products has been an effort for standardizing biological knowledge and the way it is disseminated through the scientific community.

3.1.3.3 Estimating Functional Enrichment

Therefore, to evaluate the functional enrichment of a path P, its proteins are associated with known biological processes using Gene Ontology (GO) annotations (Ashburner *et al.*, 2000). It is then

straight forward to compute the tendency of the proteins to have a common annotation using a method developed in (Sharan *et al.*, 2005). The scoring is done as follows: define a protein to be below a GO term t, if it is associated with t or any other term that is a descendant of t in the GO hierarchy. For each GO term t with at least one protein assigned to it, we computed a hypergeometric p-value based on the following quantities:(1) the number of proteins in P that are below t; (2) the total number of proteins below t; (3) the number of proteins in P that are below all parents of t; and (4) the total number of proteins below all parents of t. The p-value is then further Bonferroni-corrected for multiple testing (Bonferroni, 1936).

3.2 Constructing and Extracting Metabolic Pathways from a Biochemical Metabolic Network.

A metabolic pathway is series of chemical reactions catalyzed by enzymes and are connected by their interactions; that is, the reactants of one reaction, are the products of the previous one, and so on (http://www.biology-online.org, 2011). In other words, metabolic pathways are processes by which organism produces the energy and components it needs to survive.

A metabolic reaction is a pair (I, P) where $I = (I_1, \ldots, I_m)$ are the m input metabolites and $P = (P_1, \ldots, P_n)$ are the n product metabolites of the reaction. Each member of I and P belongs to the set M of the metabolites of the metabolic system under consideration. Note that by this definition, a metabolic reaction is directed and that we omit the stoichiometric coefficients which are not relevant for our current study. Bidirectional reactions are modeled by pairs of unidirectional reactions (I, P) and (P, I). Also note that when applying our theory, we want to

follow how the atoms are transmitted by the reactions and will therefore omit cofactor metabolites from $M, I,$ and P.

A metabolic network is given by listing the metabolic reactions that form the network. Let $R = (R_1, \ldots, R_k)$ be a set of k reactions where each $R_i = (I_i, P_i)$ for some subsets I_i and P_i of M. The corresponding metabolic graph which we also call a metabolic network, has nodes $M \; U \; R$ and arcs as follows: there is a directed arc from $M_j \in M$ to $R_i \in R \; iff \; M_j \in I_i$, and a directed arc from $R_i \in R$ to $M_j \in M \; iff \; M_j \in P_i$. We call the nodes of the network that are in M the metabolite nodes and the nodes in R the reaction nodes. Figure 3.4 gives an example graph in which the reaction nodes are shown as bullets and metabolite nodes contain abbreviated metabolite names.

A metabolic pathway in a metabolic network is a concept that is used somewhat loosely in biochemistry. It seems clear, however, that it is not sufficient to consider only simple paths in a metabolic graph. The metabolic interpretation of the network has to be taken into account: a reaction can operate only if all its input substrates are present in the system. Respectively, a metabolite can become present in a system only if it is produced by at least one reaction. We consider some (source) metabolites to be always present in a system, and denote these metabolites by A. Therefore, our metabolic network is in fact an and-or- graph (Rusel and Norvig, 2003) with reactions as and-nodes and metabolites as or-nodes. A similar interpretation of a metabolic network has been used in Ebenhoh *et al.* (Ebenhoh *et al.,* 2004). To properly take into account this interpretation, Ebenhoh *et al.* (Ebenhoh *et al.,* 2004) define distance measures for metabolite pairs that relate to the complexity of and-or-graphs connecting the pair. Let us start with reachability from source metabolites A:

A reaction $R_i = (I_i, P_i)$ is reachable from A in R, if each metabolite in I_i is reachable from A in R.

A metabolite C is reachable from A in R, if $C \in A$ or some reaction $R_j = (I_j, P_j)$ such that $C \in P_j$ is reachable from A in R.

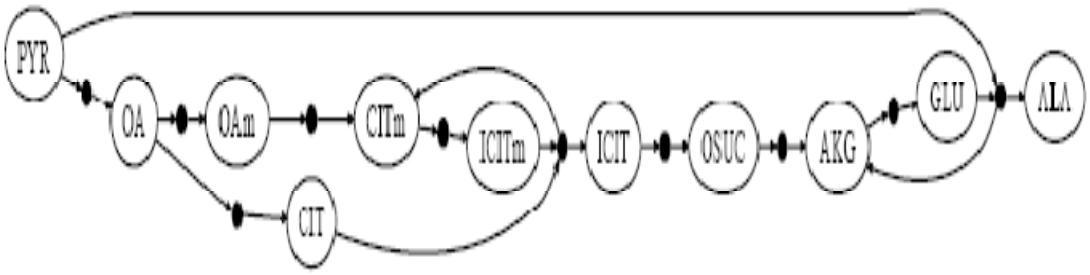


Fig. 3.4: A metabolic pathway from pyruvate (PYR) to alanine (ALA). In this network, pyruvate and glutamate (GLU) are combined to produce alanine. Here, ds(PY R;ALA) = 1.

Ebenhoh *et al.* (Ebenhoh *et al.,* 2004) define metabolic pathways from A as certain minimal sets of reactions that are reachable from A and produce the target metabolite. To this end, for any $F \subseteq R$, we let $Inputs(F)$ denote the set of the input metabolites and $Products(F)$ denote the set of the output metabolites of F. Moreover, we denoted by $W(A,F)$ the subset of R that is reachable from A in F. Hence $W(A,F)$ is the reactions in R that can be reached from A without going outside F.

A feasible metabolism from A is a set $F \subseteq R$ which satisfies (i) $F = W(A,F)$, that is, the entire F is reachable from A without going outside F itself. Specifically, a feasible metabolism from A to t is a set F for which it additionally holds that (ii) $t \in Products(F)$.

We then define that a metabolic pathway from A to t is any minimal feasible metabolism F from A to t, that is, removing any reaction from F leads to violation of requirement (i) or (ii). Thus, a metabolic pathway is a minimal subnetwork capable of performing the conversion from A to t.

Now, different distance measures can be defined. We define the metabolic distance from A to t to be the size of the smallest metabolic pathway from A to t. This distance captures the idea that the distance equals minimum number of reactions in total needed to produce t from A. The production distance from A to t is the smallest diameter taken over all metabolic pathways from A to t, where diameter of a metabolic pathway is taken as the length of the longest simple path in the pathway. Hence, production distance is the minimum number of sequential (successive) reactions needed to convert A to t. In the following, we restrict ourselves to a single source metabolite, that is $|A| = 1$. We then denoted by $d_s(A,t)$ denotes the shortest-path distance.

We use the metabolic (network) graph representation in Koenig *et al.* (Koenig *et al*., 2006). In this work, a graph was established by defining neighbours of metabolites. Two metabolites are neighbours if and only if an enzymatic reaction exists that needs one of the metabolites as input

(needed substrate) and produce the other as output (product). Explicitly, in this graph representation, we have list of all compounds $c_1, c_2, c_3, \ldots$ list of all reactions, $r_1, r_2, r_3, \ldots$ and list of definitions defining the details of all reactions, $rr_1, rr_2, rr_3, \ldots$. For example, for reaction r_1 with input and output compounds c_4, c_3 and c_1, c_2, c_5 respectively, rr_1 will be 2 (number of input compounds) c_4 c_3 3 (number of output compounds) c_1 c_2 c_5. We downloaded PlasmoCyc version 14.6 for *P. falciparum* 3D7 from Biocyc.org on the 28th July, 2010 and biochemical metabolic files for the *P. falciparum* 3D7 last updated 22nd December, 2010 from KEGG. Based on the graph representation in figure 3.4, from PlasmoCyc, we have 608 compounds and 824 reactions. And from KEGG, we have 3011 compounds and 3524 reactions. We found that not all compounds used in the reactions listed for *P. falciparum* 3D7 are listed in the compounds list for *P. falciparum* 3D7 in KEGG database. Therefore we used the file containing all compounds and found 6516 compounds and 4126 reactions. Note that the reactions that were ignored due to the fact we could not find all compounds listed in their definitions are now accounted for. This finding is totally disturbing based on the heavily negative impact of the organism, *P. falciparum,* under consideration. Another further disturbance here is the level of disparity between the graphs extracted from the two databases.

Note that the graph formation above is bipartite, having two type of nodes, namely compounds and reactions. We transformed this graph formation into one with a single type of node, namely reaction, using the following; An edge is drawn between two reactions, r_1 and r_2 if they exchange at least one compound, that is, at least one compound is an output from reaction r_1 and an input into reaction r_2. This representation leads to $R - R$ case way of viewing pathway, that is, from a source reaction to a target reaction. We took to this representation based on the fact the first work

(Croes e*t al*., 2006; Croes e*t al*., 2005) that introduced the $R-R$ concept has been the most effective of all paths finding approaches presented to date in literature (Planes And Beasley, 2009). Converting the bipartite graphs from PlasmoCyc and KEGG (the one with 6516 compounds and 4126 reactions) to our $R-R$ representation, we have a dense graph of 824 reaction nodes with 40299 edges and another heavily dense graph of 4126 reaction nodes with 780560 edges.

We assign weights to the edges on our two graphs using metabolite degrees (Croes e*t al*., 2006). The weight of an edge is the number of metabolites exchanged by the two reactions, that is, given as output by the first and taken in as input by the second. This way, we are able to avoid the problem of assigning a list of side metabolites or otherwise known elsewhere as pool metabolites. And we retain the opportunity to obtain results involving pathways which synthesize for example ATP. In a future work, instead of using the weight on the edges to evaluate the significant of a metabolic pathway, we will extract relevant pathways using atom mapping extraction (Health e*t al*., 2010).

We adapted the minimum path algorithm developed in figure 3.2 to *P. falciparum* metabolic weighted graphs (networks). Such graphs were built from BioCyc (easily updated with MPMP) and KEGG. In figure 3.2, we only applied the *Scott et al.* (Scott *et al.*, 2006) path finding technique for extracting linear pathways. In the very near future, we in-addition plan to adapt and implement their technique for extracting non-linear pathways.

CHAPTER FOUR

EXPERIMENTAL EXPERIENCES

4.1 Prediction of Signalling Pathway

4.1.1 Introduction

We applied the methods above to search for minimum pathways in the *P. falciparum* interaction (weighted graph) network. Our findings can be found in Tables 4.2–4.8 in the appendix A. A snapshot of these tables is presented in Table 4.1.

From the available literature (Doering, *et al*., 1997, Koyama, *et al.*, 2009 and Ward, *et al.*, 2004), we found the following identified classes of signal transduction pathways: cAMP dependent, cGMP dependent, MAP kinase, MAPK, phosphatidylinositol cycle, calcium signalling, protein phosphatases, calcium modulated protein kinase, cyclic nucleotide-dependent, CDK-like kinases, cell cycle kinases and the novel FIKK kinases. Searching the plasmoDB database using their text option using keywords from the above specific identified pathways names, we found 43, 12, 7, 1, 25, 85, 2, 64, 234, 10, 620, and 36 *P. falciparum* genes in these classifications, respectively. Searching for the existence of these genes in the LaCount *et al.* (LaCount *et al*., 2005) protein–protein interaction data, we found none from cAMP dependent, cGMP dependent, MAP kinase and MAPK signalling, but we found 9 from phosphatidylinositol cycle, 32 from calcium signalling, none from protein phosphatases, 28 from calcium modulated protein kinase, 55 from cyclic nucleotide- dependent, none from CDK-like kinases, 130 from cell cycle kinases and 8 from novel FIKK kinases proteins. To ensure we have exhaustively considered all putatively or

annotated signal transduction pathways genes in *P. falciparum*, we further search plasmoDB using the keyword "signal transduction". We found 1183 genes in this category and later filtered out all genes found earlier on using the known signalling pathways listed above. We also observed that there are many genes in the known pathways; which are not part of the filtered 1183 genes. We are then left with 940 genes that are not in any of the known signal transduction pathways. We call this group of genes ''unknown'' signalling pathways. Considering our modified first biological motivated problem; we evaluated the weight p-value and hypergeometric p-value (for functional enrichment) of each path extracted. Following Scott *et al*. (Scott *et al.*, 2006); we extracted pathways with lengths less than or equal to 10 and considered only pathways whose weight and hypergeometric p-values were less than 0.05.We used these criteria; since this is the very first time; this kind of analysis is being done on the only existing protein–protein interaction network (LaCount *et al*., 2005) for *P. falciparum*. We felt; it is important to be able to see explicitly all potential signalling pathways. Furthermore; the results obtained with these criteria have been biological proven to be reliable when applied to yeast protein–protein interaction network (Scott *et al.*, 2006). We found minimum pathways for the genes in the known classifications above as given in Tables 4.2–4.7 in the appendix A. For the ''unknown''; we set I to be each gene in a sequence and found also in Table 4.8 in Appendix A; the listed pathways; whose weight and hypergeometric p-values are both less than 0.05. To visualize the content of the tables diagrammatically, the highlighted (bold) ones in Tables 2–7 can be captured in the usual signalling pathways (Figs. 4.1 and 4.2). They are also highlighted in bold and tagged in Tables 4.1a–4.1d.

4.1.2 Discussion of Results

Since our work is the first attempt to predict the main chains of signal transduction pathways in *P. falciparum*, we adhere strictly to the description of most signal pathways: ‘‘the proteins would transmit the signal from the membrane, where the signal is initiated, towards the nucleus by the activation of transcription factors, which in turn lead to transcription of the final effectors’’. We thus extracted, for example, from the tables of the appendix A, the following signalling pathways, namely, calcium modulated, calcium signalling, cell cycle kinases, cyclic nucleotide, phosphatidylinositol cycle, FIKK in *P. falciparum*. They are highlighted (in red) in the tables in the appendix A and reproduced here in Tables 4.1a–4.1c. The biological validation of these pathways will certainly be useful and attractive for designing new strategies against malaria.

Vaid and Sharma (Vaid and Sharma, 2006) reported the first signalling pathway in *P. falciparum*, which involves activation of protein kinase B-like enzyme (PfPKB) by calcium/Calmodulin(CaM). This is depicted in Fig. 4.3(a) as given in Fig.7 of Vaid and Sharma (Vaid and Sharma, 2006), but it has not been characterized in term of the genes responsible. In their study, they also noted that PfPKB is expressed mainly in the schizont/ merozoite stages of *P. falciparum*, and the calcium necessary for PfPKB activation by CaM is dependent on the activation of phospholipase C(PLC). Therefore, the PfPKB pathway is regulated by CaM and phospholipase C-mediated calcium release. The erythrocyte invasion is a multistep process, which involves the interaction between the merozoites and the erythrocyte followed by reorientation of the merozoite, which leads to the formation of a tight junction between the merozoites apical end and the erythrocyte membrane (Soldati *et al.*, 2004). Vaid *et al*. (Vaid *et al*., 2008) carried out a further study, which showed that the PfPKB pathway is important for erythrocyte invasion. Their study demonstrated that PLC-

mediated control of calcium release is important for merozoites invasion and that CaM may be involved in invasion, due to the localization of CaM at the apical end of the merozoites. It was shown in their previous work that PfPKB is one of the very few CaM targets to be identified in *P. falciparum*, so it then follows that the PfPKB pathway may be important for invasion. Using these findings, we search from Tables 4.6 and 4.8 of the appendix A for pathways that contain the combination of phospholipase C/CaM/PfPKB(protein kinase B-like enzyme). We found only the first entry in Table 4.1d, as depicted in Fig. 4.3(b). We hypothesize this pathway as the Vaid and Sharma (Vaid and Sharma, 2006) Ca^{2+}/Calmodulin-PfPKB signalling pathway characterized partially in terms of the genes responsible. This partial characterization, we believe, is due to the scarcely experimental populated protein interaction network underlining our present computational platform. It is important to note that the potential corresponding Ca2+/Calmodulin-PfPKB signalling pathway of Vaid and Sharma (Vaid and Sharma, 2006) extracted by us is the only pathway that involves a merozoites surface protein among the identified phosphatidylinositol cycle proteins. Further experiments using this set of genes could lead to complete characterization of the pathway in terms of the genes responsible. We also use the keywords ‘‘merozoite’’ and ‘‘erythrocyte’’ to search all entries of Tables 4.2–4.7, and to avoid a trivial result, both of them to search Table 4.8, we found all the other entries in Table 4.1d. They are highlighted (in blue) in the tables of the appendix A. Again, we believe, the biological validation of these pathways will certainly be useful and attractive for designing new strategies against malaria. One interesting thing about the pathways extracted in Table 4.1d is that the proteins of unknown function in Tables 4.1a–4.1c are also the proteins of unknown function in Table 4.1d, except for PF11_0277. Although this is not the aim of this study, but it is worth to mention that the gene PFA0125c (Table 4.8 in Appendix A), which encodes the protein ‘‘erythrocyte binding antigen-181’’may be an

