COMPUTATIONAL METHODOLOGIES FOR MODULARIZATION OF SOME BIOCHEMICAL PATHWAYS: ANALYSIS AND ANNOTATION

Thesis submitted for the degree of
Doctor of Philosophy (Science) in
Bio-Physics, Molecular Biology and Bioinformatics
by

Losiana Nayak

Department of Bio-Physics, Molecular Biology and Bioinformatics
University of Calcutta
2012

To my Teachers, Family and Friends

ACKNOWLEDGEMENTS

A dream always takes shape from obscure ideas and vogue notions. My dream of a PhD degree took shape from the failure for qualifying an entrance test to be a Doctor of Medicines. So, first of all, acknowledgement to that failure and disappointment. Next, I acknowledge support of Dr. Rajat K. De, without whose guidance this dream would never have seen the day light. Special gratitude to him for standing beside me and helping me walk this strenuous path. In addition, I am thankful to Prof. Nitai P. Bhattacharyya of Saha Institute of Nuclear Physics for providing crucial insights and ideas for improvement of my research work.

I am thankful to Prof. Sankar K. Pal, Prof. C. A. Murthy, Prof. Sushmita Mitra, Prof. Malay K. Kundu and other faculty members of Machine Intelligence Unit, Indian Statistical Institute, Kolkata for their comments and suggestions towards my research work. I am also thankful to my friends, co-fellows and office staff of Machine Intelligence Unit for helping me around.

I am heartily thankful to Dr. Subhasis Mukhopadhyay, Prof. Abhay S. Chakraborti, Dr. Ansuman Lahiri and other faculty members of the University of Calcutta for their help and suggestive comments. I am also thankful to the office staff of the University of Calcutta for their help.

I thank my parents, grand parents, younger brother and the whole family for their support. Without their encouragement and patience, this work would not have been completed. Special thanks to my husband, Saroj, for his invaluable support while winding up the thesis. Where belief is a debate, I must thank the invisible forces guiding this world, whose non-existence I cannot prove. Thank you all, for believing in me and my dream.

(Losiana Nayak)

Contents

List of Figures

List of Tables

List of Abbreviations

$[Ca^{2+}]_i$	Intracellular free Ca^{2+} ion
AMD	Approximate Minimum Degree Ordering
APC	Adenomatosis Polyposis Coli
AXIN	AXis Inhibition Protein
BINGO tool	BIological Networks Gene Ontology tool
BioPAX	Biological PAthway eXchange
BMP	Bitmap Image Format
BMP	Bone Morphogenetic Protein
CamKII	Calmodulin-dependent Kinase II
cAMP	cyclic Adenosine Mono Phosphate
Catenin	Cadherin-Associated Protein
CREB	cAMP Response Element-Binding
CREBBP	CREB-binding protein
CSNK1E	CaSeiN Kinase 1, Epsilon
CTBP1	C-Terminal Binding Protein 1
CTNNB1	Catenin (cadherin-associated protein), Beta 1
DKK	DicKKopf
DVL	DisheVeLled, dsh homolog
EGFR	Epidermal Growth Factor Receptor
EP300	E1A binding Protein p300
ErbB	Avian erythroblastosis oncogene B
ERK	Extracellular-signal-Regulated Kinase
EVR	Exudative Vitreo Retinopathy
FE_Score	Functional Enrichment score

FOSL1	FOS-Like antigen 1
FZD	FriZzleD seven-transmembrane-span receptor
GenMAPP	Gene MicroArray Pathway Profiler
GIF	Graphics Interchange Format
GO attribute	Gene Ontology attribute
GPML	GenMAPP Pathway Markup Language
GRB2	Growth factor Receptor-Bound protein 2
GSK3	Glycogen Synthase Kinase 3
GWAS	Genome Wide Association Study
Hh	Hedgehog
$InsP_3$	Inositol 1,4,5-trisPhosphate
INT1 gene	Inturned Planar Cell Polarity Effector Homolog (Drosophila) 1 gene
ITPR1	Inositol 1,4,5-TriPhosphate Receptor, type 1
Jak	Janus kinase
JNK	Jun-N-Terminal Kinase
KEGG	Kyoto Encyclopedia of Genes and Genomes
KGML	KEGG Markup Language
LEF	Lymphoid Enhancer-binding Factor 1
LRP	Lipoprotein Receptor-related Protein
MAP2	Microtubule-Associated Protein-2
MAP3K7	Mitogen-Activated Protein Kinase Kinase Kinase 7
MAPK	Mitogen Activated Protein Kinase
MBP	Myelin Basic Protein
MEKK	Mitogen-activated protein kinase Kinase Kinase

miRNA	microRNA
MKK	Mitogen-activated protein Kinase Kinase
MMTV	Mouse Mammary Tumor Virus
Mos	Moloney sarcoma oncogene
MRCA	Most Recent Common Ancestor
Mya	Million years ago
MYC	MYeloCytomatosis oncogene
NCI	National Cancer Institute
NLK	NEMO-Like Kinase
OWL	Web Ontology Language
PID	Pathway Interaction Database
PKA	Protein Kinase A
PKC	Protein Kinase C
PLC	PhosphoLipase C
PLCD3	PhosphoLipase C, Delta 3
PNG	Portable Network Graphics
PPI network	Protein-Protein Interaction network
PPP2CA	Protein PhosPhatase 2, Catalytic subunit, Alpha isozyme
PSI-MITAB	HUPO Proteomics Standards Initiative-Molecular Interactions TAB delimited data exchange format
RAS	RAt Sarcoma
RHOA	Ras HOmolog family member A
ROCK1	RhO-associated, coiled-coil Containing protein Kinase 1
RTK	Receptor Tyrosine Kinase
RYR	Ryanodine receptor

SBML	Systems Biology Markup Language
STAT	Signal Transducers and Activators of Transcription
STP	Signal Transduction Pathway
SVG	Scalable Vector Graphic
TAK1	TGF-β Activated Kinase 1
TCF7L2	TransCription Factor 7-Like 2
TGF-β	Transforming Growth Factor-β
VEGF	Vascular Endothelial Growth Factor
Wnt	Wingless and Integrated family
XML	Extensible Markup Language

Chapter 1

Introduction and Scope of the Thesis

1.1 Introduction

A biochemical pathway is a sequence of biomolecular interactions, occurring in succession inside a cell, till the end result is achieved. Some of them are well established protocols existing in cellular environment of organisms. They are of diverse nature depending on their function, bio-molecules and type of interactions existing among the biomolecules, *i.e.*, metabolic pathways [9–12], some protein-protein interaction networks [13–17], some gene regulatory networks [18–20] and Signal transduction pathways (STPs) [1, 21, 22] among others. Metabolic pathways form a central paradigm in biology [10]. In a metabolic pathway, a principal chemical gets modified by a series of chemical reactions. Enzymes catalyze these reactions, and often require dietary minerals, vitamins and other cofactors in order to function properly [11]. These pathways can be quite elaborative because of the many chemicals (metabolites). Some traditional metabolic pathways include glycolysis, pentose phosphate pathway and the tricarboxylic acid (TCA) cycle.