important ‘‘choke point’’ in *P. falciparum*. This has not been mentioned in the analysis by LaCount *et al*. (LaCount *et al*., 2005). Several interesting hypotheses are in particular obtained from the FIKK protein family as shown in Figs. 4.1a and 4.1b and 4.2. It has been noted by Ward *et al*. (Ward *et al*., 2004) and Schneider and Mercereau-Puijalon (Schneider and Mercereau-Puijalon, 2005) that among all the *P. falciparum* protein kinases that have been identified, the FIKK protein family is particularly noteworthy. Koyama *et al*. (Koyama *et al.*, 2009) suggested that the FIKK kinases may have a role in parasite-induced signalling events because members of this family are exported into the erythrocytes where they are found associated with the Maurer’s clefts, and one of the paralogs, R45, is transported to the host cell membrane. This hypothesis is also reflected in our results as we predicted a signal transduction pathway from the FIKK family (see Fig. 4.1a) that ends upon a chloroquine resistance marker protein, PF14_0463, which indicates that interference with FIKK proteins might reverse *P. falciparum* from resistant to sensitive phenotype. The Maurer’s clefts are established by the parasite within its host cell and play an essential role in directing proteins from the parasite to the erythrocyte surface. Presently, they are appreciated as a novel type of secretory organelle. They play an important role in the export of protein from the parasite across the cytoplasm of the host cell to the erythrocyte surface. This is remarkable since erythrocytes lack secretory organelles found in other eukaryotic cells. As a result, the parasite cannot rely on the host cell for its proteins needed and therefore must establish a de-novo secretory system in the host cell cytoplasm, in a compartment outside of its own confines (Frischknecht and Lanzer, 2008). The signal pathways in Fig. 4.1a assign FIKK proteins to this pathway as enabling the resistance of the parasite by excreting chloroquine via an efflux process (Krogstad *et al.*, 1987; and Krogstad *et al.*, 1988). With respect to the Red Blood Cell, Miller *et al*. (Miller *et al*., 2002) noted that what remains completely unknown is which merozoites

surface molecules recognize the RBC surface and then signal the start of the invasion process. There was a hypothesis that suggested that RBC invasion requires the cleavage of a surface protein on the RBC by an unknown parasite serine protease. It has been noted that understanding this pathway will give insight into the parasite virulence and will facilitate rational vaccine design against merozoites invasion (Miller *et al*., 2002). The signalling pathway predicted and depicted in Fig. 4.1b suggests the transduction pathway of that process. The serine protease protein among the proteins involved in this pathway is PFA0130c. From Le Roch *et al*. (Le Roch *et al*., 2003) and Bozdech *et al*. (Bozdech *et al*., 2003a) respectively, it is known that the 48-h *P. falciparum* intraerythrocytic developmental cycle (IDC) initiates with merozoites invasion of RBCs and is followed by the formation of the parasitophorous vacuole (PV) during the ring stage. The parasite then enters a highly metabolic maturation phase, the trophozoite stage, and prior to parasite replication. In the schizont stage, the cell prepares for reinvasion of new RBCs by replicating and dividing to form up to 32 new merozoites. The ring stage, immediately after the merozoite invasion, happens between the 1st hour to the 24th hour, the trophozoite stage begins from the 8th hour to the 33rd hour, while the schizont stage picked up from the 24th hour to the 48th hour. A stage specific expression profiled (see Fig. 4.4) for PFA0130c as obtained from plasmoDB shows that this serine protease protein is highly expressed at the ring stage for all the different cultures of the parasite used in experiments. These hypotheses need of course to be experimentally validated. The popular description of most signalling pathways is: ‘‘the proteins would transmit the signal from the membrane, where the signal is initiated, towards the nucleus by activation of transcription factors, which in turn lead to transcription of the final effectors’’. We applied this to suggest the functions of some genes as depicted in the proposed signal transduction pathways of Fig. 4.2. From their position in Fig. 4.2(a–e), we hypothesize that PF11_0342 is a putative Merozoite

Surface Protein, PFE1605w, PFF022w and PFF1220 are nuclear proteins, and PF07_0056 is a transcription factor. For the other proteins in Tables 4.1a–4.1d, we predicted their functions as described in the second column of Table 4.1e. In an attempt to corroborate our prediction above, we used the DomainSweep software of del Val *et al*. (del Val *et al*., 2007) to predict the function of these proteins as listed in Tables 4.1(a–d). The result of DomainSweep on these proteins is shown in Table 4.1e. From Table 4.1e, two results are conveyed, one, DomainSweep may be able to play a vital role in the re-annotation effort on-going for *P. falciparum* proteins and two, the information extracted by our work (apart from providing information about potential signalling pathways) can be used to collaborate the results of DomainSweep. For the genes with ‘‘unkown’’ classification, the question is, which type of cellular response is transmitted by the predicted signal transduction pathways. The answer to this question will give insight into a number of other signalling pathways and help us to understand better how the malaria parasite reacts and responds to its environments.

Table 4.1a. Extracted potential important signalling transduction pathways from calcium modulated and signalling proteins. Column one indicates the name of the signalling pathway, the second column shows minimum paths extracted, while optimizing the identified number of proteins in the pathway under consideration. The third column shows the weight *p*-value and column four detailed the products (from plasmodb) of the proteins in the identified potential signalling pathways.

Name	Minimum path	*p*-Value	Gene IDs	Products
Calcium modulated				
	PFB0540w → PFB0815w → PFF0220w → PFF0590c → PF14_0632	0.044	PFB0540w PFB0815w PFF0220w PFF0590c PF14_0632	Conserved Plasmodium protein, unknown function Calcium-dependent protein kinase 1 t conserved Plasmodium protein, unknown function homologue of human HSPC025 26S proteasome subunit, putative
	PFA0110w → PFB0540w → FB0815w → FD0090c FF0220w → PFF0590c	0.049	PFA0110w PFB0540w PFB0815w PFD0090c PFF0220w PFF0590c	DNAJ protein, putative Conserved Plasmodium protein, unknown function Calcium-dependent protein kinase 1 Plasmodium exported protein (PHISTa), unknown function conserved Plasmodium protein, unknown function homologue of human HSPC025
	PFB0540w → PFB0815w → PFE0070w → PFF0675c → PF11_0111	0.044	PFB0540w PFB0815w PFE0070w PFF0675c PF11_0111	Conserved Plasmodium protein, unknown function Calcium-dependent protein kinase 1 Interspersed repeat antigen, putative Myosin E asparagine-rich antigen
	PFB0540w → PFB0815w → PFD0985w → PFF0590c → PFF0785w	0.044	PFB0540w PFB0815w PFD0985w PFF0590c PFF0785w	Conserved Plasmodium protein, unknown function Calcium-dependent protein kinase 1 Transcription factor with AP2 domain(s), putative Homologue of human HSPC025 Ndc80 homologue, putative
Calcium Signalling				
	PF10_0143 → PF11_0142 → PF11_0239 → MAL13P1.206	0.033	PF10_0143 PF11_0142 PF11_0239 MAL13P1.206	Transcriptional activator ADA2, putative ubiquitin domain containing protein calcium-dependent protein kinase, putative Na^+-dependent Pi transporter, sodium-dependent phosphate transporter
	PF11_0142 → PF11_0239 → PF13_0197 → MAL13P1.206	0.038	PF11_0142 PF11_0239 PF13_0197 MAL13P1.206	ubiquitin domain containing protein calcium-dependent protein kinase, putative Merozoite Surface Protein 7 precursor, MSP7 Na^+-dependent Pi transporter, sodium-dependent phosphate transporter
	PF11_0142 → PF11_0239 → MAL13P1.206	0.022	PF11_0142 PF11_0239 MAL13P1.206	ubiquitin domain containing protein Calcium-dependent protein kinase, putative Na^+-dependent Pi transporter, sodium-dependent phosphate transporter
	PF11_0142 → PF11_0239 → MAL13P1.206 → PF14_0059	0.046	PF11_0142 PF11_0239 MAL13P1.206 PF14_0059	Ubiquitin domain containing protein Calcium-dependent protein kinase, putative Na^+-dependent Pi transporter, sodium-dependent phosphate transporter Conserved plasmodium protein, unknown function
	PF11_0142 → PF11_0239 → MAL13P1.206 → PF14_0678	0.033	PF11_0142 PF11_0239 MAL13P1.206 PF14_0678	Ubiquitin domain containing protein Calcium-dependent protein kinase, putative Na^+-dependent Pi transporter, sodium-dependent phosphate transporter Exported protein 2
	PFD0090c → PFF0670w → PF08_0048 → PF11_0239	0.041	PFD0090c PFF0670w PF08_0048 PF11_0239	Plasmodium exported protein (PHISTa), unknown function Transcription factor with AP2 domain(s), putative ATP-dependent helicase, putative Calcium-dependent protein kinase, putative

Table 4.1b. Extracted potential important signalling transduction pathways from cell cycle, cyclic nucleotide and phosphatidylinositol cycle proteins. Column one indicates the name of the signalling pathway, the second column shows minimum paths extracted, while optimizing the identified number of proteins in the pathway under consideration. The third column shows the weight *p*-value and column four detailed the products (from plasmodb) of the proteins in the identified potential signalling pathways.

Name	Minimum paths	*p*-Value	Gene ID	Products
Cell Cycle				
	PFE1370w → PF10_0143 → PF10_0272	0.021	PFE1370w PF10_0143 PF10_0272	hsp70 interacting protein, putative Transcriptional activator ADA2, putative 60S ribosomal protein L3, putative
Cyclic nucleotide				
	PFB0190c → PFC0435w → PFE0660c → PF10_0254 → MAL13P1.202	0.033	PFB0190c PFC0435w PFE0660c PF10_0254 MAL13P1.202	Conserved Plasmodium protein, unknown function Conserved Plasmodium protein, unknown function Purine nucleotide phosphorylase, putative Conserved Plasmodium protein, unknown function Conserved Plasmodium protein, unknown function
	PFC0435w → PFE0660c → PF08_0129 → PF11_0111 → MAL13P1.202	0.028	PFC0435w PFE0660c PF08_0129 PF11_0111 MAL13P1.202	Conserved Plasmodium protein, unknown function purine nucleotide phosphorylase, putative Serine/threonine protein phosphatase, putative asparagine-rich antigen Conserved Plasmodium protein, unknown function
	PFC0435w → PFE0660c → PFL2520w → MAL13P1.202	0.016	PFC0435w PFE0660c PFL2520w MAL13P1.202	Conserved Plasmodium protein, unknown function Purine nucleotide phosphorylase, putative Reticulocyte-binding protein 3 homologue Conserved Plasmodium protein, unknown function
Phosphatidylinositol Cycle				
	PFE0750c → PFL1930w → MAL13P1.256	0.005	PFE0750c PFL1930w MAL13P1.256	RNA recognition motif, putative Conserved Plasmodium protein, unknown function Phosphatidylinositol transfer protein, putative
	PFA0110w → PFE0750c → MAL13P1.256 → PF14_0257	0.046	PFA0110w PFE0750c MAL13P1.256 PF14_0257	DNAJ protein, putative RNA recognition motif, putative phosphatidylinositol transfer protein, putative conserved protein, unknown function
	PFA0110w → PFD0090c → PFE0750c → MAL13P1.256	0.029	PFA0110w PFD0090c PFE0750c MAL13P1.256	DNAJ protein, putative Plasmodium exported protein (PHISTa), unknown function RNA recognition motif, putative Phosphatidylinositol transfer protein, putative
	PFE0750c → PFF1050w → PF10_0115 → MAL13P1.256	0.043	PFE0750c PFF1050w PF10_0115 MAL13P1.256	RNA recognition motif, putative Nascent polypeptide associated complex alpha chain, putative

Table 4.1c. Extracted potential important signalling transduction pathways from the FIKK family proteins. Column one indicates the name of the signalling pathway, the second column shows minimum paths extracted, while optimizing the identified number of proteins in the pathway under consideration. The third column shows the weight *p*-value and column four detailed the products (from plasmodb) of the proteins in the identified potential signalling pathways.

Name/Figure tag	Minimum paths	*p*-Value	Gene ID	Products
FIKK				
	PFA0130c → PFE1590w → MAL8P1.153 → PFA0215w	0.046	PFA0130c PFE1590w MAL8P1.153 PFA0220w	Serine/Threonine protein kinase, FIKK family, putative Early transcribed membrane protein 5, ETRAMP5 Transcription factor with AP2 domain(s), putative Ubiquitin carboxyl-terminal hydrolase, putative
	PFA0130c → PFE1590w → MAL8P1.153 → PF10_0075	0.039	PFA0130c PFE1590w MAL8P1.153 PF10_0075	Serine/Threonine protein kinase, FIKK family, putative Early transcribed membrane protein 5, ETRAMP5 Transcription factor with AP2 domain(s), putative Transcription factor with AP2 domain(s), putative
Fig. 2a	**PFA0130c → PFE1590w → MAL8P1.153 → PF11_0342**	**0.036**	**PFA0130c PFE1590w MAL8P1.153 PF11_0342**	**Serine/Threonine protein kinase, FIKK family, putative Early transcribed membrane protein 5, ETRAMP5 transcription factor with AP2 domain(s), putative conserved Plasmodium protein, unknown function**
	PFA0130c → PFE1590w → PF10_0232 → PF11_0506	0.036	PFA0130c PFE1590w PF10_0232 PF11_0506	Serine/Threonine protein kinase, FIKK family, putative Early transcribed membrane protein 5, ETRAMP5 Chromodomain-helicase-DNA-binding protein 1 homolog, putative Antigen 332, DBL-like protein
Fig. 1b	**PFA0130c → PFE1590w → PFF0590c → MAL8P1.153 → PFL1385c**	**0.046**	**PFA0130c PFE1590w PFF0590c MAL8P1.153 PFL1385c**	**Serine/Threonine protein kinase, FIKK family, putative Early transcribed membrane protein 5, ETRAMP5 Homologue of human HSPC025 Transcription factor with AP2 domain(s), putative Merozoite Surface Protein 9, MSP-9**
Fig. 1a	**PFA0130c → PFE1590w → MAL8P1.153 → MAL8P1.23 → PF14_0463**	**0.039**	**PFA0130c PFE1590w MAL8P1.153 MAL8P1.23 PF14_0463**	**Serine/Threonine protein kinase, FIKK family, putative Early transcribed membrane protein 5, ETRAMP5 Transcription factor with AP2 domain(s), putative Ubiquitin-protein ligase 1, putative Chloroquine resistance marker protein**
Fig. 2b	**PFA0130c → PFE1590w → PFE1605w → MAL8P1.153**	**0.036**	**PFA0130c PFE1590w PFE1605w MAL8P1.153**	**Serine/Threonine protein kinase, FIKK family, putative Early transcribed membrane protein 5, ETRAMP5 Plasmodium exported protein (PHISTb), unknown function Transcription factor with AP2 domain(s), putative**
Fig. 2c	**PFA0130c → PFE1590w → PFF0220w → PFF0590c → MAL8P1.153**	**0.046**	**PFA0130c PFE1590w PFF0220w PFF0590c MAL8P1.153**	**Serine/Threonine protein kinase, FIKK family, putative Early transcribed membrane protein 5, ETRAMP5 Conserved Plasmodium protein, unknown function Homologue of human HSPC025 Transcription factor with AP2 domain(s), putative**
Fig. 2d	**PFA0130c → PFE1590w → PFF1220w → MAL8P1.153**	**0.036**	**PFA0130c PFE1590w PFF1220w MAL8P1.153**	**Serine/Threonine protein kinase, FIKK family, putative Early transcribed membrane protein 5, ETRAMP5 Conserved Plasmodium protein, unknown function transcription factor with AP2 domain(s), putative**
Fig. 2e	**PFA0130c → PFE1590w → PF07_0056 → MAL8P1.153 → MAL8P1.23**	**0.036**	**PFA0130c PFE1590w PF07_0056 MAL8P1.153 MAL8P1.23**	**Serine/Threonine protein kinase, FIKK family, putative Early transcribed membrane protein 5, ETRAMP5 Conserved Plasmodium protein, unknown function Transcription factor with AP2 domain(s), putative Ubiquitin-protein ligase 1, putative**
	PFA0130c → PFE1590w → MAL8P1.153	0.007	PFA0130c PFE1590w MAL8P1.153	Serine/threonine protein kinase, FIKK family, putative Early transcribed membrane protein 5, ETRAMP5 Transcription factor with AP2 domain(s), putative
	PFA0130c → PFE1590w → MAL8P1.153 → PF08_0034	0.041	PFA0130c PFE1590w MAL8P1.153 PF08_0034	Serine/threonine protein kinase, FIKK family, putative Early transcribed membrane protein 5, ETRAMP5 Transcription factor with AP2 domain(s), putative histone acetyltransferase GCN5, putative
	PFA0130c → PFE1590w → MAL8P1.153 → MAL8P1.23	0.036	PFA0130c PFE1590w MAL8P1.153 MAL8P1.23	Serine/threonine protein kinase, FIKK family, putative Early transcribed membrane protein 5, ETRAMP5 Transcription factor with AP2 domain(s), putative Ubiquitin-protein ligase 1, putative

Table 4.1d. Vaid and Sharma (2006) and Vaid et al. (2008) motivated extracted potential important signalling transduction pathways. Column one indicates the name of the signalling pathway, the second column shows minimum paths extracted, while optimizing the identified number of proteins in the pathway under consideration. The third column shows the weight *p*-value and column four detailed the products (from plasmodb) of the proteins in the identified potential signalling pathways.