In a protein-protein interaction (PPI) network, nodes represent proteins and edges represent their associations, based on experimental evidence [13]. Protein interaction networks summarize a large amount of protein-protein interaction data, both from individual, small-scale experiments and from automated high-throughput screens [17]. They provide insights into the relationships between the proteins of an organism thereby contributing to a better understanding of cellular processes [16]. Due to development in high throughput experimental procedures, genomes of many prokaryotes and a few eukaryotes are available in public domain. Finding genes, functional relationships among these genes, and their way of interactions with others are important questions that led to construction of gene regulatory networks while analyzing genomic data. In a gene regulatory network, the interaction between two genes does not necessarily imply a physical interaction, but refers to an indirect regulation at transcriptional level via proteins and metabolites that have not been measured directly [19]. They explain exactly how genomic sequences encode the regulation of expression of the sets of genes that progressively generate developmental patterns and execute the

construction of multiple states of differentiation [18].

STPs mediate the sensing and processing of stimuli. These molecular circuits detect, amplify and integrate diverse external signals to generate responses such as changes in enzyme activity, gene expression, and/or ion-channel activity [22]. They are either two-component [21, 23, 24] or multi-component systems [22,25]. Two-component signal transduction systems are one of the most prevalent means by which bacteria sense, respond and adapt to changes in their environment or in their intracellular states [21]. Multi-component STPs are non-linear, exist as complex webs, and function by serial and successive interactions among a large number of vital biomolecules and biochemical compounds [1]. In some multi-component STPs, the ligands are small lipophilic molecule that can enter inside the cell without any hindrance and bind with receptors. However, in some cases, ligands cannot enter inside a cell due to their size. They bind to transmembrane receptors and the information gets passed from one biomolecule to another in a cascade till it reaches its destination; the destination being genes/transcription factors in most of the cases. These cascades are known as STPs. They are generally named after their central controlling biomolecule(s)/ion(s), *i.e.*, MAPK [4, 26]/Ca^{2+} [2, 27, 28]/Wnt [29, 30] STP. Mechanism of a STP involves exponential amplification of signal(s) unlike other biochemical pathways, where mostly transcription factors or enzymes emanate and disseminate signal(s) in proportioned measures. They are frequently probed and prodded since their association with multiple types of human cancer was discovered. Since the present thesis deals with analysis and annotation of biochemical pathways, especially some signal transduction pathways, we shall describe them in more details.

Modularization is one of the ways for analyzing and annotating these biochemical pathways. Modularization is a process that divides a network into smaller units for better understanding and analysis of the original network. There is no single definition available for a module. We define a module as a subset of the original biochemical network, which tends to be self-sufficient in terms of biological function and has minimal dependency on the rest part of the network. Unlike studying a signaling pathway as a whole, this enables

one to study the individual modules (less complex smaller units) easily and to have a better view of the entire pathway. The justification for dividing a network into a number of modules lies in the fact that the complexity of each module is much less than that of the entire pathway and becomes an easier means of studying the entire network by parts. Thus by analyzing all the modules generated from a pathway separately, one can have a better operational view of the whole network [1, 2]. In the present thesis, we consider some STPs for modularization and subsequently analyze them from various point of views.

The proposed thesis comprises seven chapters including four contributory ones. Chapter 1 introduces the thesis while Chapter 7 concludes it. Chapter 2 includes an extensive review on MAPK, Ca^{2+} and WNT STPs as we have considered these three STPs in the subsequent chapters of the thesis. In Chapter 3, we describe an algorithm for modularization of STPs [1, 2, 4]. We partition the human Wnt STP, in Chapter 4, into multiple feasible subpathways or modules by five algorithms inspired from different concepts. The comparison among these algorithms has been done by considering two types of gene ontology supported scores, *viz.*, 'valid attribute score' and 'functional enrichment score'. Chapter 5 deals with evolution of the Wnt STPs, based on the modules obtained above, in different species. Here, we create multiple phylogenetic trees from different sets of factors associated with 48 species-specific Wnt STPs. These trees are compared with reference phylogenetic trees to find the level of similarity between evolution of a pathway and general course of evolution in multiple species. Chapter 6 emphasizes on a human Wnt diseasome (disease map). This disease map showcases comorbidity among human diseases caused by human Wnt STP genes.

1.2 Cell Signaling and Signal Transduction Pathways (STPs)

STPs reflect crucial mechanisms that allow cells to sense and respond to extracellular stimuli [31, 32]. They are integral conserved protocols of cell

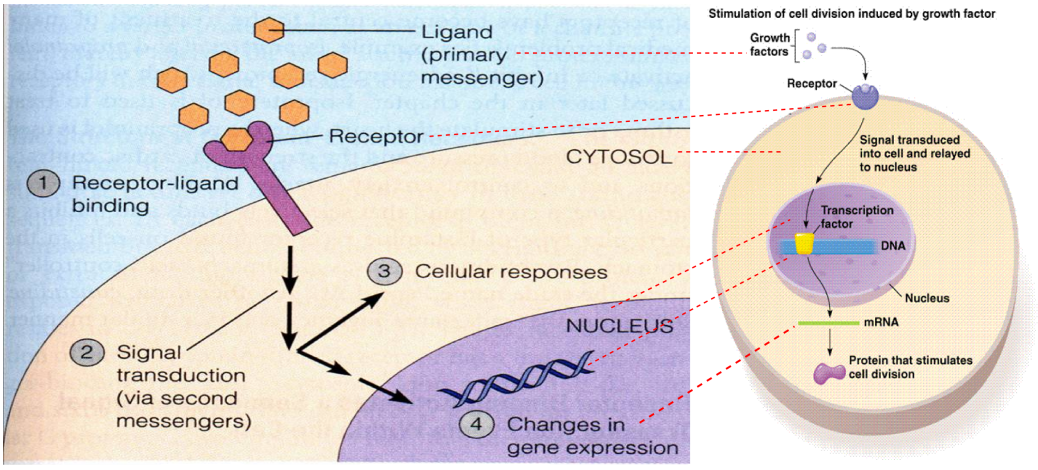


Figure 1.1: The overall flow of information during cell signaling
Cells must respond appropriately to external stimuli to survive. They respond to stimuli via cell signaling. Binding of ligand by a receptor activates a series of events known as signal transduction, which relays the signal to the interior of the cell, resulting in specific cellular responses and/or changes in gene expression.

signaling. In general signals are sensed by receptors and changed by transducers which are passed on to effectors that trigger the final response [33]. Here we describe components of a standard cell signaling mechanism (Figure 1.1). A signaling mechanism starts when ligands bind with their unique receptors.

Ligands can exert their effects from outside or enter into the cell in order to elicit a response. Some smaller ligands, *viz.*, steroids (cortisol, estrogen, progesterone), retinoids, thyroid hormone and vitamin D can easily enter into a cell [34, 35]. Generally lipophilic molecules having smaller size can cross the plasma membrane easily (Figure 1.2). They directly bind with intracellular receptors. Otherwise, signals generated outside the cell are sensed when large-sized ligands (that cannot travel inside the cell directly) bind with transmembrane receptor cells.

Receptors undergo structural changes upon sensing the signal. Internal receptors sense the signal molecules which enter into the cell. External receptors are found on the cell surface. They transmit the signal via activation of STPs. The same signal molecule can elicit different responses from different cells due to receptor diversity. Cytosolic receptors are soluble in nature and

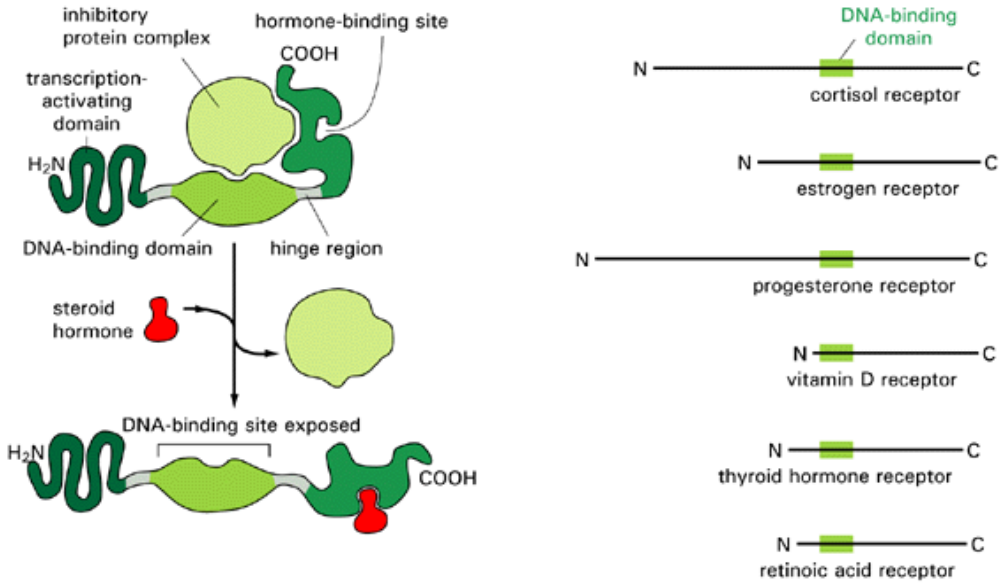


Figure 1.2: Some small ligands which can travel inside a cell
Lipophilic molecules having smaller size can cross the plasma membrane easily and directly bind with intracellular receptors.

sense intracellular signals [22, 36]. Transmembrane receptors span the cell membrane sensing extracellular signals and triggering responses inside the cell. These receptors can be classified into 3 major types [37] based on the action mechanism as shown in Figure 1.3. Ion channeled receptors trigger responses via an ion channel [38], *i.e.*, Acetyl choline receptor [39]. G-protein coupled receptors (GPCRs) represent one of the most important classes of protein due to their critical role in cell signaling in response to hormones and neurotransmitters [40–42]. GPCRs are linked with G-proteins [43] which in turn activates the other intracellular enzymes through second messengers such as cAMP and Ca^{2+} ion among others. Enzyme coupled receptors [37] activate an enzyme by coupling with it, *i.e.*, Phospholipase C, Tyrosine kinase.

These cell-surface receptor proteins act as signal transducers. They transduce the signal to intracellular environment. The signal then travels from one biomolecule to another in succession, till the desired response is not generated. Characteristics of an STP include specificity, amplification, integration,