Name/Figure tag	**Minimum paths**	***p*-Value**	**Gene ID**	**Products**
Phosphatidylinositol cycle				
Fig. 3b	**PFE0750c → PFL1385c → MAL13P1.256**	**0.005**	**PFE0750c PFL1385c MAL13P1.256**	**RNA recognition motif, putative Merozoite Surface Protein 9, MSP-9 Phosphatidylinositol transfer protein, putative**
Calcium modulated				
	PFB0540w → PFB0815w → PFF0675c → PF10_0345 → PF11_0111	0.044	PFB0540w PFB0815w PFF0675c PF10_0345 PF11_0111	Conserved Plasmodium protein, unknown function Calcium-dependent protein kinase 1 myosin E merozoite surface protein 3 asparagine-rich antigen
	PFB0540w → PFB0815w → PFF0220w → PFF0590c → PFL1385c	0.024	PFB0540w PFB0815w PFF0220w PFF0590c PFL1385c	Conserved Plasmodium protein, unknown function Calcium-dependent protein kinase 1 Conserved Plasmodium protein, unknown function Homologue of human HSPC025 Merozoite Surface Protein 9, MSP-9
	PFB0540w → PFB0815w → PFF1365c → MAL7P1.12	0.029	PFB0540w PFB0815w PFF1365c MAL7P1.12	Conserved Plasmodium protein, unknown function Calcium-dependent protein kinase 1 HECT-domain (ubiquitin-transferase), putative Erythrocyte membrane-associated antigen
Cell Cycle				
	PF10_0254 → PF10_0272 → PFL1385c	0.021	PF10_0254 PF10_0272 PFL1385c	Conserved plasmodium protein, unknown function 60S ribosomal protein L3, putative Merozoite Surface protein 9, MSP-9
cyclic nucleotide				
	PFC0435w → PFE0660c → PF10_0281 → PF11_0224	0.016	PFC0435w PFE0660c PF10_0281 PF11_0224	Conserved Plasmodium protein, unknown function Purine nucleotide phosphorylase, putative Merozoite TRAP-like protein, MTRAP Circumsporozoite-related antigen
Unknown signal transduction groups				
	PFA0125c → PFE0570w → PF11_0277 → PFL1385c	0.027	PFA0125c PFE0570w PF11_0277 PFL1385c	Erythrocyte binding antigen-181 RNA pseudouridylate synthase, putative Conserved Plasmodium protein, unknown function Merozoite Surface Protein 9, MSP-9

Table 4.1e: DomainSweep functional prediction for the proteins with unknown function in Tables 4.1a–4.1d above. m.p = membrane protein, n.p = nuclear protein, t.f = transcription factor, m.s.p = merozoite surface protein. The third column indicates putative hits that do not fulfill the criteria of a significant hit but have a score above a certain threshold. A significant hit has at least two hits of domains which are described in two protein family databases AND which are members of the same INTERPRO family/domain, or at least two motif hits or two block hits in correct order as described in an individual entry of the Prints or the Blocks database. We listed the first two as predicted from DomainSweep. Columns four and five indicate selected hits from and name of the specific domains or families, respectively.

Gene ID	Our prediction from predicted signalling pathways	Putative hits	Selected protein domains and families hits	The selected protein domains and families
PFB0540w	m.p.	GPCR, family 3, metabotropic glutamate receptor 3 Involucrin repeat	Ribosomal protein L35	PFAM A
PFF0220w	n.p.	Mycobacterial pentapeptide repeat Ribosomal protein VAR1	Anticodon nuclease activator family Bipartite nuclear lo	PFAM A PROSITE-PROFILES
PF14_0059	t.f.	Protein of unknown function DUF1754, eukaryotic Daxx protein	Transcription factor IIA, alpha/bet Transcription elongation factor Elf1	BLOCKS PFAM A
PFB0190c	m.p.	Sel1 repeat	Mitochondrial ribosomal protein (VAR1) Plasmodium histidine-rich protein (HRP) Putative stress-responsive nuclear e Bipartite nuclear lo IF-2: translation initiation factor I S8e: ribosomal protein S8.e	PFAM APROSITE-PROFILES TIGRFAMS
PFC0435w	t.p.	Botulinum neurotoxin	Mitochondrial ribosomal protein (VAR1) YL1 nuclear protein Bipartite nuclear lo ETRAMP: early transcribed membrane	PFAM APROSITE-PROFILES TIGRFAMS
PF10_0254	n.p.	Bipartite nuclear lo Asparagine-rich regi	Ribosomal protein S15 Transcription factor S-II (TFIIS), ce Heat shock factor binding protein 1 Plasmodium histidine-rich protein (HRP) Ribosomal protein S26e rho: transcription termination factor	PFAM A TIGRFAMS
MAL13P1.202	t.f.	Clostridium neurotoxin, translocation Phosphatidylinositol-4, 5-bisphosphate phosphodiesterase beta, conserved site	phage_rinA: phage transcriptional reg rho: transcription termination factor	TIGRFAMS
PFL1930w	t.f.	Uso1/p115 like vesicle tethering protein, head region	bZIP transcription factor Putative stress-responsive nuclear en Asparagine-rich regi Bipartite nuclear lo	PFAM APROSITE-PROFILES
PF14_0257	t.f.	Transcription factor IIA, alpha/beta subunit Translation initiation factor eIF3 subunit	P21_Cbot: transcriptional regulator	TIGRFAMS
PFD0090c	t.f.	Exported protein, PHISTa/c, conserved domain, Plasmodium Basic helix–loop–helix, Nulp1-type	Myb-like DNA-binding domain	PFAM A
PF11_0342	m.s.p.	P60-like TAFII55 protein, conserved region	Merozoite surface protein (SPAM)	PFAM A
PFE1605w	n.p.	Apoptosis regulator, Bcl-2 relatedANTIGEN SURFACE MALARIA	Nuclear factor I protein pre-N-termin Putative stress-responsive nuclear en	PFAM APFAM A
PFF1220w	n.p.	Botulinum neurotoxin Asparagine-rich regi Lysine-rich region p	Mitochondrial ribosomal protein (VAR1)	PFAM A
PF07_0056	t.f.	Subtilin biosynthesis protein SpaC Putative 5-3 exonuclease	Poxvirus Late Transcription Factor VL	PFAM A
PF11_0277	t.f.	Autophagy-related protein 6 Uso1/p115 like vesicle tethering protein, head region	ribosomal protein L29	TIGRFAMS

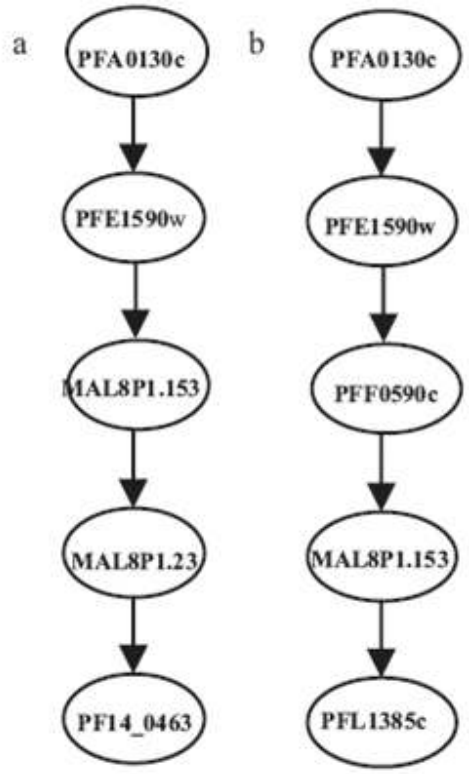


Fig. 4.1: Potential vital signalling pathways from FIKK family proteins as extracted into table 4.1c. (a) Potential chloroquine resistance signalling pathway and (b) Potential signalling pathway that may have signal the start of the invasion process of Red Blood Cell (RBC) by the merozoites

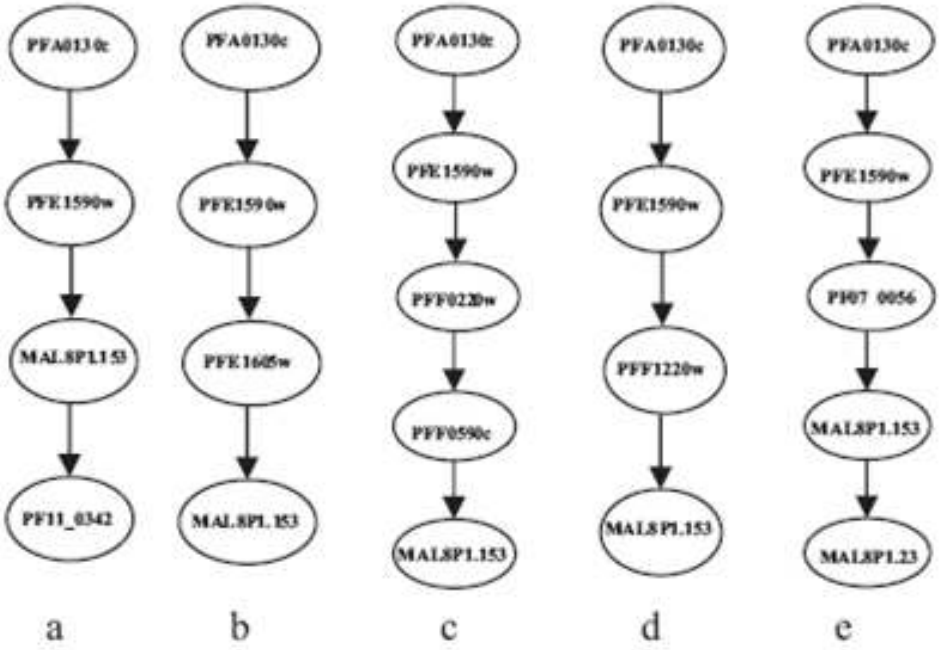


Fig. 4.2: Hypothetical functional predictions from some predicted signalling pathways from the FIKK family proteins as extracted into Table 4.1a. (a) P11_0342 was predicted to be a Merozoite Surface Protein. (b) PFE1605w, (c) PFF0220w and (d) PFF1220w as nucleus proteins and (e) PF07_0056 as a transcription factor.

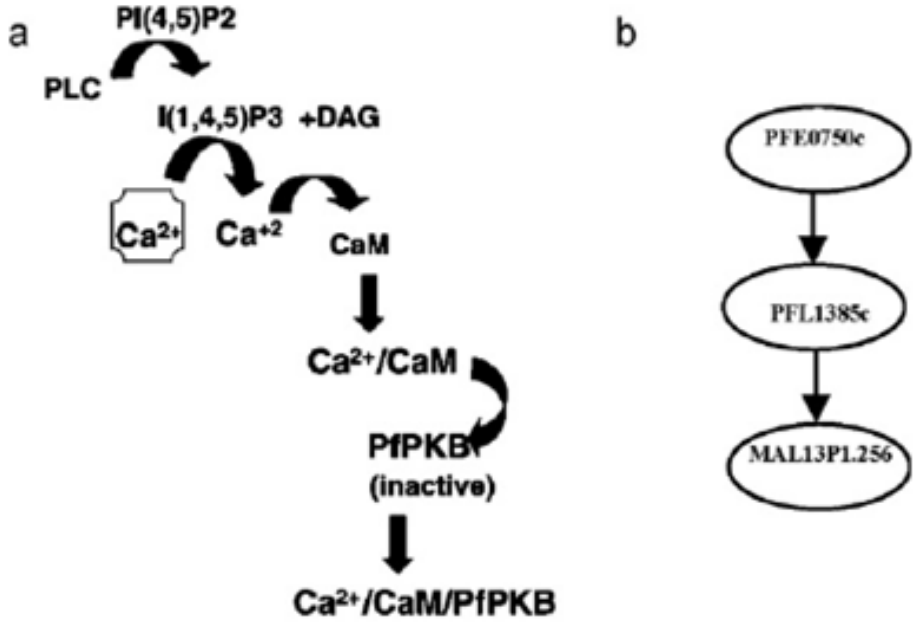


Fig. 4.3: The Ca^{2+}/Calmodulin-PfPKB signalling pathway as biologically dissected by Vaid and Sharma (2006). (b) The potential corresponding Ca^{2+}/Calmodulin-PfPKB signalling pathway of Vaid and Sharma (2006) from the protein-protein interaction data of LaCount *et al*. (LaCount *et al*., 2005).

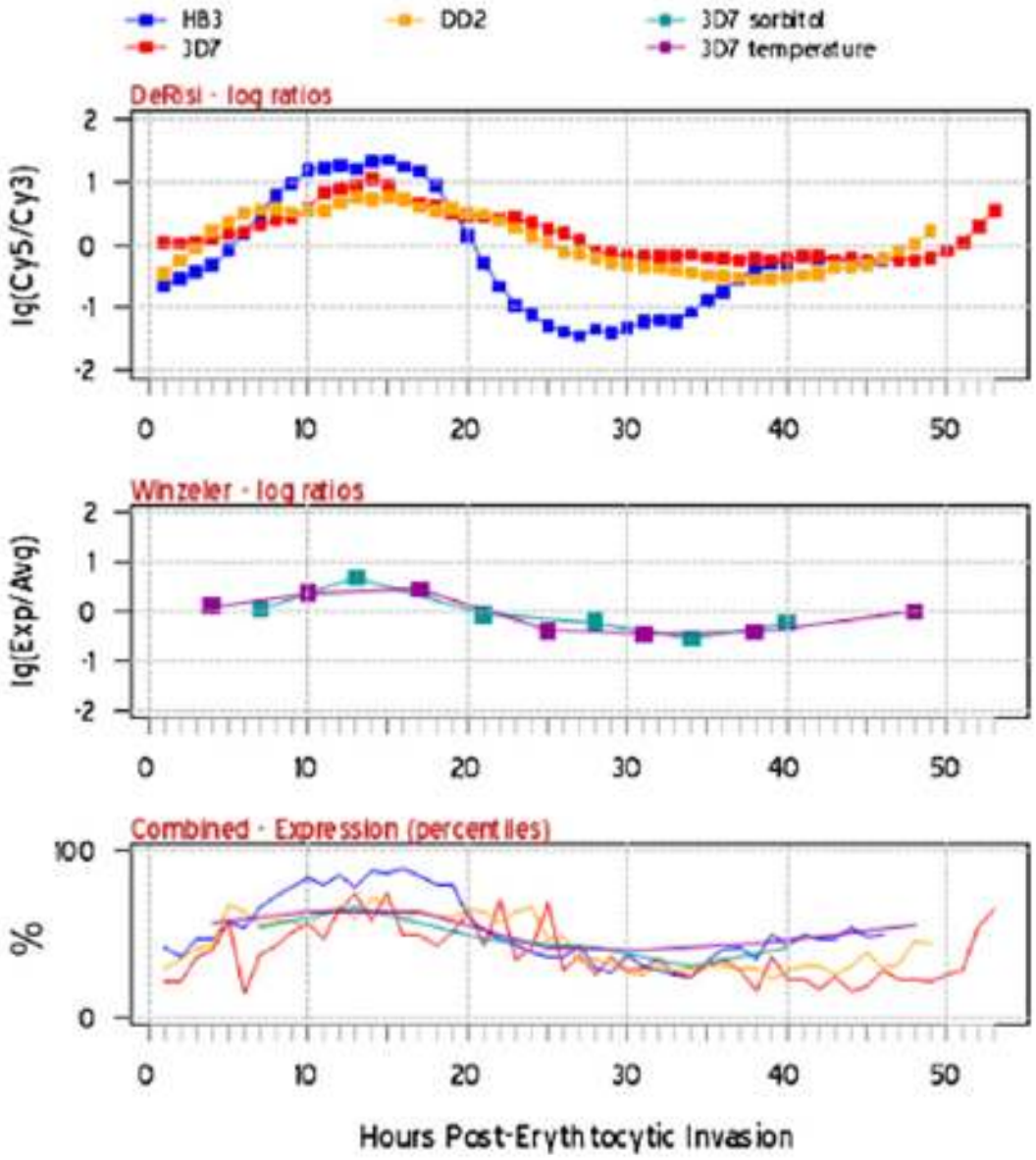


Fig. 4.4: Stage specific expression profile data for PFA0130c as obtained from plasmoDB shows that this serine protease protein is highly expressed at the ring stage for all the different cultures (HB3, 3D7, DD2 in Bozdech *et al*., 2003) of the parasite used in experiments.

4.2 Prediction of Metabolic Pathways

In metabolic pathway, we applied the methods in section 3.2.1 to extract metabolic pathways in the *P. falciparum* metabolic weighted graph (network) of section 3.2. PlasmoCyc provided an enzrxn flat file that enables us to mine out genes that catalyse each reaction in the network. We found that some reactions do not have genes encoding their enzymes, while some have many (up to 10 in some cases). This naturally allows us to incorporate measurement data, such as gene or protein expression data, into our analysis. This is specifically important in discovering condition-specific pathways (Pitkaenen *et al*.,2009).

In a pilot exercises, we tested our algorithm (for four selected pathways: Pyruvate, Glutamate, Glycolysis and Mitochondrial TCA) on the metabolic graph from KEGG and compare our results with the results obtained from ReTrace (Pitkaenen *et al*.,2009) and atommetanet (Health e*t al*., 2010). Our results compare favourably with the results from the two algorithms. We however compare the results with genes classified into these pathways from Plasmodb and found a lot of false positiveness. We however compare the runs of our algorithm on metabolic graphs from KEGG and PlasmoCyc (from BioCyc). The results are remarkably different and the results from PlasmoCyc produce less false positiveness when compared to the results from Plasmodb. We identify 2, 1, 2, 4 gene(s) in addition to belong to Pyruvate, Glutamate, Glucolysis and Mitochondrial TCA respectively. Some of the genes have not been classified earlier to any known metabolic pathways.

CHAPTER FIVE

CONCLUSIONS AND FUTURE WORK

In this work, we have been able for the first time to mine signal transduction pathways from the most deadly malaria parasite, *P. falciparum*. We have been able to use these results to suggest important hypotheses that can help to explain the mechanisms that signal chloroquine resistance process by the malaria via an efflux process, and which signals start the invasion process of RBC by the merozoites. One of our predicted pathways may also have provided the Vaid and Sharma (Vaid A.and Sharma P., 2006) Ca2+/Calmodulin-PfPKB signalling pathway characterized in terms of the genes responsible. The PfPKB pathway has been shown to be important to the erythrocyte invasion (Vaid *et al.*, 2008). We have also been able to use our results to predict functionality for some proteins.