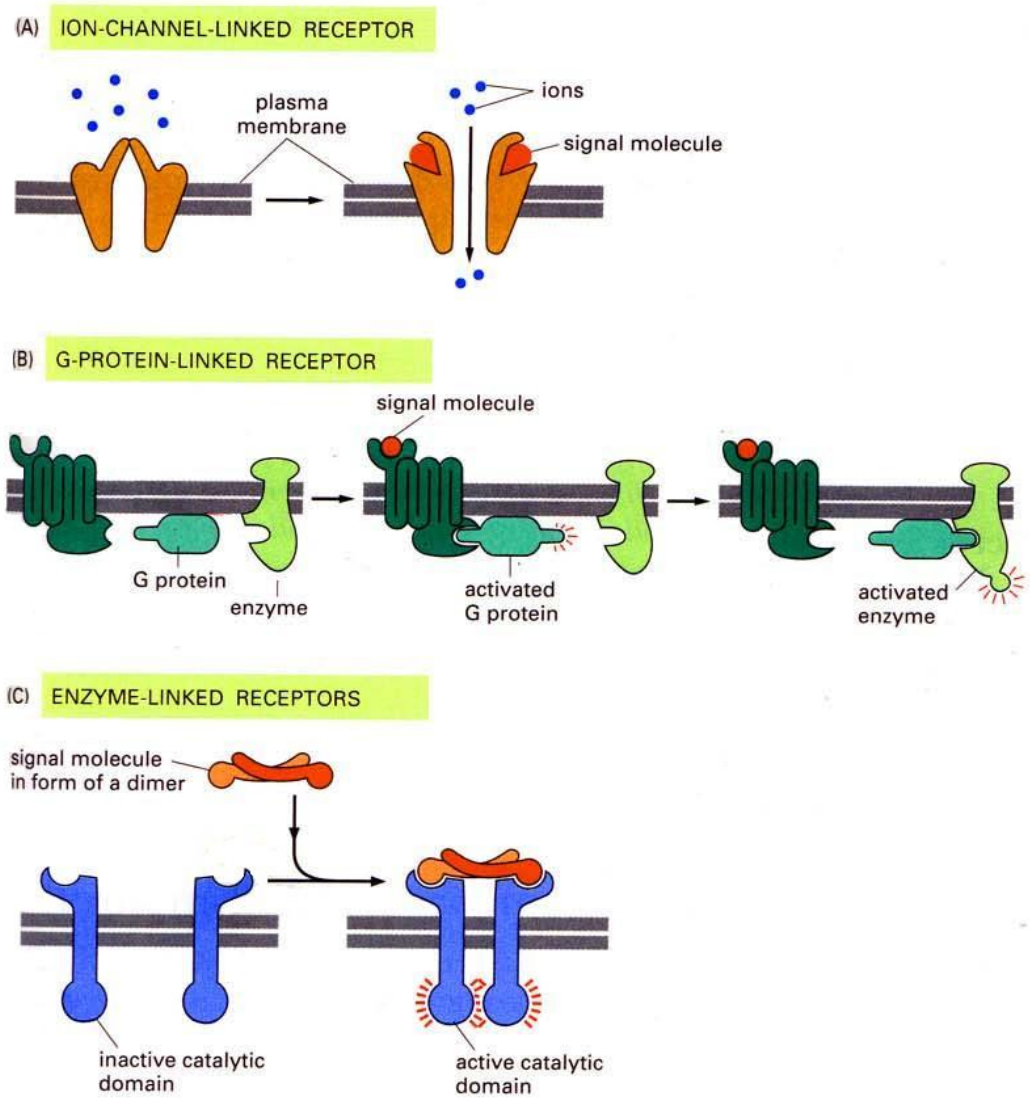


Figure 1.3: Types of transmembrane receptors
(A) Responses are triggered via ion channels. (B) GPCRs are linked with G-protein induced signaling cascades. (C) Enzyme coupled receptors [37] activate an enzyme by coupling with it.

inhibition and adaptation. They convert an extracellular ligand binding event into intracellular signals that alter the behavior of the target cell [37]. Signal molecules are highly specific to receptors. Once the receptor is bound with a ligand, the ligand triggers a change that desensitizes the receptor. So that, the receptor cannot bind with another ligand(s). This response may either remove the receptor from the surface or shut down it. When there are two or more signals of antagonistic metabolic functions, the cumulative response of the STP is the result of integration of such signals from multiple pathways. Amplifiers increase the signal strength. One molecule of signal is amplified into numerous outgoing signals (secondary signals). Examples include G-proteins, kinases and cyclases. Multiple amplifications can be found in an STP. Integrators accumulate multiple signals from different pathways to a common effector molecule. For example, phosphorylase kinase gets activated by Ca^{2+} ion as well as cAMP molecule. When either of the signal molecules is in abundance or in the presence of both molecules, the enzyme gets activated to a common response. Effectors trigger the result or final response. Just as integrators can sense multiple signals, the same signal can trigger multiple pathways. This is possible due to the presence of different effector molecules. Inhibitors block the signaling pathways. The effect of presence of an inhibitor is the same as removal of the signal molecule/inactivation of signaling. For example, phosphatases inhibit kinases; Cyclic nucleotides get hydrolyzed by phosphodiesterases. Now we briefly describe some important STPs.

1.2.1 Some important STPs

The major STPs (MAPK [4, 26, 44–46], Ca^{2+} [2, 27, 28], Wnt [29, 30], ErbB [47–49], Notch [50–54], Hedgehog [55–58], TGF-β [59–61], VEGF [62–64] and Jak STAT [65–68]) are conserved to a remarkable extent in all animals [69]. They regulate embryonic development, proliferation and differentiation and even remain active in adult organism to control differential gene expression and maintain stem cell reserves [70]. A brief description of these STPs is given here.

MAPK STP: The Mitogen-Activated Protein Kinase (MAPK) cascade is a highly conserved module that is involved in various cellular functions, including cell proliferation, differentiation, migration, division and death [4, 26, 45, 71].

Ca^{2+} STP: Ca^{2+} ion influx-outflux is a common signaling mechanism. The ion exerts allosteric regulatory effects on many enzymes and proteins in cytoplasm after entering inside a cell. Ca^{2+} ions can transduce signal after influx resulting from activation of ion channels or as a second messenger. The influx of Ca^{2+} ions from the environment or release from internal stores causes a very rapid and dramatic increase in cytoplasmic calcium concentration, which has been widely exploited for signal transduction [2, 27, 28].

Wnt STP: Wnt proteins are secreted morphogens that are required for basic developmental processes, such as cell-fate specification, progenitor-cell proliferation and the control of asymmetric cell division, in many different species and organs. Wnt STP describes a network of proteins most well known for their roles in embryogenesis and cancer, but also involved in normal physiological processes in adult animals [29, 30].

ErbB STP: The ErbB family of receptor tyrosine kinases (RTKs), otherwise known as epidermal growth factor receptors (EGFRs), facilitate binding of extracellular growth factor ligands to intracellular STPs regulating diverse biologic responses, including proliferation, differentiation, cell motility, and survival [47, 49].

Notch STP: Notch STP is essential for proper embryonic development in all metazoan organisms in the Animal kingdom. It mediates communication between neighboring cells to control cell fate decisions during embryogenesis and in postnatal life [50], and regulates lymphocyte development [52]. It is a cell-cell communication process, which allows the establishment of patterns of gene expression and differentiation, regulates binary cell fate choice and maintains stem cell populations [51].

Hedgehog STP: The Hedgehog (Hh) family of secreted signaling proteins plays a crucial role in development and morphogenesis of a variety of tissues and organs. It controls cell proliferation, differentiation, limb formation, bone differentiation, neural tube development in vertebrates and regulation of stem cell homeostasis in adult tissues [58].

TGF-β STP: The transforming growth factor-β (TGF-β) family members, which include TGF-βs, activins and bone morphogenetic proteins (BMPs), are structurally related secreted cytokines found in species ranging from worms and insects to mammals. A wide spectrum of cellular functions such as proliferation, apoptosis, differentiation and migration are regulated by TGF-β family members [59–61].

VEGF STP: VEGF STP is considered to be a crucial in both physiologic and pathologic angiogenesis. Vascular endothelial growth factor (VEGF) regulates multiple endothelial cell functions, including mitogenesis [62, 64].

Jak-STAT STP: The Janus kinase/Signal Transducers and Activators of Transcription (JAK/STAT) pathway is one of a handful of pleiotropic cascades used to transduce a multitude of signals for development and homeostasis in animals, ranging from humans to flies. In mammals, it is the principal signaling mechanism for a wide array of cytokines and growth factors.

Since the present thesis deals with MAPK, Ca^{2+} and Wnt STPs, we describe them in more details in Chapter 2.

1.2.2 STP related diseases and disorders

STPs have role in causing disease(s)/disorder(s) when member gene(s) in these pathways show abnormal behavior [70]. Here we list some disease(s) caused by the aforementioned STPs. Disease like conditions arise when STP member(s) show abnormal behavior due to aberrant expression. The diseases are listed year and author wise (delimited by ';') in decreasing manner.

- *Some MAPK STP related diseases*: Alzheimer's disease, Parkinsons disease, and amyotrophic lateral sclerosis [72]; the RAS/MAPK syndromes [73]; RASopathies [74]; diabetes, polycystic kidney, and cardio-facio-cutaneous syndrome [75]; chronic dental pain and periodontal diseases [76]; asthma, autoimmunity related diseases, rheumatoid arthritis [77]; Noonan syndrome [78]; cancer (breast [79], lung, colon, pancreas, prostate, and kidney), acute leukemia, and malignant glioma [80].

- *Some Ca^{2+} STP related diseases*: heart failure [81]; hereditary deafness [82–84]; polycystic kidney disease [85], cardiac hypertrophy, Alzheimer's disease [86,87], Parkinson's disease, amyotrophic lateral sclerosis, Huntington's disease and spinocerebellar ataxia [88]; severe combined immunodeficiency (SCID), X-linked agammaglobulinaemia (XLA), common variable immunodeficiency (CVID), and WiskottAldrich syndrome (WAS) [89]; malignant hyperthermia, central core disease (CCD), and Brody's disease [90].

- *Some Wnt STP related diseases*: human tumors [91–93]; cancers [94–99]; osteoarthritis [100]; tetra-amelia [101], bone density defects, rheumatoid arthritis [102], vascular defects in the eye, Osteoporosis-Pseudoglioma syndrome [103, 104], familial exudative vitreoretinopathy [105], non-syndromic tooth agenesis [106, 107]; RubinsteinTaybi syndrome [108]; colorectal cancer [109–111]; neuroepithelial brain tumor [112]; familial adenomatous polyposis and sporadic hepatoblastoma [113]; type II diabetes [114, 115]; kidney cancers, renal fibrosis, cystic kidney diseases, acute renal failure, diabetic nephropathy, and ischaemic injury [116]; myeloid leukemia [117]; split-hand/foot malformation [118]; odonto-onchyo-dermal hypoplasia [119]; desmoid tumor, hypertrophic scar formation, aggressive fibromatoses, and Lederhose disease [120]; coronary artery disease [121]; hepatic fibrosis, hepatoblastoma, hepatic adenoma, hepatocellular carcinoma, and cholangiocarcinoma [122]; obesity [123]; psoriatic plaques [124]; Fuhrmann syndrome, and Al-Awadi/Raas-Rothschild/Schinzel phocomelia syndrome [125]; osteoporosis [126,127]; leiomyoma [128]; synovial sarcomas [129]; High Bone Mass (HBM) trait

[130]; Van Buchem disease, autosomal dominant sclerosteosis, and osteoporosis type I syndrome [131]; skeletal malformation [132]; schizophrenia [133, 134].

- *Some ErbB STP related diseases*: glioma [135, 136]; neurodegenerative diseases (multiple sclerosis and Alzheimer's disease) and breast cancer [137]; breast, lung, and ovary cancers [138, 139].

- *Some Notch STP related diseases*: aortic valve disease [140]; T-cell acute lymphoblastic leukemia and cerebral autosomal dominant arteriopathy with sub-cortical infarcts and leukoencephalopathy [141, 142]; breast cancer [143]; cancer [144]; multiple sclerosis, Tetralogy of Fallot; Alagille syndrome, spondylocostal dysostosis, and CADASIL[1] syndrome [145].

- *Some Hedgehog STP related diseases*: osteoarthritis [146]; anaplastic large cell lymphoma [147]; myeloid leukemia [148]; Gorlin's syndrome (basal cell carcinoma [149], medulloblastoma, and rhabdomyosarcoma) and non-Gorlin's tumors (small-cell lung cancer and carcinomas of the oesophagus, stomach, pancreas, biliary tract, and Prostate) [150].

- *Some TGF-β STP related diseases*: familial thoracic aortic aneurysm syndrome and abdominal aortic aneurysm [151]; renal fibrosis [152]; Hereditary hemorrhagic telangiectasia, Loeys-Dietz Syndrome, arterial tortuosity syndrome, atherosclerosis, Marfan syndrome, Camurati-Engelmann disease, osteoporosis, sclerosteosis, Van Buchem disease, juvenile polyposis syndrome, Bannayan-Riley-Ruvalcaba and Cowden syndromes, cancer (breast [153], colorectal [154], pancreatic, lung, prostate, head and neck squamous cell [155, 156]), cleft palate, Situs Inversus, Situs Ambiguus [157]; restenosis [61].