For the metabolic pathways, following our experimental experiences from our pilot run, we have been able to identify additional genes to four key pathways and also in the process annotate genes of putative hypothetic functions. We plan to build a very accurate and comprehensive metabolic network for this important organism, the malaria parasite, *P. falciparum*. The following findings necessitate this lead. Recall that using the graph representation of Koenig *et al*. (Koenig *et al*., 2006), there are two major setbacks observed from the graphs derived from KEGG and PlasmoCyc (see pages 93-95). In a separate work, we found about 30 reactions (Adeoye *et al*., 2010, unpublished manuscript) from MPMP that were not listed in PlasmoCyc. Furthermore, the compounds in a good number of the 30 reactions definitions were not listed in PlasmoCyc. We also found that a number of genes with functional classifications in Plasmodb are without these

classifications in PlasmoCyc. Finally we also found out genes with functional classifications in KEGG (confirmed by our methods to be correct) that are without these classifications either in PlasmoCyc or Plasmodb.

Presently, from plasmodb, for *P. falciparum*, we have 137 metabolic pathways covering 2521 genes. This is just about half of the annotated genes of *P. falciparum*. Therefore, we plan in the nearest future to deploy our techniques (section 3.2) at a large scale for all known pathways. This we know from the findings, will help to both reconfirm existing classifications and classify genes of unknown functions into functional modules - metabolic pathways. We will also find paths in attempts to engineer the discovery of unknown metabolic pathways in *P. falciparum.*

To further address the problem of data scarcity (in particular with regard to the protein–protein interaction information available for the malaria parasite), we need to develop techniques to deal with missing edges, i.e. protein–protein interaction that have never been observed but exist in reality. One way to do this is to integrate transcription factors into the derived network, resulting into what has been called an integrated cellular weighted network of transcription–regulation and protein–protein interaction (Yerger-Lotem *et al*., 2004). For the malaria parasite *P. falciparum*, only about a third of the number of transcription-associated proteins (TAPs) usually found in the genome of a free-living eukaryote is presently known (Coulson *et al.*, 2004).

REFERENCES

Adebiyi, E. F. (2006). *On specific system biology computational tools for Plasmodium falciparum.* DAAD grant.

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K. and Watson, J. D. (1994). *Molecular Biology of the Cell (3rd ed.).* Garland Pub. Inc., New York.

Albert R. (2005). Scale-free networks in cell biology *J. Cell Sci.* 118, 4947–57.

Alm E. and Arkin A. P. (2003). Biological networks. *Curr. Opin. Struct. Biol., 13(2), 193–202*.

Almaas, E., Kovacs, B., Vicsek, T., Oltvai, Z.N. and Barabási, A.L. (2004). Global organization of metabolic fluxes in the bacterium Escherichia coli. *Nature. Vol. 427, 839.*

Anamika N, Srinivasan N, Krupa A (2005). A genomic perspective of protein kinases in Plasmodium falciparum *PROTEINS: Structure, Function and Bioinformatics; Vol.58, pg.180-189.*

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

Ashburner M. and Drysdale R. (1994). *The Drosophila genetic database, Development.* FlyBase Vol.120, 2077-2079.

Bader, G.D. and Hogue, C.W. (2003). *An automated method for finding molecular complexes in large protein interaction networks.* BMC Bioinformatics, Vol. 4, pg. 2.

Bader, J., Chaudhuri, A., Rothberg, J., and Chant, J. (2004). *Gaining confidence in high-throughput protein interaction networks.* Nat. Biotechnology, Vol. 22, pg. *78-85.*

Bannister LH, Hopkins JM, Fowler RE, Krishna S, Mitchell GH (2000). *A brief illustrated guide to the ultrastructure of Plasmodium falciparum asexual blood stages,* Parasitology today; Vol. 16, pg.427-433.

Barabasi, A.-L. and Albert, R(1999). *Emergence of scaling in random networks.* Science Vol. 286(5439), pg. 509.512.

Baraba´si1 A. L, Oltvai Z. N.(2004). *Network biology: Understanding the cell's functional organization.* Nat. Rev. Genet, Vol. 5, pg. 101–113.

Barthelemy, M., Gondran, B. and Guichard, E. (2003). *Spatial structure of the Internet traffic.* Physica A. Vol. 319, pg. 633-42.

Bebek, G and Yang, J. (2007). *PathFinder: mining signal transduction pathway segments from protein-protein interaction networks.* BMC Bioinformatics Vol. 8, pg. 335.

Ben Mamoun, C., Gluzman, L.Y. and Hott, C. (2001). *Coordinated programme of gene expression during asexual intraerythrocytic development of the human malaria parasite plasmodium falciparum revealed by microarray analysis.* Mol. Microbiol., Vol.39(1), pg. 26–36.

Bonferroni CE (1936). *Teoria statistica delle classi e calcolo delle probabilità.* Pubblicazioni del Istituto Superiore di Scienze Economiche e Commerciali di Firenze, Vol. 8, pg.3-62.

Bozdech Z., Linas, M., Pulliam B. I., Wong E. D., Zhu J., DeRisi J. I., (2003a). *The transcriptome of the intraerythroecytic developmental cycle of plasmodium falciparum,* PLoS Biol., 1, E5.

Breman, J.G., Alilio, M.S. and Mills, A. (2004). *Conquering the intolerable burden of malaria: what's new, what's needed: a summary. Am J. Trop. Med. Hyg., Vol. 71(2), pg.1–15.*

Brun, C., Chevenet, F., Martin, D., Wojcik, J., Guenoche, A. and Jacq, B. (2003). *Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network.* Genome Biol.,Vol. 5, R6.

Bu D, Zhao Y, Cai L, Xue H, Zhu X, Lu H, Zhang J, Sun S, Ling L, Zhang N, Li G, Chen R. (2003). *Topological structure analysis of the protein-protein interaction network in budding yeast.* Nucleic Acids Res., Vol. 31, pg. 2443-2450.

Carter, Richard and Kamini N. Mendis(2002). *Evolutionary and historical aspects of the burden of malaria.* Clin Microbiol Rev. Vol. 15(4), pg.565 – 591.

Cherry JM, Adler C, Ball C, Chervitz SA, Dwight SS, Hester ET, Jia Y, Juvik G, Roe T, Schroeder M, Weng S, Botstein D. (1998). *SGD: Saccharomyces Genome Database.* Nucleic Acids Res 1998 26(1):73-80.

Chishti AH, Maalouf GJ, Marfatia S, Palek J, Wang W, Fisher D, Liu SC (1994). *Phosphorylation of protein 4.1 in Plasmodium falciparum-infected human red blood cells,* Blood; Vol. 83, pg. 3339-3345.

Croes, D., Couche, F., Wodak, S. J. and van Helden, J. (2005). *Metabolic PathFinding: inferring relevant pathways in biochemical networks.* Nucleic Acids Res, Vol. 33, pg.326–30.

Croes D, Couche F, Wodak SJ, van Helden J. (2006). *Inferring meaningful pathways in weighted metabolic networks.* J Mol Biol., Vol.356(1), pg.222-36.

Coulson, R.M.R., Hall, N., Ouzounis, C.A. (2004).*Comparative genomics of transcriptional control in the human malaria parasite P. falciparum.* GenomeRes. Vol.14, pg. 1548–1554.

Curtisa RK, Oresˇicˇ M, Vidal-Puiga A. (2005). *Pathways to the analysis of microarray data.* Trends Biotechnol Vol. 23, pg. 429–35.

Daaron Acemoglu and Asu Ozdaglar (2009). *Lecture note on Graph Theory and Social Networks,* MIT OpenCourseWare.

del Val, C., Ernst, P., Falkenhahn, M., Fladerer, C., Glatting, K.H., Suhai, S., Hotz-Wagenblatt, A. (2007). *ProtSweep, 2Dsweep and DomainSweep: protein analysis suite at DKFZ.* Nucleic Acids Res. 35 (Web Server issue), W444–W450.

de Menezes, M.A. and Barabási, A.L. (2004). *Fluctuations in network dynamics.* Phys Rev Lett. Vol. 92, pg. 028701.

Deng, M., Sun, F., and Chen, T. (2003). *Assessment of the reliability of protein-protein interactions and protein function prediction.* Proceeding PSB. Vol. 8, pg. 140-151.

Derrida, B. and Flyvbjerg, H. (1987). *Statistical properties of randomly broken objects and of multivalley structures in disordered-systems.* J. Phys. A: Math Gen. Vol.20, pg. 5273-88.

Dezso, Z., Oltvai, Z.N. and Barabási, A.L. (2003). *Bioinformatics analysis of experimentally determined protein complexes in the yeast, Saccharomyces cerevisiae.* Genome Res. Vol. 13, pg. 2450-4.

Doering, C. D. (1997). *Signal transduction in malaria parasites.* Parasitology Today, Vol. 13(8), pg.307- 313.

Dorogovtsev, S.N., Goltsev, A.V. and Mendes, J.F.F. (2002). *Pseudofractal scalefree web.* Phys Rev E. Vol. 65, pg. 066122.

Dorogovtsev, S.N. and Mendes, J.F.F. (2003). *Evolution of networks : From biological nets to the Internet and WWW.* Oxford University Press, Oxford.

Douglas J. LaCount, Marissa Vignali, Rakesh Chettier, Amit Phansalkar, Russell Bell, Jay R. Hesselberth, Lori W. Schoenfeld, Irene Ota, Sudhir Sahasrabudhe, Cornelia Kurschner, Stanley Fields and Robert E. Hughes. (2005). *A protein interaction network of the malaria parasite Plasmodium falciparum.* Nature, Vol. 438, pg. 103-107.

Dunn, R., Dudbridge, F. and Sanderson, C.M. (2005). *The use of edge-betweenness*

Clustering to investigate biological function in protein interaction networks. BMC Bioinformatics, Vol. 6, pg. 39.

Dyer MD, Murali TM, Sobral BW (2007). *Computational prediction of host-pathogen protein interactions,* Bioinformatics, Vol. 23, pg. 159-166.

Ebenhoh, O. Handorf, T. and Heinrich, R. (2004). *Structural analysis of expanding metabolic networks.* Genome Informatics, Vol. 15(1), pg. 35-45.

Edwards, J.S. and Palsson, B.O. (2000). *The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities.* Proc Natl Acad Sci.Vol. 97, pg. 5528-33.

Edwards, J.S., Ibarra, R.U. and Palsson, B.O. (2001). *In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data.* Nat Biotechnol. Vol. 19, pg. 125-30.

Emmerling, M., Dauner, M., Ponti, A., Fiaux, J., Hochuli, M., Szyperski, T., Wuthrich, K., Bailey, J.E. and Sauer, U. (2002). *Metabolic flux responses to pyruvate kinase knockout in Escherichia coli.* J Bacteriol. Vol. 184, pg. 152-64.

Eppstein, D. and Wang, J.(2002). *A steady state model for graph power laws.*

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

Estrada E.(2006). *Virtual identification of essential proteins within the protein interaction network of yeast.* Proteomics, Vol. 6, pg. 35–40.

Farkas, I.J., Jeong, H., Vicsek, T., Barabási, A.L. and Oltvai, Z.N. (2003). *The topology of the transcription regulatory network in the yeast, Saccharomyces cerevisiae.* Physica A. Vol. 318, pg. 601-12.

Frischknecht, F.,Lanzer, M. (2008). *The Plasmodium falciparum Maurer's clefts in 3D.* Mol. Microbiol. Vol.6(4), pg. 687–691.

Gallup, J.L. and Sachs, J.D.(2001). *The economic burden of malaria.* Am J. Trop. Med. Hyg., Vol. 64, pg. 85–96.

Gangman Y., Sing, H. S. and Michael, R. T. (2007). *Identifying clusters of functionally related genes in genomes.* BMC Vol. 23(9) pg. 1053-1060.

Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S, Paulsen IT, James K, Eisen JA, Rutherford K, Salzberg SL, Craig A, Kyes S, Chan MS, Nene V, Shallom SJ, Suh B, Peterson J, Angiuoli S, Pertea M, Allen J, Selengut J, Haft D, Mather MW, Vaidya AB, Martin DM, Fairlamb AH, Fraunholz MJ, Roos DS, Ralph SA, McFadden GI, Cummings LM, Subramanian GM, Mungall C, Venter JC, Carucci DJ, Hoffman SL, Newbold C, Davis RW, Fraser CM, Barrell B. (2002). *Genome sequence of the human malaria parasite Plasmodium falciparum.* Nature, Vol. 419(6906), pg. 498-511.

Ge, H., Liu, Z., Church, G.M. and Vidal, M. (2001). *Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae.* Nature Genet. Vol.29, pg.482-6.

Ginsburg, H.(2006). *Progress in in silico functional genomics: the malaria Metabolic Pathways database.* Trend Parasitol. Vol. 22, pg. 238-240.

Goeman, J. J. and Peter B. U.(2006). *Analyzing gene expression data in terms of gene sets:* Methodological issues Vol. 0, pg. 1–7.

Goh, K.-I., Kahng, B. and Kim, D. (2002). *Fluctuation-driven dynamics of the internet topology.* Phys Rev Lett. Vol. 88, pg.108701.

Goldberg, D. and Roth, F. (2003). *Assessing experimentally derived interactions in a small world.* Proceeding National Academy of Science USA Vol. 100, pg. 4372-4376.

Green, ML and Karp (2007). *PD Using genome-content data to identify specific types of functional associations in pathway/genome databases.* Bioinf. Vol.23, i205-i211.

Grigoriev, A. (2001). *A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the S. cerevisiae.* Nucleic Acids Res. Vol. 29, pg. 3513-3519.

Grindrod P, Kibble M.(2004). *Review of uses of network and graph theory concepts within proteomics.* Expert Rev Proteomics Vol.1, pg. 229–38.

Grogoriev, A. (2001). *A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and yeast Saccharomyces cerevisiae.* Nucleic Acids Res. Vol. 29, pg. 3513-9.

Guimer`a, R. and L.A. Nunes, L. A.(2005). *Functional cartography of complex metabolic networks.* Nature, Vol. 433(7028), pg. 895–900.

Guo X, Liu R, Shriver CD, *et al*.(2006). *Assessing semantic similarity measures for the characterization of human regulatory pathways.* Bioinformatics, Vol. 22, pg. 967–73.

Hanks S. K and Hunter T. (1995). *Protein kinase 6. The eukaryotic protein kinase*

superfamily: kinase (catalytic) domain structure and classification. The Journal of the Federation of American Societies for Experimental Biology, Vol. 9, pg. 576-596

Han, J.-D.J., Dupuy, D., Bertin, N., Cusick, M.E. and Vidal, M. (2005). *Effect of sampling on topology predictions of protein-protein interaction networks.* Nat. Biotechnol., Vol. 23, pg. 839–44.

Harris TW, Lee R, Schwarz E, Bradnam K, Lawson D, Chen W, Blasier D, Kenny E, Cunningham F, Kishore R, Chan J, Muller HM, Petcherski A, Thorisson G, Day A, Bieri T, Rogers A, Chen CK, Spieth J, Sternberg P, Durbin R, Stein LD. (2003). *WormBase: a cross-species database for comparative genomics.* Nucleic Acids Res. Vol.31(1), pg.133-7.

Hartwell LH, Hopfiled JJ, Leibler S, and Murray A. (1999). *From molecular to modular cell biology.* Nature, Vol. 402(6761), pg.47–52.

Heath A., Bennett G., and Kavraki L.(2010). *Finding metabolic pathways using atom tracking.* Bioinformatics,Vol. 26(12), pg. 1548-1555.

Hubbard MJ and Cohen P. (1993). *On target with a new mechanism for the regulation of protein phosphorylatio.* Trends in Biochemical Science; Vol. 18, pg.172-177

Hunter and Borg(2003). *Integration from proteins to organs:* The physiome project, Nature Molecular Cell Biology.

Ibarra, R.U., Edwards, J.S. and Palsson, B.O. (2002). *Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth.* Nature. Vol. 420, Vol. 186-9.

Ideker T, Ozier O, Schwikowski B, and Siegel AF (2002). *Discovering regulatory and signalling circuits in molecular interaction networks.* Bioinformatics, Vol. 18(1), pg. 233–40.

Jan Porekar (2002). *Random Networks.*

Jansen, R., Greenbaum, D. and Gerstein, M. (2002). *Relating whole-genome expression data with protein-protein interactions.* Genome Res. Vol.12, pg. 37-46.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N. and Barabási, A.L. (2000). *The large-scale organization of metabolic networks.* Nature. Vol. 407,pg. 651-4.

Jeong, H., Mason, S.P., Barabási, A.L. and Oltvai, Z.N. (2001). *Lethality and centrality in protein networks.* Nature. Vol. 411, pg. 41-2.

Jiang, D., Pei, J. and Zhang, A.(2003) *Interactive Exploration of Coherent :Patterns in Time-series Gene Expression Data.* In Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, pg. 24-27.

Jiang, D., Pei, J. and Zhang, A.(2003). *DHC: A Density-based Hierachical Clustering Method for Time-series Gene Expression Data*. In Proceeding of BIBE2003: 3RD IEE International Symposium on Bioinformatics and Bioengineering, Bethesda, Maryland, March 10-12.

John, David T., and William A. Petri(2006). *J. Markell and Voge's Medical Parasitology.* 9th ed. St. Louis: Saunders Elsevier.

Joyce A.R, Palsson B.O.(2006). *The model organism as a system: integrating 'omics' data sets.* Nat. Rev. Mol. Cell Biol. Vol. 7, pg. 198–210.

Kanehisa M.(2002). T*he KEGG database.* Novartis Found Symp. 247:91-101, discussion 101-3, 119- 28, 244-52.