- *Some VEGF STP related diseases*: proteinuria [158]; Atherosclerosis [159], ischemic heart disease, pulmonary hypertension and vascular restenosis [160].

[1] cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy

- *Some Jak-STAT STP related diseases*: cancer [65]; diabetic nephropathy [161]; atherosclerosis, hypertension and neointima formation [66]; asthma [162]; myocardial ischemia [163]; lymphoma, myeloma, severe combined immunodeficiency, severe congenital neutropenia and Fanconi anemia [164].

1.2.3 STP resources

Among many pathway databases[2], here, we describe some important and popularly used STP databases along with their advantages and shortcomings.

1. BioCarta[3]: BioCarta [165] contains dynamic graphical models. The online maps depict molecular relationships from various areas of active research. In an "open source" approach, this community-fed forum constantly integrates emerging proteomic information from the scientific community. It also catalogs and summarizes important resources providing information for over 120,000 genes from multiple species. Pictorial views are available in GIF[4] format.

2. INOH[5]: The Integrating Network Objects with Hierarchies (INOH) database [166] is a highly structured and manually curated database of STPs including Mammalia, *X. laevis*, *D. melanogaster* and *C. elegans*. The database focuses on curating and encoding textual knowledge into a machine-processable form. INOH ontologies, 73 signal transduction and 29 metabolic pathway diagrams (including over 6155 interactions and 3395 protein entities) are freely available in INOH XML[6] and BioPAX[7] formats.

[2]http://www.pathguide.org/

[3]http://www.biocarta.com

[4]The Graphics Interchange Format: a bitmap image format that has widespread usage on the World Wide Web due to its wide support and portability.

[5]http://www.inoh.org/

[6]Extensible Markup Language: a metalanguage that allows users to define their own customized markup languages in order to display documents on the World Wide Web

[7]Biological Pathway Exchange: a Resource Description Framework (RDF)/Web Ontology Language (OWL)-based standard language to represent biological pathways at the molecular and cellular level.

3. KEGG /PATHWAY[8]: Kyoto Encyclopedia of Genes and Genomes [167] is maintained by the Kanehisa Laboratories, Bioinformatics Center, Kyoto University and the Human Genome Center, University of Tokyo. It is a bioinformatics resource for linking genomes to life and the environment. KEGG/PATHWAY database is a collection of manually curated pathway maps. It also contains species-specific STPs (maximum number of species-specific STPs covered in any database at present), whose XML data files along with their KGML[9] and PNG[10] diagrams are publicly accessible. But some XML files are incomplete and some information of these files is not experimentally verified.

4. NetPath[11]: NetPath [168] is a collaborative effort of the PandeyLab, Johns Hopkins University and the Institute of Bioinformatics. The database is a curated resource of human STPs (10 immune and 10 cancer STPs). These 10 cancer STPs have been developed in collaboration with the Computational Biology Center, Memorial Sloan-Kettering Cancer Center and Gary Bader's laboratory, University of Toronto for the 'Cancer Cell Map'. Individual pathways can be downloaded in BioPAX level 2.0, PSI-MI version 2.5 and SBML[12] version 2.1 formats. Transcriptionally regulated genes of STPs can be downloaded in Excel and tab delimited text formats. However, no species-specific data is available in this database other than human.

5. NetSlim[13]: NetSlim [169] is a new resource that contains this 'core' subset of reactions for each pathway for easy visualization and manipulation. The pathways in NetPath contain a large number of molecules and reactions which can sometimes be difficult to visualize or inter-

[8]http://www.genome.jp/kegg/pathway.html#environmental

[9]The KEGG Markup Language: an exchange format of the KEGG graph objects, especially the KEGG pathway maps that are manually drawn and updated.

[10]Portable Network Graphics: a bitmapped image format that employs lossless data compression.

[11]http://www.netpath.org

[12]The Systems Biology Markup Language is a machine-readable language, based on XML, for representing models of biological processes. SBML can represent metabolic networks, cell-signaling pathways, regulatory networks, and many other kinds of systems.

[13]http://www.netpath.org/netslim

pret given their complexity. Unlike Netpath, NetSlim is a collection of high-confidence signaling maps generated due to more stringent curation and inclusion criteria. At present, NetSlim versions are available for 10 immune and 10 cancer STPs. All these 20 NetSlim maps are freely available for download in GenMAPP[14], GPML[15], PDF and SVG[16] formats. However, no species-specific data is available in this database other than human.

6. PID[17]: Pathway Interaction Database [170] contains biomolecular interactions and cellular processes assembled into authoritative human STPs. It contains three kinds of pathways, *i.e.*, NCI[18]-Nature curated, BioCarta and Reactome imported pathways. Pathway data is available in XML and BioPAX formats. Network visualization can be done by a third party source VISIBIOweb[19] [171] using BioPAX files. No species-specific data is available other than human.

7. Reactome[20]: Reactome [172, 173] is a curated database of biological pathways. The basic unit of the database is a reaction. They are grouped into causal chains to form pathways. The whole database is publicly available for download. Some specific datasets are also available to be downloaded in tab delimited text files, PSI-MITAB[21], and BioPAX 2 and 3 formats. However, there is no option to download the

[14]Gene MicroArray Pathway Profiler: GenMAPP is a free stand-alone computer program designed for viewing and analyzing gene expression data in the context of biological pathways.

[15]GenMAPP Pathway Markup Language: GPML is an XML-based format. It can be used to define a pathway consisting of purely graphical elements (such as lines and shapes) or graphical elements with added biological information (such as genes, proteins and datanodes).

[16]Scalable Vector Graphic: a graphic file format written in XML for images that scale smoothly to different sizes.

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

[18]National Cancer Institute

[19]http://visibioweb.patika.org

[20]http://www.reactome.org

[21]The HUPO Proteomics Standards Initiative (PSI) defines community standards for data representation in proteomics to facilitate data comparison, exchange and verification. One of its work groups deals with Molecular Interactions (PSI-MI), with MITAB being their tab delimited data exchange format.

molecular interactions of species-specific STPs.

8. Database of Cell Signaling[22] (STKE): STKE database [174, 175] provides information on the components of cellular signaling pathways and their relations with one another, which are organized into pathways, called 'Connections Maps'. These maps serve as the graphical interface of the database. Visualization of pathways can be saved as BMP[23] and SVG files. Pathway data in XML format can be availed by submitting a request form to appropriate authorities free of cost after registration. Data for a few species are available here (*e.g.*, *D. melanogaster*, *D. rerio*, *C. elegans* and *H. sapiens*).

Among the aforesaid databases, KEGG/PATHWAY database [167] provides the maximum number of easy to use species-specific pathways. These pathways have been used as raw data in our work described in various chapters of this thesis.

1.3 Partitioning Biochemical Pathways

There exist various approaches for partitioning networks/pathways. Some of these approaches are based on hierarchical clustering techniques [176–178], graph partitioning techniques [179,180], block modeling methods [181], differential equation based methods [182], cartographic representations [183] and Bayesian partition method [184]. Among them, approaches based on graph partitioning [179, 180, 185], community structure detection [186–188] and module detection [184,189–203] are popular. Algorithms based on these concepts have been used to divide, study and analyze networks. Some of these methods have been applied to biological networks also.

Graph[24] partitioning algorithms have been applied mostly to non-biological networks including a few biochemical pathways. These networks are related

[22]http://stke.sciencemag.org/cm/

[23]Bitmap Image Format: an uncompressed file format; hence, the file sizes are much larger than figures of other formats.

[24]Graph is an abstract representation of a set of objects where some pairs of the objects are connected by links. The interconnected objects are called vertices, and the links that connect some pairs of vertices are called edges.

to VLSI (Very Large Scale Integration) [204, 205], CAD (Computer Aided Design) [206, 207], Hypertext Browsing [208], geographic information services [209], parallel computing [207], integrated circuit designing [210], and some biological problems like physical mapping of DNA (Deoxyribo Nucleic Acid) [211] among others. On the other hand, a wide variety of community detection techniques have been developed based on the notion of centrality measures [212, 213], flow models [214], random walks [215, 216], resistor networks [188], and optimization [217]. Some of them have been applied to biochemical pathways.

Module detection techniques are "go-between" methods between the graph partitioning and community finding techniques. They have been applied to various kinds of biochemical pathways for elucidation of drug-disease associations [184], detection of co-evolution of cancer genes [189], mining interactome modules [190], core module biomarker identification with network exploration in breast cancer metastasis data [191], mining functional gene modules linked with rheumatoid arthritis using a SNP-SNP Network [192], identifying biochemical network modules based on shortest retroactive distances [193], inferring biologically meaningful gene modules [194], identifying cooperative functional modules [195], identifying miRNA(miRNA)-mRNA modules based on microarray data [196], identifying conserved coexpression modules among organisms [197]. They have also been used for identification of dense network modules from the yeast protein complex interaction network [198], detection of overlapping modules from protein-protein interaction networks [199], discovery of phenotype-associated protein functional modules [200], dense module search for genome-wide association studies in protein-protein interaction networks [201], inferring functional modules of protein families [202] and mining bi-sparse and cohesive modules in protein interaction networks [203]. Since the present thesis deals with some graph partitioning, community finding and modularization techniques, we provide further details on them.

1.3.1 Graph partitioning techniques

Graph partitioning techniques consider division of a set of tasks among the processors of a parallel computer so as to minimize the necessary amount of interprocessor communication [187]. In such an application, the number of processors is usually known in advance along with an approximate figure of the number of tasks that each processor can handle. Thus we know the number and size of the groups into which the network is to be split. Moreover, the goal is usually to find the best division of the network regardless of the fact whether a good division even exists or not.

Vast and complex biochemical networks have inherent non-local features that require the global structure to be taken into account in the decomposition procedure. It is important to know the naturally occurring subnetworks of a network, while studying its functionality as a whole. Holme et al. [218] have proposed an algorithm for decomposing biochemical networks into subnetworks based on the global network structure. They have analyzed full hierarchical organization of biochemical networks (metabolic and cellular networks) of 43 organisms taken from the WIT database. The investigation of Jeong et al. [219] suggests presence of the same topological scaling properties in metabolic networks that show striking similarities to the inherent organization of complex non-biological systems. Identifying recurrent patterns across multiple networks is also an important step to discover biological modules, especially from microarray datasets. Most of the existing algorithms are very costly in time and space for frequent pattern mining as the pattern sizes and network numbers increase.

Hu et al. [220] have developed a novel algorithm, called "CODENSE", to efficiently mine frequent coherent dense subgraphs across a large number of massive graphs. Unlike the other methods, this algorithm is scalable in the number and size of the input graphs, and adjustable in terms of exact or approximate pattern mining. Graph theoretical algorithms can also be used to identify backbone clusters of residues in proteins. The identified clusters show protein sites with the highest degree of interactions. Patra and Vishveshwara [221] have devised a method for identifying highly interacting

centers (clusters) in proteins. This method can be applied to the problems such as identification of domains and recognition of structural similarities in proteins. Pathway analysis of large metabolic networks meets with the problem of combinatorial explosion of pathways. Schuster et al. [222] have developed an algorithm for metabolic pathway decomposition based on local connectivity of the metabolites. Applicability of the method is analyzed with metabolic networks of *M. pneumoniae*. Several studies have also been done on networks of *E. coli* and *C. elegans* by Wagner et al. [223].

1.3.2 Community finding techniques

Community structure detection, by contrast, is used to shed light on the structure of large-scale networks, such as social networks, internet and web data, or biochemical networks [187]. Community structure detection methods normally assume that the network of interest divides naturally into subgroups, if any, and the experimenter's job is to find these groups. The number and size of the subgroups are thus determined by the network itself and not by the experimenter.

In many networks, nodes are joined together in tightly knit groups, between which there are a few loose connections. Girvan et al. [212] have proposed a method for detecting communities in such networks. They have used the idea of centrality indices to find community boundaries. Community finding algorithms can also be applied to a network of relations among genes [224]. Wilkinson and Huberman [225] have studied a network of gene co-occurrences for colon cancer accumulated from the literature, and partitioned it into communities of related genes. Their method identifies communities where the component genes of each community are related by their functions. They have designed the partitioning procedure to be particularly applicable to large networks in which individual nodes may play a role in more than one community.