Karp D, Ouzounis A., Moore-Kochlacs C., Goldovsky L., Kaipa P., Ahrén D., Tsoka S, Darzentas N., Kunin V, and López-Bigas N (2005). *Expansion of the BioCyc collection of pathway/genome databases to 160 genomes.* Nucleic Acids Res. Vol. 33, pg. 6083-6089.

King, A.D., Pr·zulj, N., Jurisica, I., (2004). *Protein complex prediction via cost-based clustering.* Bioinformatics, Vol. 20(17), pg. 3013-3020.

Kitano H.(2002). *Systems biology:* A brief overview. Science, Vol. 295 pg.1662–1664.

Koenig R, Schramm G, Oswald M, Seitz H, Sager S, Zapatka M, Reinelt G, Eils R. (2006). *Discovering functional gene expression patterns in the metabolic network of Escherichia coli with wavelets transforms.* BMC Bioinformatics. Vol.7, pg.119.

Koyama, F. C., Chakrabarti, D. and Garcia, C. R. S. (2009). *Molecular machinery of signal transduction and cell cycle regulation in Plasmodium.* Mol & Biochem. Parasitology, Vol.165 pg. 1-7.

Krogstad, D.J., Gluzman, I.Y., Kyle, D.E., Odunola, A.M.J., Martin, S.K., Milhous, W.K., Schlesinger, P.H. (1987). *Efflux of chloroquine from Plasmodium falciparum: mechanism of chloroquine resistance.* Science Vol. 238, pg.1283–1285.

Krogstad, D.J., Schlesinger, P.H., Herwaldt, B.L. (1988). *Antimalarial agents: mechanism of chloroquine resistance.* Antimicrob. Agents Chemother.799–801.

Kutznetsov, V.A., Knott, G.D. and Bonner, R.F. (2002). *General statistics of stochastic processes of gene expression in eukaryotic cells.* Genetics. Vol. 161, pg. 1321-32.

LaCount, D.J.,Marissa,V.,Rakesh,C.,Amit,P.,Russell,B.,Jay,R.,Hesselberth,Lori,W., Schoenfeld, Irene,O.,Sudhir,S.,Cornelia,K.,Stanley,F.,Robert,E.,Hughes (2005). *A protein interaction network of the malaria parasite Plasmodium falciparum.* Nature Vol. 438, pg. 103–107.

Lappe M, Holm L(2004). *Unraveling protein interaction networks with near-optimal efficiency.* Nat Biotechnol., Vol. 22, pg. 98–103.

Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, Zeitlinger J, Jennings EG, Murray HL, Gordon DB, Ren B, Wyrick JJ, Tagne JB, Volkert TL, Fraenkel E, Gifford DK, Young RA.(2002). *Transcriptional regulatory networks in Saccharomyces cerevisiae.* Science, Vol.298, pg. 799–804.

LeRoch K.G., Zhou Y., Blair P. I., Grainger M., Moch J. K., *et al* (2003). *Discovery of gene function by expression profiling of the malaria parasite life cycle.* Science, Vol. 301, pg. 1503-1508.

Lu, H., Zhu, X., Liu, H., Skogerbo, G., Zhang, J., Zhang, Y., Cai, L., Zhao, Y., Sun, S., Xu, J., Bu, D., and Chen, R. (2004). *The interactome as a tree-an attempt to visualize the protein-protein interaction network in yeast.* Nucleic Acids Res., Vol. 32(16), pg. 4804-4811.

Ma H. W., Zhao X. M., Yuan Y. J., and Zeng A. P. (2004). *Decomposition of metabolic network into functional modules based on the global connectivity structure of reaction graph.* Bioinformatics, Vol. 20(12, pg. 1870–1876.

Miller, L.H., Baruch, D.I., Marsh, K., Doumbo, O.K. (2002). *The pathogenic basis of malaria.* Nature Vol. 415, pg. 673–679.

Milo R, Shen-Orr S, Itzkovitz S, Kashtan N., Chklovskii D., and Alon U.(2002). *Network motifs: simple building blocks of complex networks.* Science, Vol.298, pg. 824–7.

Nunes MC, Goldring D, Doerig C, Scherfl A. (2007). *A novel protein kinase family in Plasmodium falciparum is differentially transcribed and secreted to various cellular compartments of the host cell.* Molecular Microbiology, Vol. 63, pg. 391-403.

Othmer, Hans G. (2006). *Lecture note on Analysis of Complex Reaction Networks in Signal Transduction, Gene Control and Metabolism.* School of Mathematics, University of Minnesota Minneapolis, MN.

Oyelade, J. O., Adebiyi, E. F., Yah, S. C. and Olaseinde, G. I. (2008). *Computational Identification of functional related gene in malaria parasites.* Poster proceeding, ISMB.

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

Palmer, M. (2006). *Chem333: Metabolism Lecture Notes.* Unpublished lecture notes.

Palumbo MC, Colosimo A, Giuliani A, *et al*.(2005). *Functional essentiality from topology features in metabolic networks:* A case study in yeast. FEBS Letters, Vol. 579, pg. 4642–6.

Papin J. A., Reed J. L., and Palsson B. O. (2004). *Hierarchical thinking in network biology: the unbiased modularization of biochemical networks.* Trends Biochem. Sci., Vol. 29(12), pg. 641–647.

Papin J.A., Stelling J., Price N. D., Klamt S., Schuster S., and Palsson B.O. (2004). *Comparison of network-based pathway analysis methods.* Trends Biotechnol., Vol. 22(8), pg. 400–405.

Papin J.A., Hunter T., Palsson B.O, *et al*.(2005). *Reconstruction of cellular signalling networks and analysis of their properties.* Nat. Rev. Mol. Cell Biol., Vol. 6, pg.99–111.

Per Kraulis (2003). *Lecture notes: KTH Bioinformatics, Stockholm* Bioinformatics Center.

Pereira-Leal, J.B., Enright, A.J., Ouzounis, C.A. (2004). *Detection of functional modules from protein interaction networks.* Proteins, Vol. 54, pg. 49-57.

Pitkaenen, E., Jouhten P., and Rousu J.(2009). *Inferring branching pathways in genome-scale metabolic networks.* BMC Systems Biology, Vol.3, pg.103.

Pinney,J., Shirley, M., McConkey,G., and Westhead,D. (2005). *metaSHARK: software for automated metabolic network prediction from DNA sequence and its application to the genomes of Plasmodium falciparum and Eimeria tenella.* NAR, Vol. 33, pg. 1399-1409.

Pinney J., Papp B., Hyland C., Wambua L., Westhead D., and McConkey G. (2007). *Metabolic reconstruction and analysis for parasite genomes.*Trends Parasitol. Vol. 23, pg. 548-554.

Planes, F.J. and Beasley, J.E.(2009). *Path finding approaches and metabolic pathways.* Discrete Applied Mathematics, Vol. 157, pg. 2244-2256.

Przulj N, Corneil D.G., Jurisica I.(2004). *Modeling interactome: scale-free or geometric?* Bioinformatics, Vol. 20, pg. 3508–15.

Przulj N, Wigle D. A., Jurisica I.(2004). *Functional topology in a network of protein interactions.* Bioinformatics, Vol. 20, pg. 340–8.

Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N. and Barabási, A.L. (2002). *Hierarchical organization of modularity in metabolic networks.* Science. Vol. 297, pg. 1551-5.

Ravasz, E. and Barabási, A.L. (2003). *Hierarchical organization in complex networks.* Phys Rev E. Vol. 67, pg. 026112.

Rives, A.W. and Galitski, T. (2003). *Modular organization of cellular networks.* Proc. Natl Acad. Sci., USA, Vol. 100, pg. 1128-1133.

Ron Caspi, Hartmut Foerster, Carol A. Fulcher, Pallavi Kaipa, Markus Krummenacker, Mario Latendresse, Suzanne Paley, Seung Y. Rhee, Alexander G. Shearer, Christophe Tissier, Thomas C. Walk, Peifen Zhang, and Peter D. Karp (2008). T*he MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases.* NAR, 36 (Database issue).

Rusell S. J. and Norvig P. (2003). *Artificial Intelligence: A Modern Approach.* Prentice Hall, 2nd edition.

Samal A, Singh S, Giri V, *et al*.(2006). *Low degree metabolites explain essential reactions and enhance modularity in biological networks.* BMC Bioinformatics, Vol.7, pg. 118.

Shivaram Narayanan(2005). *The betweenness centrality of biological networks,* M.Sc. Thesis, Virginia Polytechnic Institute and State University.

Schneider, A.G., Mercereau-Puijalon, O. (2005). *A new Apicomplexa-specific protein kinase family: multiple members in Plasmodium falciparum, all with an export signature.* BMC Genomics Vol. 6(1), pg. 30.

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, Vol. 18(2), pg. 351–361.

Scott J, Ideker T, Karp RM, Sharan R.(2006). *Efficient algorithms for detecting signaling pathways in protein interaction networks.* J Comput Biol. Vol. 13(2), pg.133-44.

Segre, D., Vitkup, D. and Church, G.M. (2002). *Analysis of optimality in natural and perturbed metabolic networks.* Proc Natl Acad Sci.Vol. 99, pg. 15112-7.

Sharan, R., Suthram, S., Kelley, R.M., Kuhn, T., McCuine, S., Sittler, T., Karp, R.M., Ideker, T. (2005). *Conserved patterns of protein interaction in multiple species.* PNAS Vol. 102(6), pg. 1974–1979.

Shen-Orr S. S., Milo R., Mangan S., and Alon U. (2002). *Network motifs in the transcriptional regulation network of Escherichia coli.* Nat. Genet., Vol. 31(1), pg.64–68.

Shlomi, T., Segal, D., Ruppin, E., and Sharan, R. (2006). *QPath: a method for querying pathways in a protein–protein interaction network.* BMC Bioinformatics, Vol. 7, pg. 199.

Soldati, D., Foth, B.J., Cowman, A.F. (2004). *Molecular and functional aspects of parasite invasion.* Trends Parasitol. Vol. 20, pg. 567–574.

Spirin, V. and Mirny, L.A. (2003). *Protein complexes and functional modules in molecular networks.* Proc.Natl Acad. Sci., USA, Vol. 100(21), pg. 12123-12126.

Steffen, M., Petti, A., Aach, J., D'haeseleer, P., Church, G.(2002). *Automated modelling of signal transduction networks.* BMC Bioinformatics, Vol.3, pg. 34–44.

Stumpf M. P., Wiuf C., May R. M.(2005). *Subnets of scale-free networks are not scale-free: sampling properties of networks.* Proc Natl Acad Sci USA Vol. 102, pg. 4221–4.

Suetterlin BW, Kappes B, Franklin RM (1991). *Localisation and stage specific phosphorylation of Plasmodium falciparum phosphoproteins during the intraerythrocytic cycle*. Molecular and Biochemical Parasitology; Vol. 46, pg. 113-122.

Tero Aittokallio and Benno Schwikowski (2006). *Graph-based methods for analyzing networks in cell biology.* Briefings in Bioinformatics. Vol. 7(3), pg. 243-255.

Vaid, A., Sharma, P. (2006). *PfPKB, a protein kinaseB-like enzyme from Plasmodium falciparum. II. Identification of calcium/calmodulin as its upstream activator and dissection of a novel signalling pathway.* J. Biol. Chem. Vol. 281(37), pg. 27126–27133.

Vaid, A.,Thomas,D.C.,Sharma,P.(2008). *RoleofCa2+/Calmodulin-PfPKB signalling pathway inerthrocycteinvasionby Plasmodium falciparum.* J.Biol.Chem. Vol. 283(9), pg. 5589–5597.

Vázquez, A., Pastor-Satorras, R. and Vespignani, A. (2002). *Large-scale topological and dynamical properties of the Internet.* Phys Rev E. Vol. 65, pg. 066130.

Vazquez A, Dobrin R, Sergi D, Eckmann J-P, Oltvai ZN and Barabasi A-L (2004). *The topological relationship between the large-scale attributes and local interaction patterns of complex networks.* Proc. Natl. Acad. Sci. USA., Vol. 101, pg. 17940–5.

Von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S., Fields, S., and Bork, P. (2002). Nature, Vol. 417, pg. 399-403.

Ward, P., Equinet, L., Packer, J., Doerig, C. (2004). *Protein kinases of the human malaria parasite Plasmodium falciparum: the kinome of a divergent eukaryote.* BMC Genomics Vol.5, pg. 79.

Watts, D.J. and Strogatz, S.H. (1998). *Collective dynamics of small-world networks.* Nature. Vol. 393, pg. 440-2.

Watts, D. J. and Strogatz, S. H(1998). *Collective dynamics of 'small-world' networks.* Nature Vol. 393(6684), pg.440-442.

Wolf D.M. and Arkin A.P. (2003). *Motifs, modules and games in bacteria.* Curr. Opin. Microbiol., 6(2):125–134, Apr 2003.

Yeger-Lotem E, Sattath S, Kashtan N, Itzkovitz S, Milo R, Pinter RY, Alon U, Margalit H (2004). *Network motifs in integrated cellular networks of transcription-regulation and protein–protein interaction.* Proc. Natl. Acad. Sci. U. S. A. Vol.101, pg. 5934–5939

Zhou, Y., Young, J. A., Santrosyan, A., Chen, K., Frank Yan, S., Winzeler, E. (2005). *In-silico gene function prediction using ontology-based pattern identification.* Bioinformatics, Vol. 21(7), pg. 1237-1245.

Zhou X, Kao MC, Wong W.H.(2002). *Transitive functional annotation by shortest-path analysis of gene expression data.* Proc. Natl. Acad. Sci. USA., Vol. 99, pg. 12783–8.

http://www.dpd.ede.gov/dpdx , 2011.

http://www.plasmodb.org, 2011.

http://sites.huji.ac.il/malaria/, 2011.

http://www.biology-online.org, 2011.

http://en.wikipedia.org/wiki/centrlity, 2011.

http://en.wikipedia.org/wiki/protein, 2011.

APPENDIX A

Table 4.2 Predicted minimum pathways (potential signal transduction pathways) for the calcium modulated proteins: Column one indicates the name of the signalling pathway, the second column shows minimum paths extracted, while optimizing the identified number of proteins in the pathway under consideration. The third column shows the weight p-value and column four detailed the products (fromplasmodb) of the proteins in the identified potential signalling pathways.

Name	Minimum path	p-value	Details of genes	
			Genes IDS	Products
Calcium Modulated	PFA0110w---> PFB0540w---> FB0815w---> FF0220w---> PFF0590c	0.029	PFA0110w PFB0540w PFB0815w PFF0220w PFF0590c	DNAJ protein, putative conserved Plasmodium protein, unknown function Calcium-dependent protein kinase 1 conserved Plasmodium protein, unknown function homologue of human HSPC025
	PFB0540w --->PFB0815w --->PFF0220w--->PFF0590c ---> PF10_0194	0.044	PFB0540w PFB0815w PFF0220w PFF0590c PF10_0194	conserved Plasmodium protein, unknown function Calcium-dependent protein kinase 1 conserved Plasmodium protein, unknown function homologue of human HSPC025 NOP12-like protein
	PFB0540w --->PFB0815w ---> PFF0675c---> PF10_0345---> PF11_0111	**0.044**	**PFB0540w PFB0815w PFF0675c PF10_0345 PF11_0111**	**conserved Plasmodium protein, unknown function Calcium-dependent protein kinase 1 myosin E merozoite surface protein 3 asparagine-rich antigen**
	PFB0540w --->PFB0815w ---> PF11_0111	0.013	PFB0540w PFB0815w PF11_0111	conserved Plasmodium protein, unknown function Calcium-dependent protein kinase 1 asparagine-rich antigen
	PFB0540w --->PFB0815w ---> PFF1365c--->PF11_0241	0.027	PFB0540w PFB0815w PFF1365c PF11_0241	conserved Plasmodium protein, unknown function Calcium-dependent protein kinase 1 HECT-domain (ubiquitin-transferase), putative Myb-like DNA-binding domain, putative
	PFB0540w --->PFB0815w ---> PFF0220w---> PFF0590c --->PFL0305c	0.035	PFB0540w PFB0815w PFF0220w PFF0590c PFL0305c	conserved Plasmodium protein, unknown function Calcium-dependent protein kinase 1 conserved Plasmodium protein, unknown function homologue of human HSPC025 IMP-specific 5'-nucleotidase
	PFB0540w --->PFB0815w ---> PFF0220w--->PFF0590c ---> PFL1385c	**0.024**	**PFB0540w PFB0815w PFF0220w PFF0590c PFL1385c**	**conserved Plasmodium protein, unknown function Calcium-dependent protein kinase 1 conserved Plasmodium protein, unknown function homologue of human HSPC025 Merozoite Surface Protein 9, MSP-9**
	PFB0815w ---> PFL1795c	0.009	PFB0815w PFL1795c	Calcium-dependent protein kinase 1 conserved Plasmodium protein, unknown function
	PFB0540w --->PFB0815w ---> PFF1365c--->PF13_0139	0.029	PFB0540w PFB0815w PFF1365c PF13_0139	conserved Plasmodium protein, unknown function Calcium-dependent protein kinase 1 HECT-domain (ubiquitin-transferase), putative conserved Plasmodium protein, unknown function
	PFB0540w --->PFB0815w ---> PFD0985w---> PFF0590c--->PF14_0029	0.044	PFB0540w PFB0815w PFD0985w PFF0590c PF14_0029	conserved Plasmodium protein, unknown function Calcium-dependent protein kinase 1 transcription factor with AP2 domain(s), putative homologue of human HSPC025 conserved Plasmodium protein, unknown function
	PFB0815w ---> PF14_0170	0.002	PFB0815w PF14_0170	Calcium-dependent protein kinase 1 NOT family protein, putative
	PFB0815w ---> PFF0505c ---> PF14_010	0.02	PFB0815w PFF0505c PF14_010	Calcium-dependent protein kinase 1 conserved Plasmodium protein, unknown function p23 co-chaperone, putative

	PFB0540w --->PFB0815w ---> PFF0220w---> PFF0590c ---> PF14_0632	**0.044**	**PFB0540w PFB0815w PFF0220w PFF0590c PF14_0632**	**conserved Plasmodium protein, unknown function Calcium-dependent protein kinase 1 t conserved Plasmodium protein, unknown function homologue of human HSPC025 26S proteasome subunit, putative**
	PFB0540w --->PFB0815w	0.002	PFB0540w PFB0815w	Calcium-dependent protein kinase 1 conserved Plasmodium protein, unknown function
	PFA0110w---> PFB0540w---> FB0815w---> FD0090c FF0220w---> PFF0590c	**0.049**	**PFA0110w PFB0540w PFB0815w PFD0090c PFF0220w PFF0590c**	**DNAJ protein, putative conserved Plasmodium protein, unknown function Calcium-dependent protein kinase 1 Plasmodium exported protein (PHISTa), unknown function conserved Plasmodium protein, unknown function homologue of human HSPC025**
	PFB0815w ---> PFD0795w---> PFF1440w	0.016	PFB0815w PFD0795w PFF1440w	Calcium-dependent protein kinase 1 conserved Plasmodium protein, unknown function SET domain protein, putative
	PFB0540w --->PFB0815w ---> PFD0985w--->PFF0590c	0.024	PFB0540w PFB0815w PFD0985w PFF0590c	conserved Plasmodium protein, unknown function Calcium-dependent protein kinase 1 transcription factor with AP2 domain(s), putative homologue of human HSPC025
	PFB0540w --->PFB0815w ---> PFE0070w---> PFF0675c --->PF11_0111	**0.044**	**PFB0540w PFB0815w PFE0070w PFF0675c PF11_0111**	**conserved Plasmodium protein, unknown function Calcium-dependent protein kinase 1 interspersed repeat antigen, putative myosin E asparagine-rich antigen**
	PFB0540w --->PFB0815w ---> PFF0220w--->PFF0590c	0.024	PFB0540w PFB0815w PFF0220w PFF0590c	conserved Plasmodium protein, unknown function Calcium-dependent protein kinase 1 conserved Plasmodium protein, unknown function homologue of human HSPC025
	PFB0815w ---> PFF0505c	0.004	PFB0815w PFF0505c	Calcium-dependent protein kinase 1 conserved Plasmodium protein, unknown function
	PFB0540w --->PFB0815w ---> PFF0590c	0.013	PFB0540w PFB0815w PFF0590c	conserved Plasmodium protein, unknown function Calcium-dependent protein kinase 1 homologue of human HSPC025
	PFB0540w --->PFB0815w ---> PFF0675c--->PF11_0111	0.029	PFB0540w PFB0815w PFF0675c PF11_0111	conserved Plasmodium protein, unknown function Calcium-dependent protein kinase 1 myosin E asparagine-rich antigen
	PFB0540w --->PFB0815w ---> PFD0985w---> PFF0590c -->PFF0785w	**0.044**	**PFB0540w PFB0815w PFD0985w PFF0590c PFF0785w**	**conserved Plasmodium protein, unknown function Calcium-dependent protein kinase 1 transcription factor with AP2 domain(s), putative homologue of human HSPC025 Ndc80 homologue, putative**
	PFB0540w --->PFB0815w ---> PFF1365c	0.016	PFB0540w PFB0815w PFF1365c	conserved Plasmodium protein, unknown function Calcium-dependent protein kinase 1 HECT-domain (ubiquitin-transferase), putative
	PFB0815w ---> PFF1440w	0.002	PFB0815w PFF0505c PFF1440w	Calcium-dependent protein kinase 1 conserved Plasmodium protein, unknown function SET domain protein, putative
	PFB0540w --->PFB0815w ---> PFF1365c---> PFF1470c	0.029	PFB0540w PFB0815w PFF1365c PFF1470c	conserved Plasmodium protein, unknown function Calcium-dependent protein kinase 1 HECT-domain (ubiquitin-transferase), putative DNA polymerase epsilon, catalytic subunit a, putative
	PFB0540w --->PFB0815w ---> PFF1365c---> MAL7P1.12	**0.029**	**PFB0540w PFB0815w PFF1365c MAL7P1.12**	**conserved Plasmodium protein, unknown function Calcium-dependent protein kinase 1 HECT-domain (ubiquitin-transferase), putative erythrocyte membrane-associated antigen**
	PFB0540w --->PFB0815w ---> PF07_0053---> PF11_0111	0.027	PFB0540w PFB0815w PF07_0053 PF11_0111	conserved Plasmodium protein, unknown function Calcium-dependent protein kinase 1 conserved Plasmodium protein, unknown function asparagine-rich antigen

Table 4.3 Predicted minimum pathways (potential signal transduction pathways) for the calcium signalling proteins: Column one indicates the name of the signalling pathway, the second column shows minimum paths extracted, while optimizing the identified number of proteins in the pathway under consideration. The third column shows the weight p-value and column four detailed the products (from plasmodb) of the proteins in the identified potential signalling pathways.

Name	Minimum path	p-value	Details of genes	
			Genes IDS	Products
Calcium Signaling	PFA0515w --->PF11_0239 ---> PF14_0252	0.021	PFA0515w PF11_0239 PF14_0252	phosphatidylinositol-4-phosphate-5-kinase calcium-dependent protein kinase, putative conserved Plasmodium protein, unknown function
	PF10_0143 --->PF11_0142 ---> PF11_0239 --->MAL13P1.206	**0.033**	**PF10_0143 PF11_0142 PF11_0239 MAL13P1.206**	**transcriptional activator ADA2, putative ubiquitin domain containing protein calcium-dependent protein kinase, putative Na+ -dependent Pi transporter, sodium-dependent phosphate transporter**
	PF11_0142 ---> PF11_0239	0.002	PF11_0142 PF11_0239	ubiquitin domain containing protein calcium-dependent protein kinase, putative
	PF11_0239--->PFL0280c	0.013	PF11_0239 PFL0280c	calcium-dependent protein kinase, putative histone binding protein, putative
	PF11_0239--->PFL1745c--->PF14_0637	0.019	PF11_0239 PFL1745c PF14_0637	calcium-dependent protein kinase, putative clustered-asparagine-rich protein rhoptry protein, putative
	PF11_0239--->PFL2420w	0.002	PF11_0239 PFL2420w	calcium-dependent protein kinase, putative conserved Plasmodium protein, unknown function
	PF11_0142 ---> PF11_0239---> PF13_0197---> MAL13P1.206	**0.038**	**PF11_0142 PF11_0239 PF13_0197 MAL13P1.206**	**ubiquitin domain containing protein calcium-dependent protein kinase, putative Merozoite Surface Protein 7 precursor, MSP7 Na+ -dependent Pi transporter, sodium-dependent phosphate transporter**
	PF08_0048 ---> PF11_0239---> PF13_0219	0.024	PF08_0048 PF11_0239 PF13_0219	ATP-dependent helicase, putative calcium-dependent protein kinase, putative conserved Plasmodium protein, unknown function
	PF11_0142 ---> PF11_0239---> MAL13P1.206	**0.022**	**PF11_0142 PF11_0239 MAL13P1.206**	**ubiquitin domain containing protein calcium-dependent protein kinase, putative Na+ -dependent Pi transporter, sodium-dependent phosphate transporter**
	PF11_0142 ---> PF11_0239---> MAL13P1.206---> PF14_0059	**0.046**	**PF11_0142 PF11_0239 MAL13P1.206 PF14_0059**	**ubiquitin domain containing protein calcium-dependent protein kinase, putative Na+ -dependent Pi transporter, sodium-dependent phosphate transporter conserved Plasmodium protein, unknown function**
	PF11_0239--->PF14_0252	0.002	PF11_0239 PF14_0252	calcium-dependent protein kinase, putative conserved Plasmodium protein, unknown function
	PF08_0048 ---> PF11_0239---> PF14_0391	0.029	PF08_0048 PF11_0239 PF14_0391	ATP-dependent helicase, putative calcium-dependent protein kinase, putative 60S ribosomal protein L1, putative
	PF11_0239--->PF14_0500	0.005	PF11_0239 PF14_0500	calcium-dependent protein kinase, putative SNARE protein, putative
	PF11_0239--->PF14_0637	0.002	PF11_0239 PF14_0637	calcium-dependent protein kinase, putative rhoptry protein, putative
	PF11_0142 ---> PF11_0239---> MAL13P1.206---> PF14_0678	**0.033**	**PF11_0142 PF11_0239 MAL13P1.206 PF14_0678**	**ubiquitin domain containing protein calcium-dependent protein kinase, putative Na+ -dependent Pi transporter, sodium-dependent phosphate transporter exported protein 2**
	PFB0915w--->PF11_0239	0.011	PFB0915w PF11_0239	liver stage antigen 3 calcium-dependent protein kinase, putative
	PFC0235w--->PF11_0239	0.011	PFC0235w PF11_0239	conserved Plasmodium protein, unknown function calcium-dependent protein kinase, putative
	PFD0090c---> PF08_0048 ---> PF11_0239	0.019	PFD0090c PF08_0048 PF11_0239	Plasmodium exported protein (PHISTa), unknown function ATP-dependent helicase, putative

				calcium-dependent protein kinase, putative
	PFD0090c---> PFF0670w---> PF08_0048 ---> PF11_0239	**0.041**	**PFD0090c PFF0670w PF08_0048 PF11_0239**	**Plasmodium exported protein (PHISTa), unknown function transcription factor with AP2 domain(s), putative ATP-dependent helicase, putative calcium-dependent protein kinase, putative**
	PFF0785w--->PF11_0239	0.006	PFF0785w PF11_0239	Ndc80 homologue, putative calcium-dependent protein kinase, putative
	PFF1185w---> PF11_0142 ---> PF11_0239---> MAL13P1.206	0.037	PFF1185w PF11_0142 PF11_0239 MAL13P1.206	Smarca -related protein ubiquitin domain containing protein calcium-dependent protein kinase, putative Na+ -dependent Pi transporter, sodium-dependent phosphate transporter
	PFF1395c---> PF08_0048 ---> PF11_0239	0.024	PFF1395c PF08_0048 PF11_0239	Glutamyl-tRNA(Gln) amidotransferase subunit B, putatative ATP-dependent helicase, putative calcium-dependent protein kinase, putative
	PF08_0048 --->PF11_0239	0.006	PF08_0048 PF11_0239	ATP-dependent helicase, putative calcium-dependent protein kinase, putative

Table 4.4 Predicted minimum pathways (potential signal transduction pathways) for the cell cycle kinases proteins: Column one indicates the name of the signalling pathway, the second column shows minimum paths extracted, while optimizing the identified number of proteins in the pathway under consideration. The third column shows the weight p-value and column four detailed the products (from plasmodb) of the proteins in the identified potential signalling pathways.

Name	**Minimum path**	**p-value**	**Details of genes**	
			Genes IDS	Products
Cell Cycle	**PFE1370w ---> PF10_0143 ---> PF10_0272**	**0.021**	**PFE1370w PF10_0143 PF10_0272**	**hsp70 interacting protein, putative transcriptional activator ADA2, putative 60S ribosomal protein L3, putative**
	PF10_0254 --->PF11_0272	0.005	PF10_0254 PF11_0272	Conserved plasmodium protein, unknown function 60S ribosomal protein L3, putative
	PF10_0272--->PFL0185c	0.009	PF10_0272 PFL0185c	60S ribosomal protein L3, putative Nucleosome assembly protein 1, putative
	PF10_0254--->PF10_0272--->PFL1385c	**0.021**	**PF10_0254 PF10_0272 PFL1385c**	**Conserved plasmodium protein, unknown function 60S ribosomal protein L3, putative Merozoite Surface protein 9, MSP-9**
	PF10_0254--->PF10_0272--->PFL1845c	0.016	PF10_0254 PF10_0272 PFL1845c	Conserved plasmodium protein, unknown function 60S ribosomal protein L3, putative Calcyclin binding protein, putative
	PFD0090c ---> PFE1370w---> PF10_0272	0.023	PFD0090c PFE1370w PF10_0272	Plasmodium exported protein (PHISTa), unknown function Hsp70 interacting protein, putative 60S ribosomal protein L3, putative
	PFE1370w ---> PF10_0272	0.007	PFE1370w PF10_0272	Hsp70 interacting protein, putative 60S ribosomal protein L3, putative

Table 4.5 Predicted minimum pathways (potential signal transduction pathways) for the cyclic nucleotide proteins: Column one indicates the name of the signalling pathway, the second column shows minimum paths extracted, while optimizing the identified number of proteins in the pathway under consideration. The third column shows the weight p-value and column four detailed the products (from plasmodb) of the proteins in the identified potential signalling pathways.

Name	Minimum path	p-value	Details of genes	
			Genes IDS	Products
Cyclic Nucleotide	**PFB0190c--->PFC0435w--->PFE0660c ---> PF10_0254 ---> MAL13P1.202**	**0.033**	**PFB0190c PFC0435w PFE0660c PF10_0254 MAL13P1.202**	**conserved Plasmodium protein, unknown function conserved Plasmodium protein, unknown function purine nucleotide phosphorylase, putative conserved Plasmodium protein, unknown function conserved Plasmodium protein, unknown function**
	PFC0435w--->PFE0660c --->PF10_0281---> PF11_0224	**0.016**	**PFC0435w PFE0660c PF10_0281 PF11_0224**	**conserved Plasmodium protein, unknown function purine nucleotide phosphorylase, putative merozoite TRAP-like protein, MTRAP circumsporozoite-related antigen**
	PFC0435w--->PFE0660c ---> PF08_0129---> PF11_0111 --->MAL13P1.202	**0.028**	**PFC0435w PFE0660c PF08_0129 PF11_0111 MAL13P1.202**	**conserved Plasmodium protein, unknown function purine nucleotide phosphorylase, putative serine/threonine protein phosphatase, putative asparagine-rich antigen conserved Plasmodium protein, unknown function**
	PFC0435w---> PFE0660c---> PF11_0224	0.007	PFC0435w PFE0660c PF11_0224	conserved Plasmodium protein, unknown function purine nucleotide phosphorylase, putative circumsporozoite-related antigen
	PFC0435w---> PFE0660c---> PFL2520w---> MAL13P1.202	**0.016**	**PFC0435w PFE0660c PFL2520w MAL13P1.202**	**conserved Plasmodium protein, unknown function purine nucleotide phosphorylase, putative reticulocyte-binding protein 3 homologue conserved Plasmodium protein, unknown function**
	PFC0435w---> PFE0660c---> MAL13P1.202	0.009	PFC0435w PFE0660c MAL13P1.202	conserved Plasmodium protein, unknown function purine nucleotide phosphorylase, putative conserved Plasmodium protein, unknown function
	PFB0190c---> PFC0435w---> PFE0660c---> MAL13P1.202		PFB0190c PFC0435w PFE0660c MAL13P1.202	conserved Plasmodium protein, unknown function conserved Plasmodium protein, unknown function purine nucleotide phosphorylase, putative conserved Plasmodium protein, unknown function
	PFC0435w---> PFE0660c	0.003	PFC0435w PFE0660c	conserved Plasmodium protein, unknown function purine nucleotide phosphorylase, putative
	PFC0435w---> PFE0660c---> PF08_0129---> MAL13P1.202	0.016	PFC0435w PFE0660c PF08_0129 MAL13P1.202	conserved Plasmodium protein, unknown function purine nucleotide phosphorylase, putative serine/threonine protein phosphatase, putative conserved Plasmodium protein, unknown function

Table 4.6 Predicted minimum pathways (potential signal transduction pathways) for the phosphatidylinositol cycle proteins: Column one indicates the name of the signalling pathway, the second column shows minimum paths extracted, while optimizing the identified number of proteins in the pathway under consideration. The third column shows the weight p-value and column four detailed the products (from plasmodb) of the proteins in the identified potential signalling pathways.

Name	Minimum path	p-value	Details of genes	
			Genes IDS	Products
Phosphatidylinositol cycle	PFA0110w --->PFE0750c---> MAL13P1.256	0.005	PFA0110w PFE0750c MAL13P1.256	DNAJ protein, putative RNA recognition motif, putative phosphatidylinositol transfer protein, putative
	PFA0285c---> PFE0750c---> MAL13P1.256	0.029	PFA0285c PFE0750c MAL13P1.256	conserved Plasmodium protein, unknown function RNA recognition motif, putative phosphatidylinositol transfer protein, putative
	PFA0635c---> PFE0750c---> MAL13P1.256	0.043	PFA0635c PFE0750c MAL13P1.256	Plasmodium exported protein (hyp1), unknown function RNA recognition motif, putative phosphatidylinositol transfer protein, putative
	PFE0750c---> PF10_0115---> MAL13P1.256	0.005	PFE0750c PF10_0115 MAL13P1.256	RNA recognition motif, putative QF122 antigen phosphatidylinositol transfer protein, putative
	PFE0750c---> PF11_0175---> MAL13P1.256	0.019	PFE0750c PF11_0175 MAL13P1.256	RNA recognition motif, putative heat shock protein 101, putative phosphatidylinositol transfer protein, putative
	PFE0750c---> PFL0830w---> MAL13P1.256	0.019	PFE0750c PFL0830w MAL13P1.256	RNA recognition motif, putative RNA binding protein, putative phosphatidylinositol transfer protein, putative
	PFE0750c---> PFL1385c--> MAL13P1.256	**0.005**	**PFE0750c PFL1385c MAL13P1.256**	**RNA recognition motif, putative Merozoite Surface Protein 9, MSP-9 phosphatidylinositol transfer protein, putative**
	PFE0750c---> PFL1930w---> MAL13P1.256	**0.005**	**PFE0750c PFL1930w MAL13P1.256**	**RNA recognition motif, putative conserved Plasmodium protein, unknown function phosphatidylinositol transfer protein, putative**
	PFE0750c---> MAL13P1.56-→ MAL13P1.256	0.019	PFE0750c MAL13P1.56 MAL13P1.256	RNA recognition motif, putative m1-family aminopeptidase phosphatidylinositol transfer protein, putative
	PFE0750c ---> PF13_0091--> MAL13P1.256	0.043	PFE0750c PF13_0091 MAL13P1.256	RNA recognition motif, putative conserved Plasmodium protein, unknown function phosphatidylinositol transfer protein, putative
	PFE0750c ---> PF10_0115---> PF13_0165 --> MAL13P1.256	0.048	PFE0750c PF10_0115 PF13_0165 MAL13P1.256	RNA recognition motif, putative QF122 antigen conserved Plasmodium protein, unknown function phosphatidylinositol transfer protein, putative
	PFE0750c---> MAL13P1.256--> PF13_0315	0.043	PFE0750c MAL13P1.256 PF13_0315	RNA recognition motif, putative phosphatidylinositol transfer protein, putative RNA binding protein, putative
	PFA0110w---> PFE0750c--> MAL13P1.256--> PF14_0257	**0.046**	**PFA0110w PFE0750c MAL13P1.256 PF14_0257**	**DNAJ protein, putative RNA recognition motif, putative phosphatidylinositol transfer protein, putative conserved protein, unknown function**
	PFE0750c ---> PF10_0115 ---> MAL13P1.256--> PF14_0344	0.043	PFE0750c PF10_0115 MAL13P1.256 PF14_0344	RNA recognition motif, putative QF122 antigen phosphatidylinositol transfer protein, putative conserved Plasmodium protein, unknown function
	PFC0425w---> PFE0750c---> MAL13P1.256	0.01	PFC0425w PFE0750c MAL13P1.256	conserved Plasmodium protein, unknown function RNA recognition motif, putative phosphatidylinositol transfer protein, putative
	PFA0110w---> PFD0090c--> PFE0750c--> MAL13P1.256	**0.029**	**PFA0110w PFD0090c PFE0750c MAL13P1.256**	**DNAJ protein, putative Plasmodium exported protein (PHISTa), unknown function RNA recognition motif, putative phosphatidylinositol transfer protein, putative**

	PFD0835c---> PFE0750c ---> PF10_0115---> MAL13P1.256	0.043	PFD0835c PFE0750c PF10_0115 MAL13P1.256	LETM1-like protein, putative RNA recognition motif, putative QF122 antigen phosphatidylinositol transfer protein, putative
	PFD0950w---> PFE0750c---> MAL13P1.256	0.019	PFD0950w PFE0750c MAL13P1.256	ran binding protein 1, putative RNA recognition motif, putative phosphatidylinositol transfer protein, putative
	PFE0750c---> MAL13P1.256	0.005	PFE0750c MAL13P1.256	RNA recognition motif, putative phosphatidylinositol transfer protein, putative
	PFE0750c--> PFF0785w--> MAL13P1.256	0.007	PFE0750c PFF0785w MAL13P1.256	RNA recognition motif, putative Ndc80 homologue, putative phosphatidylinositol transfer protein, putative
	PFE0750c--> PFF1050w--> PF10_0115--> MAL13P1.256	**0.043**	**PFE0750c PFF1050w PF10_0115 MAL13P1.256**	**RNA recognition motif, putative nascent polypeptide associated complex alpha chain, putative QF122 antigen phosphatidylinositol transfer protein, putative**
	PFE0750c-->PFF1100c-->MAL13P1.256	0.017	PFE0750c PFF1100c MAL13P1.256	RNA recognition motif, putative transcription factor with AP2 domain(s), putative phosphatidylinositol transfer protein, putative

Table 4.7 Predicted minimum pathways (potential signal transduction pathways) for the FIKK proteins: Column one indicates the name of the signalling pathway, the second column shows minimum paths extracted, while optimizing the identified number of proteins in the pathway under consideration. The third column shows the weight p-value and column four detailed the products (from plasmodb) of the proteins in the identified potential signalling pathways.

Name	Minimum path	p-value	Details of genes	
			Genes IDS	Products
FIKK	**PFA0130c ---> PFE1590w ---> MAL8P1.153---> PFA0215w**	**0.046**	**PFA0130c PFE1590w MAL8P1.153 PFA0215w**	**Serine/Threonine protein kinase, FIKK family, putative early transcribed membrane protein 5, ETRAMP5 transcription factor with AP2 domain(s), putative Nill**
	PFA0130c ---> PFE1590w ---> MAL8P1.153---> PF10_0075	**0.039**	**PFA0130c PFE1590w MAL8P1.153 PF10_0075**	**Serine/Threonine protein kinase, FIKK family, putative early transcribed membrane protein 5, ETRAMP5 transcription factor with AP2 domain(s), putative transcription factor with AP2 domain(s), putative**
	PFA0130c ---> PF10_0143	0.005	PFA0130c PF10_0143	Serine/Threonine protein kinase, FIKK family, putative transcriptional activator ADA2, putative
	PFA0130c ---> PFE1590w ---> PF10_0232	0.009	PFA0130c PFE1590w PF10_0232	Serine/Threonine protein kinase, FIKK family, putative early transcribed membrane protein 5, ETRAMP5 Chromodomain-helicase-DNA-binding protein 1 homolog, putative
	PFA0130c ---> PFE1590w ---> PF11_0142	0.009	PFA0130c PFE1590w PF11_0142	Serine/Threonine protein kinase, FIKK family, putative early transcribed membrane protein 5, ETRAMP5 ubiquitin domain containing protein
	PFA0130c ---> PFE1590w ---> PF11_0302	0.014	PFA0130c PFE1590w PF11_0302	Serine/Threonine protein kinase, FIKK family, putative early transcribed membrane protein 5, ETRAMP5 conserved Plasmodium protein, unknown function
	PFA0130c ---> PFE1590w ---> MAL8P1.153---> PF11_0342	**0.036**	**PFA0130c PFE1590w MAL8P1.153 PF11_0342**	**Serine/Threonine protein kinase, FIKK family, putative early transcribed membrane protein 5, ETRAMP5 transcription factor with AP2 domain(s), putative conserved Plasmodium protein, unknown function**
	PFA0130c ---> PFE1590w ---> PF10_0232---> PF11_0506	**0.036**	**PFA0130c PFE1590w PF10_0232 PF11_0506**	**Serine/Threonine protein kinase, FIKK family, putative early transcribed membrane protein 5, ETRAMP5 Chromodomain-helicase-DNA-binding protein 1 homolog, putative Antigen 332, DBL-like protein**
	PFA0130c--->PFE1590w--->MAL8P1.104---> PFL0350c	0.046	PFA0130c PFE1590w MAL8P1.104 PFL0350c	Serine/Threonine protein kinase, FIKK family, putative early transcribed membrane protein 5, ETRAMP5 CAF1 family ribonuclease, putative conserved Plasmodium protein, unknown function
	PFA0130c--->PFE1590w--->PFF0590c---> MAL8P1.153--> PFL1385c	**0.046**	**PFA0130c PFE1590w PFF0590c**	**Serine/Threonine protein kinase, FIKK family, putative early transcribed membrane protein 5, ETRAMP5 homologue of human HSPC025**

			MAL8P1.153 **PFL1385c**	**transcription factor with AP2 domain(s), putative** **Merozoite Surface Protein 9, MSP-9**
	PFA0130c--->PFE1590w--->MAL8P1.153---> PFL1395c	0.036	PFA0130c PFE1590w MAL8P1.153 PFL1395c	Serine/Threonine protein kinase, FIKK family, putative early transcribed membrane protein 5, ETRAMP5 transcription factor with AP2 domain(s), putative conserved Plasmodium protein, unknown function
	PFA0130c--->PFE1590w--->PF10_0232--->PFL1705w	0.039	PFA0130c PFE1590w PF10_0232 PFL1705w	Serine/Threonine protein kinase, FIKK family, putative early transcribed membrane protein 5, ETRAMP5 Chromodomain-helicase-DNA-binding protein 1 homolog, putative RNA binding protein, putative
	PFA0130c--->PFE1590w--->PFL1900w	0.014	PFA0130c PFE1590w PFL1900w	Serine/Threonine protein kinase, FIKK family, putative early transcribed membrane protein 5, ETRAMP5 transcription factor with AP2 domain(s), putative
	PFA0130c--->PFE1590w--->PFL2520w--->MAL13P1.202	0.036	PFA0130c PFE1590w PFL2520w MAL13P1.202	Serine/Threonine protein kinase, FIKK family, putative early transcribed membrane protein 5, ETRAMP5 reticulocyte-binding protein 3 homologue conserved Plasmodium protein, unknown function
	PFA0130c--->PFE1590w--->PFF1185w--->PF13_0036	0.036	PFA0130c PFE1590w PFF1185w PF13_0036	Serine/Threonine protein kinase, FIKK family, putative early transcribed membrane protein 5, ETRAMP5 Smarca -related protein DNAJ protein, putative
	PFA0130c--->PFE1590w--->PF13_0044	0.03	PFA0130c PFE1590w PF13_0044	Serine/Threonine protein kinase, FIKK family, putative early transcribed membrane protein 5, ETRAMP5 carbamoyl phosphate synthetase
	PFA0130c--->PFE1590w--->MAL13P1.202	0.007	PFA0130c PFE1590w MAL13P1.202	Serine/Threonine protein kinase, FIKK family, putative early transcribed membrane protein 5, ETRAMP5 conserved Plasmodium protein, unknown function
	PFA0130c--->PFE1590w--->MAL8P1.153--> MAL13P1.275	0.046	PFA0130c PFE1590w MAL8P1.153 MAL13P1.275	Serine/Threonine protein kinase, FIKK family, putative early transcribed membrane protein 5, ETRAMP5 transcription factor with AP2 domain(s), putative protein phosphatase, putative
	PFA0130c--->PFE1590w--->MAL8P1.153---> PF14_0031	0.039	PFA0130c PFE1590w MAL8P1.153 PF14_0031	Serine/Threonine protein kinase, FIKK family, putative early transcribed membrane protein 5, ETRAMP5 transcription factor with AP2 domain(s), putative conserved Plasmodium protein, unknown function
	PFA0130c--->PFE1590w--->MAL8P1.153---> MAL8P1.23--->PF14_0463	**0.039**	**PFA0130c PFE1590w MAL8P1.153 MAL8P1.23 PF14_0463**	**Serine/Threonine protein kinase, FIKK family, putative early transcribed membrane protein 5, ETRAMP5 transcription factor with AP2 domain(s), putative ubiquitin-protein ligase 1, putative chloroquine resistance marker protein**
	PFA0130c--->PFE1590w--->PF14_0636	0.036	PFA0130c PFE1590w PF14_0636	Serine/Threonine protein kinase, FIKK family, putative early transcribed membrane protein 5, ETRAMP5 conserved Plasmodium protein, unknown function
	PFA0130c--->PFE1590w--->PF10_0232--> PF14_0644	0.039	PFA0130c PFE1590w PF10_0232 PF14_0644	Serine/Threonine protein kinase, FIKK family, putative early transcribed membrane protein 5, ETRAMP5 Chromodomain-helicase-DNA-binding protein 1 homolog, putative conserved Plasmodium protein, unknown function
	PFA0130c--->PFE1590w--->MAL7P1.19--->PF14_0678	0.039	PFA0130c PFE1590w MAL7P1.19 PF14_0678	Serine/Threonine protein kinase, FIKK family, putative early transcribed membrane protein 5, ETRAMP5 ubiquitin transferase, putative exported protein 2
	PFA0130c--->PFE1590w---> MAL8P1.153---> PF14_0679	0.046	PFA0130c PFE1590w MAL8P1.153 PF14_0679	Serine/Threonine protein kinase, FIKK family, putative early transcribed membrane protein 5, ETRAMP5 transcription factor with AP2 domain(s), putative inorganic anion exchanger, inorganic anion antiporter
	PFA0130c--->PFB0190c--->PFE1590w---> MAL13P1.202	0.039	PFA0130c PFB0190c PFE1590w MAL13P1.202	Serine/Threonine protein kinase, FIKK family, putative conserved Plasmodium protein, unknown function early transcribed membrane protein 5, ETRAMP5 conserved Plasmodium protein, unknown function
	PFA0130c--->PFC0390w--->PFE1590w	0.009	PFA0130c PFC0390w PFE1590w	Serine/Threonine protein kinase, FIKK family, putative N2227-like protein, putative early transcribed membrane protein 5, ETRAMP5
	PFA0130c-->PFD0835c-->PFE1590w--> MAL8P1.153	0.036	PFA0130c PFD0835c PFE1590w MAL8P1.153	Serine/Threonine protein kinase, FIKK family, putative LETM1-like protein, putative early transcribed membrane protein 5, ETRAMP5 transcription factor with AP2 domain(s), putative
	PFA0130c-->PFD0885c-->PFE1590w	0.007	PFA0130c PFD0885c PFE1590w	Serine/Threonine protein kinase, FIKK family, putative conserved Plasmodium protein, unknown function early transcribed membrane protein 5, ETRAMP5
	PFA0130c-->PFD0985w-->PFE1590w--> MAL8P1.153	0.036	PFA0130c PFD0985w PFE1590w	Serine/Threonine protein kinase, FIKK family, putative transcription factor with AP2 domain(s), putative early transcribed membrane protein 5, ETRAMP5

			MAL8P1.153	transcription factor with AP2 domain(s), putative
	PFA0130c-->PFE0070w-->PFE1590w--> MAL7P1.19	0.039	PFA0130c PFE0070w PFE1590w MAL7P1.19	Serine/Threonine protein kinase, FIKK family, putative interspersed repeat antigen, putative early transcribed membrane protein 5, ETRAMP5 ubiquitin transferase, putative
	PFA0130c-->PFE1225w-->PFE1590w	0.036	PFA0130c PFE1225w PFE1590w	Serine/Threonine protein kinase, FIKK family, putative organelle ribosomal protein L7/L12 precursor, putative early transcribed membrane protein 5, ETRAMP5
	PFA0130c-->PFE1590w	0.005	PFA0130c PFE1590w	Serine/Threonine protein kinase, FIKK family, putative early transcribed membrane protein 5, ETRAMP5
	PFA0130c-->PFE1590w-->PFE1605w--> MAL8P1.153	**0.036**	**PFA0130c PFE1590w PFE1605w MAL8P1.153**	**Serine/Threonine protein kinase, FIKK family, putative early transcribed membrane protein 5, ETRAMP5 Plasmodium exported protein (PHISTb), unknown function transcription factor with AP2 domain(s), putative**
	PFA0130c-->PFE1590w-->PFF0220w-->PFF0590c --> MAL8P1.153	**0.046**	**PFA0130c PFE1590w PFF0220w PFF0590c MAL8P1.153**	**Serine/Threonine protein kinase, FIKK family, putative early transcribed membrane protein 5, ETRAMP5 conserved Plasmodium protein, unknown function homologue of human HSPC025 transcription factor with AP2 domain(s), putative**
	PFA0130c-->PFE1590w-->PFF0590c--> MAL8P1.153	0.018	PFA0130c PFE1590w PFF0590c MAL8P1.153	Serine/Threonine protein kinase, FIKK family, putative early transcribed membrane protein 5, ETRAMP5 homologue of human HSPC025 transcription factor with AP2 domain(s), putative
	PFA0130c-->PFE1590w-->PFF0835w	0.036	PFA0130c PFE1590w PFF0835w	Serine/Threonine protein kinase, FIKK family, putative early transcribed membrane protein 5, ETRAMP5 conserved Plasmodium protein, unknown function
	PFA0130c-->PFE1590w-->PFF0920c-->PF10_0232	0.036	PFA0130c PFE1590w PFF0920c PF10_0232	Serine/Threonine protein kinase, FIKK family, putative early transcribed membrane protein 5, ETRAMP5 conserved Plasmodium protein, unknown function Chromodomain-helicase-DNA-binding protein 1 homolog, putative
	PFA0130c-->PFE1590w-->PFF1185w	0.007	PFA0130c PFE1590w PFF1185w	Serine/Threonine protein kinase, FIKK family, putative early transcribed membrane protein 5, ETRAMP5 Smarca -related protein
	PFA0130c-->PFE1590w--> PFF1220w--> MAL8P1.153	**0.036**	**PFA0130c PFE1590w PFF1220w MAL8P1.153**	**Serine/Threonine protein kinase, FIKK family, putative early transcribed membrane protein 5, ETRAMP5 conserved Plasmodium protein, unknown function transcription factor with AP2 domain(s), putative**
	PFA0130c-->PFE1590w--> MAL7P1.19	0.014	PFA0130c PFE1590w MAL7P1.19	Serine/Threonine protein kinase, FIKK family, putative early transcribed membrane protein 5, ETRAMP5 ubiquitin transferase, putative
	PFA0130c-->PFE1590w--> PF07_0044	0.03	PFA0130c PFE1590w PF07_0044	Serine/Threonine protein kinase, FIKK family, putative early transcribed membrane protein 5, ETRAMP5 conserved Plasmodium protein, unknown function
	PFA0130c-->PFE1590w--> PF07_0056--> MAL8P1.153--> MAL8P1.23	**0.036**	**PFA0130c PFE1590w PF07_0056 MAL8P1.153 MAL8P1.23**	**Serine/Threonine protein kinase, FIKK family, putative early transcribed membrane protein 5, ETRAMP5 conserved Plasmodium protein, unknown function transcription factor with AP2 domain(s), putative ubiquitin-protein ligase 1, putative**
	PFA0130c--> PF07_0115	0.005	PFA0130c PF07_0115	Serine/Threonine protein kinase, FIKK family, putative cation transporting ATPase, cation transporter
	PFA0130c-->PFE1590w--> MAL7P1.171--> MAL8P1.104	0.039	PFA0130c PFE1590w MAL7P1.171 MAL8P1.104	Serine/Threonine protein kinase, FIKK family, putative early transcribed membrane protein 5, ETRAMP5 Plasmodium exported protein, unknown function CAF1 family ribonuclease, putative
	PFA0130c-->PFE1590w--> MAL8P1.153	**0.007**	**PFA0130c PFE1590w MAL8P1.153**	**Serine/Threonine protein kinase, FIKK family, putative early transcribed membrane protein 5, ETRAMP5 transcription factor with AP2 domain(s), putative**
	PFA0130c-->PFE1590w--> PF08_0129--> MAL13P1.202	0.036	PFA0130c PFE1590w PF08_0129 MAL13P1.202	Serine/Threonine protein kinase, FIKK family, putative early transcribed membrane protein 5, ETRAMP5 serine/threonine protein phosphatase, putative conserved Plasmodium protein, unknown function
	PFA0130c-->PFE1590w--> MAL8P1.104	0.014	PFA0130c PFE1590w MAL8P1.104	Serine/Threonine protein kinase, FIKK family, putative early transcribed membrane protein 5, ETRAMP5 CAF1 family ribonuclease, putative
	PFA0130c-->PFE1590w--> MAL8P1.153--> PF08_0034	**0.041**	**PFA0130c PFE1590w MAL8P1.153 PF08_0034**	**Serine/Threonine protein kinase, FIKK family, putative early transcribed membrane protein 5, ETRAMP5 transcription factor with AP2 domain(s), putative histone acetyltransferase GCN5, putative**
	PFA0130c-->PFE1590w--> MAL8P1.153--> MAL8P1.23	**0.036**	**PFA0130c PFE1590w MAL8P1.153 MAL8P1.23**	**Serine/Threonine protein kinase, FIKK family, putative early transcribed membrane protein 5, ETRAMP5 transcription factor with AP2 domain(s), putative ubiquitin-protein ligase 1, putative**

Table 4.8: Predicted minimum pathways (potential signal transduction pathways) for the unknown signal transduction groups proteins. Column one indicates the name of the signalling pathway, the second column shows minimum paths extracted, while optimizing the identified number of proteins in the pathway under consideration. The third column shows the weight p-value and column four detailed the products (from plasmodb) of the proteins in the identified potential signalling pathways.

Name	Minimum path	p-value	Details of genes	
			Genes IDS	Products
Unknown genes	PFA0125c --->PFA0515w---> PF08_0034--> PF11_0504	0.045	PFA0125c PFA0515w PF08_0034 PF11_0504	erythrocyte binding antigen-181 phosphatidylinositol-4-phosphate-5-kinase histone acetyltransferase GCN5, putative Plasmodium exported protein (hyp11), unknown function
	PFA0125c---> PF08_0034---> PF10_0146--> PF11_0504	0.039	PFA0125c PF08_0034 PF10_0146 PF11_0504	erythrocyte binding antigen-181 histone acetyltransferase GCN5, putative conserved Plasmodium protein, unknown function Plasmodium exported protein (hyp11), unknown function
	PFA0125c---> PF08_0034--> PF10_0232--> PF11_0504	0.029	PFA0125c PF08_0034 PF10_0232 PF11_0504	erythrocyte binding antigen-181 histone acetyltransferase GCN5, putative Chromodomain-helicase-DNA-binding protein 1 homolog, putative Plasmodium exported protein (hyp11), unknown function
	PFA0125c---> PFE0570w---> PF11_0055 --> PF11_0277	0.03	PFA0125c PFE0570w PF11_0055 PF11_0277	erythrocyte binding antigen-181 RNA pseudouridylate synthase, putative conserved protein, unknown function conserved Plasmodium protein, unknown function
	PFA0125c---> PF08_0034--> PF11_0504--> PFL0815w	0.033	PFA0125c PF08_0034 PF11_0504 PFL0815w	erythrocyte binding antigen-181 histone acetyltransferase GCN5, putative Plasmodium exported protein (hyp11), unknown function DNA-binding chaperone, putative
	PFA0125c ---> PFE0570w ---> PF11_0277 --> PFL1385c	**0.027**	**PFA0125c PFE0570w PF11_0277 PFL1385c**	**erythrocyte binding antigen-181 RNA pseudouridylate synthase, putative conserved Plasmodium protein, unknown function Merozoite Surface Protein 9, MSP-9**
	PFA0125c ---> PFE0570w--> PF11_0277--> PFL1565c	0.03	PFA0125c PFE0570w PF11_0277 PFL1565c	erythrocyte binding antigen-181 RNA pseudouridylate synthase, putative conserved Plasmodium protein, unknown function CG2-related protein, putative
	PFA0125c ---> PF11_0277--> PFL2520w--> chr13_1000012.gen_6	0.03	PFA0125c PF11_0277 PFL2520w Chr13_1000012.gen_6	erythrocyte binding antigen-181 conserved Plasmodium protein, unknown function reticulocyte-binding protein 3 homologue nil
	PFA0125c ---> PF11_0277--> MAL13P1.135--> chr13_1000012.gen_6	0.032	PFA0125c PF11_0277 MAL13P1.135 chr13_1000012.gen_6	erythrocyte binding antigen-181 conserved Plasmodium protein, unknown function SNARE protein, putative nil
	PFA0125c---> PF08_0034--> PF11_0504--> PF13_0161	0.032	PFA0125c PF08_0034 PF11_0504 PF13_0161	erythrocyte binding antigen-181 histone acetyltransferase GCN5, putative Plasmodium exported protein (hyp11), unknown function conserved Plasmodium protein, unknown function
	PFA0125c---> PF08_0034--> PF11_0504---> PF13_0173	0.049	PFA0125c PF08_0034 PF11_0504 PF13_0173	erythrocyte binding antigen-181 histone acetyltransferase GCN5, putative Plasmodium exported protein (hyp11), unknown function conserved Plasmodium protein, unknown function
	PFA0125c---> PFE0570w---> PFE1590w--> PF11_0277--> MAL13P1.202	0.036	PFA0125c PFE0570w PFE1590w PF11_0277 MAL13P1.202	erythrocyte binding antigen-181 RNA pseudouridylate synthase, putative early transcribed membrane protein 5, ETRAMP5 conserved Plasmodium protein, unknown function conserved Plasmodium protein, unknown function
	PFA0125c---> PF08_0034---> PF11_0504---> MAL13P1.269	0.035	PFA0125c PF08_0034 PF11_0504 MAL13P1.269	erythrocyte binding antigen-181 histone acetyltransferase GCN5, putative Plasmodium exported protein (hyp11), unknown function tryptophan-rich antigen, putative
	PFA0125c---> PFE0570w--> PF11_0277--> PF13_0322	0.029	PFA0125c PFE0570w PF11_0277 PF13_0322	erythrocyte binding antigen-181 RNA pseudouridylate synthase, putative conserved Plasmodium protein, unknown function falcilysin
	PFA0125c---> PF08_0034 ---> PF11_0504--->	0.043	PFA0125c	erythrocyte binding antigen-181

	PF14_0230		PF08_0034 PF11_0504 PF14_0230	histone acetyltransferase GCN5, putative Plasmodium exported protein (hyp11), unknown function 60S ribosomal protein L5, putative
	PFA0125c---> PF08_0034---> PF11_0504--> PF14_0510	0.032	PFA0125c PF08_0034 PF11_0504 PF14_0510	erythrocyte binding antigen-181 histone acetyltransferase GCN5, putative Plasmodium exported protein (hyp11), unknown function p23 co-chaperone, putative
	PFA0125c---> PF08_0034---> PF11_0504--> PF14_0678	0.032	PFA0125c PF08_0034 PF11_0504 PF14_0510	erythrocyte binding antigen-181 histone acetyltransferase GCN5, putative Plasmodium exported protein (hyp11), unknown exported protein 2
	PFA0125c--> PFB0095c--> PF08_0034--> PF11_0504	0.049	PFA0125c PFB0095c PF08_0034 PF11_0504	erythrocyte binding antigen-181 erythrocyte membrane protein 3 histone acetyltransferase GCN5, putative Plasmodium exported protein (hyp11), unknown function
	PFA0125c--> PFD0835c--> PF08_0034--> PF11_0504	0.03	PFA0125c PFD0835c PF08_0034 PF11_0504	erythrocyte binding antigen-181 LETM1-like protein, putative histone acetyltransferase GCN5, putative Plasmodium exported protein (hyp11), unknown function
	PFA0125c--> PFD0885c--> PFE0570w--> PFE1590w--> PF11_0277	0.036	PFA0125c PFD0885c PFE0570w PFE1590w PF11_0277	erythrocyte binding antigen-181 conserved Plasmodium protein, unknown function RNA pseudouridylate synthase, putative early transcribed membrane protein 5, ETRAMP5 conserved Plasmodium protein, unknown function
	PFA0125c-->PFE0770w-->PF11_0277--> chr13_1000012.gen_6	0.027	PFA0125c PFE0770w- PF11_0277 chr13_1000012.gen_6	erythrocyte binding antigen-181 conserved Plasmodium protein, unknown function conserved Plasmodium protein, unknown function nil
	PFA0125c-->PFE0845c-->PF08_0034--> PF11_0504	0.048	PFA0125c PFE0845c PF08_0034 PF11_0504	erythrocyte binding antigen-181 60S ribosomal protein L8, putative histone acetyltransferase GCN5, putative Plasmodium exported protein (hyp11), unknown function
	PFA0125c-->PFE1225w-->PF11_0277--> chr13_1000012.gen_6	0.027	PFA0125c PFE1225w PF11_0277 chr13_1000012.gen_6	erythrocyte binding antigen-181 organelle ribosomal protein L7/L12 precursor, putative conserved Plasmodium protein, unknown function nil
	PFA0125c-->PFE1465w-->PF11_0277--> chr13_1000012.gen_6	0.03	PFA0125c PFE1465w PF11_0277 chr13_1000012.gen_6	erythrocyte binding antigen-181 conserved Plasmodium protein, unknown function conserved Plasmodium protein, unknown function nil
	PFA0125c -->PFE0570w -->PFE1590w--> PF11_0277	0.024	PFA0125c PFE0570w PFE1590w PF11_0277	erythrocyte binding antigen-181 RNA pseudouridylate synthase, putative early transcribed membrane protein 5, ETRAMP5 conserved Plasmodium protein, unknown function
	PFA0125c-->PFE0570w-->PFE1590w--> PFF1185w--> PF11_0277	0.042	PFA0125c PFE0570w PFE1590w PFF1185w PF11_0277	erythrocyte binding antigen-181 RNA pseudouridylate synthase, putative early transcribed membrane protein 5, ETRAMP5 Smarca -related protein conserved Plasmodium protein, unknown function
	PFA0125c-->PFE0570w--> PFE1590w--> MAL8P1.153--> PF11_0277	0.043	PFA0125c PFE0570w PFE1590w MAL8P1.153 PF11_0277	erythrocyte binding antigen-181 RNA pseudouridylate synthase, putative early transcribed membrane protein 5, ETRAMP5 transcription factor with AP2 domain(s), putative conserved Plasmodium protein, unknown function
	PFA0125c -->PF08_0127 -->PF08_0034 --> PF11_0504	0.043	PFA0125c PF08_0127 PF08_0034 PF11_0504	erythrocyte binding antigen-181 conserved Plasmodium protein, unknown function histone acetyltransferase GCN5, putative Plasmodium exported protein (hyp11), unknown function
	PFA0125c-->PF08_0034-->PFI0495w--> PF11_0504	0.045	PFA0125c PF08_0034 PFI0495w PF11_0504	erythrocyte binding antigen-181 histone acetyltransferase GCN5, putative conserved Plasmodium protein, unknown function Plasmodium exported protein (hyp11), unknown function
	PFA0125c-->PF08_0034--> PFI1715w--> PF10_0232--> PF11_0504	0.046	PFA0125c PF08_0034 PFI1715w PF10_0232 PF11_0504	erythrocyte binding antigen-181 histone acetyltransferase GCN5, putative Plasmodium exported protein ,unknown function Chromodomain-helicase-DNA-binding protein 1 homolog, putative Plasmodium exported protein (hyp11), unknown function
	PFA0125c-->PFE0570w--> PFE1590w--> PF11_0277--> MAL13P1.86	0.049	PFA0125c PFE0570w PFE1590w PF11_0277 MAL13P1.86	erythrocyte binding antigen-181 RNA pseudouridylate synthase, putative early transcribed membrane protein 5, ETRAMP5 conserved Plasmodium protein, unknown function cholinephosphate cytidylyltransferase
	PFA0125c--> PF08_0034--> PF11_0504-->	0.049	PFA0125c	erythrocyte binding antigen-181

	MAL13P1.88		PF08_0034 PF11_0504 MAL13P1.88	histone acetyltransferase GCN5, putative Plasmodium exported protein (hyp11), unknown function conserved Plasmodium protein, unknown function
	PFA0125c-->PF10_0081	0.009	PFA0125c PF10_0081	erythrocyte binding antigen-181 26S proteasome regulatory subunit 4, putative
	PFA0125c-->PF10_0183-->PF11_0277	0.013	PFA0125c PF10_0183 PF11_0277	erythrocyte binding antigen-181 eukaryotic translation initiation factor subunit eIF2A, putative conserved Plasmodium protein, unknown function
	PFA0125c--> PF11_0277	0.001	PFA0125c PF11_0277	erythrocyte binding antigen-181 conserved Plasmodium protein, unknown function
	PFA0125c--> PF08_0034--> PF11_0504	0.017	PFA0125c PF08_0034 PF11_0504	erythrocyte binding antigen-181 histone acetyltransferase GCN5, putative Plasmodium exported protein (hyp11), unknown function
	PFA0125c-->PFL0830w	0.003	PFA0125c PFL0830w	erythrocyte binding antigen-181 RNA binding protein, putative
	PFA0125c-->PF14_0241	0.006	PFA0125c PF14_0241	erythrocyte binding antigen-181 basic transcription factor 3b, putative
	PFA0125c--> PFC0465c	0.006	PFA0125c PFC0465c	erythrocyte binding antigen-181 pre-mRNA splicing factor, putative
	PFA0125c--> PFD0795w-->PF11_0277	0.013	PFA0125c PFD0795w PF11_0277	erythrocyte binding antigen-181 RNA pseudouridylate synthase, putative conserved Plasmodium protein, unknown function
	PFA0125c--> PFE0570w -->PF11_0277	0.013	PFA0125c PFE0570w PF11_0277	erythrocyte binding antigen-181 RNA pseudouridylate synthase, putative conserved Plasmodium protein, unknown function
	PFA0125c -->PFF1050w -->PF14_0241	0.026	PFA0125c PFF1050w PF14_0241	erythrocyte binding antigen-181 nascent polypeptide associated complex alpha chain, putative basic transcription factor 3b, putative
	PFA0125c -->MAL7P1.172 -->PF10_0081	0.024	PFA0125c MAL7P1.172 PF10_0081	erythrocyte binding antigen-181 Plasmodium exported protein (PHISTc), unknown function 26S proteasome regulatory subunit 4, putative
	PFA0125c-->PF08_0034	0.001	PFA0125c PF08_0034	erythrocyte binding antigen-181 histone acetyltransferase GCN5, putative
	PFA0125c--> PFI0635c -->PF11_0277	0.016	PFA0125c PFI0635c PF11_0277	erythrocyte binding antigen-181 conserved Plasmodium protein, unknown function conserved Plasmodium protein, unknown function
	PFA0125c--> PFI1090w	0.003	PFA0125c PFI1090w	erythrocyte binding antigen-181 S-adenosylmethionine synthetase
	PFA0125c--> PF11_0277--> chr13_1000012.gen_6	0.013	PFA0125c PF11_0277 chr13_1000012.gen_6	erythrocyte binding antigen-181 conserved Plasmodium protein, unknown function nil

LIST OF PUBLICATIONS IN THIS RESEARCH WORK

Oyelade, O. J., Adebiyi, E. F., Yah S.C., and Olaseinde G. I. Computational Identification of functional related gene in Malaria parasites. Poster proceeding of the ISMB 2008.

Jelili Oyelade, Ezekiel Adebiyi, Svetlana Bulashevska, Benedikt Brors and Roland Eils. Computational Identification of functional modules in *Plasmodium falciparum*. Proceeding of the 1st Joint ISCB Africa ASBCB Conference on Bioinformatics of Infectious Diseases, MRTC, Bamako, Mali, November 30 – December 2, 2009.

Oyelade, Olanrewaju Jelili, Ezekiel Adebiyi, Benedict Brors and Roland Eils. Computational Identification of functional related genes in Malaria. Poster proceeding of the EMBnet-RiBio, 2009.

Jelili Oyelade, Itunu Ewejobi, Benedikt Brors, Roland Eils, and Ezekiel Adebiyi.Computational Identification of Signalling Pathways in *Plasmodium falciparum*. Infection, Genetics and Evolution-Elsevier Journal, 2010.

Ezekiel Adebiyi, **Jelili Oyelade**, Itunu Ewejobi, Benedikt Brors and Roland Eils. Computational Identification of Metabolic Pathways in the Malaria Parasite, *Plasmodium falciparum*- In preparation