Biological networks can be of different kinds. A metabolic network represents metabolic substrates and products with directed edges joining them. Protein interaction networks convey mechanistic physical interactions among

proteins [226]. Expression of a gene may be controlled by other proteins (activators and inhibitors) in a genetic regulatory network. Hence a genome can be viewed as a switching network with vertices representing the proteins and directed edges conveying dependence of protein production on the proteins at other vertices [226]. A robust approach to partition a network involves maximization of a benefit function called "modularity" over possible divisions of the network as proposed by Newman [186].

Metabolic and signaling pathways are shaped by the networks of interacting proteins whose production, in turn, is controlled by gene regulatory networks. Maslov and Sneppen [227] have quantified correlations among connectivities of interacting nodes and compared them to a null model of a network (a network with all links randomly rewired). They have found that for both protein interaction and gene regulatory networks, links between highly connected proteins are systematically suppressed, whereas those between a highly connected and lowly connected pairs of proteins are favored. This effect decreases the likelihood of cross talk between different functional modules of the cell and increases the overall robustness of a network by localizing effects of some perturbations. Stelling et al. [228] have devised a theoretical method for simultaneously predicting key aspects of network functionality, robustness and gene regulation from network structure alone. They have determined and demonstrated that the non-decomposable pathways are able to operate coherently at steady state by using *E. coli* central metabolism as an illustration.

A gene may have several connections, circuits and pathways that may crosslink and represent connected components. Guelzim et al. [229] have created a network of 909 genetically or biochemically established interactions among 491 yeast genes. After thorough analysis of the interaction network, it has been found that the number of regulating proteins per regulated gene has a narrow distribution with an exponential decay, while the number of regulated genes per regulating protein has a broader distribution with a decay resembling to a power law. As a whole, the yeast transcriptional regulatory network combines a small maximal diameter, an elevated local semi-clustering, a high number of feedback circuits and a global fragmen-

tation. Here each small connected piece indicates towards implementation of a biological function, and the global fragmentation serves to limit inter-functional crosstalk at the transcriptional level.

Clustering properties of the reaction networks can be obtained from maps of known metabolic pathways. Raine and Norris [230] have investigated random connection model, random cluster model and accumulation model for construction of metabolic networks. The random cluster and accumulation models exhibited "small-world"[25] features, in agreement with the structure of real biological networks, while random cluster and accumulation models also depict a long-tailed distribution of nodes of the original networks.

Milo et al. [231] have defined "network motifs" as patterns of interconnections occurring in complex networks. Such motifs were found in networks from biochemistry, neurobiology, ecology and engineering. The motifs shared by ecological food webs were distinct from the motifs shared by the genetic networks of *E. coli* and *S. cerevisiae*, or from those found in the world wide web. Similar motifs were found in networks that perform information processing, even though they describe elements as different as biomolecules within a cell and synaptic connections between neurons in *C. elegans*. The authors have used motifs to define universal classes of networks. It is worth to detect and understand network motifs in order to gain insight into their dynamical behavior and to define classes of networks and network homologies. Motif detection in *E. coli* transcriptional regulatory networks has also have been carried out by Shen-Orr et al. [232].

Newman et al. have proposed a series of algorithms [186–188, 212, 217, 224, 226] to find communities in various kinds of networks. These algorithms have been designed to optimize a network's divisions based on the properties of the network itself. We have compared our method with an interesting community finding algorithm of Newman [187], which has already been applied to metabolic pathways along with other kind of networks. Newman's algorithm optimizes a quality function known as "modularity" over possi-

[25]Small worlds are networks that are linked in such a way that they exhibit a high degree of clustering like ordered networks but a relatively short average number of links between any two nodes like random networks.

ble divisions of a given network. Modularity value is directly dependent on the network architecture in terms of adjacency matrix and eigenvalues of a symmetric matrix calculated from the adjacency matrix. Positive value of modularity indicates possible presence of modules in a network. One important aspect of the algorithm is that it refuses to divide a network if no good division exists. In other words, a negative value of modularity indicates no possible division of the given network.

1.3.3 Modularization techniques

A sound strategy to study Biochemical pathways is to decompose them into smaller units or modules. Such modules facilitate studying of biological processes by deconstructing complex biological networks into conceptually simple entities/modules. Module-wise presentation of biological data has its own advantages. It can discover subtle associations between biological elements that are too weak to be detected by considering all their features as a whole. Understanding how modules are organized within molecular interaction networks both physically and in terms of functional dependencies can lead to a better understanding of coordination among cellular and developmental processes. In a broad sense, a network module is a subset of nodes that forms a sub-network inside a larger network. Any subset of nodes inside a larger network can be considered as a module. Definition of network modules is often based on intuitive reasoning. While a network module in computation simply refers to a group of nodes exhibiting network proximity, in biology it refers to a certain functional units such as subset of genes, protein complexes, signaling or metabolic pathways, transcriptional programs and miRNA clusters that could independently implement a specific function as a whole .

Bayesian networks and probabilistic models are already used for identifying regulatory modules from gene expression data to identify functionally coherent modules and their correct regulators in *S. cerevisiae* [233]. Repeated random walk (RRW) based methods are also used for discovering functional modules within large-scale protein interaction networks. They can find multifunctional proteins by allowing overlapping clusters [234]. Netsplitter [235]

creates partitions progressively and the interactive visual matrix presentation allows considerable control over the process by the user, while incorporating special strategies to maintain the network integrity and minimize the information loss due to partitioning. Structural Clustering Algorithm for Networks (SCAN) finds clusters or functional modules, hubs, and outliers in complex biological networks [236]. Cartographic representation of networks can be used to find functional modules in metabolic networks [183]. MOdularized NETwork learning (MONET) draws a whole network into overlapping modules and then tries to get the global picture by integrating the learned sub-networks [237]. Deterministic Modularization of Networks (dMoNet), a new agglomerative algorithm, finds even better modules in large-scale yeast and human protein interaction networks [238].

Jayaswal et al. [196] have proposed a miRNA-mRNA module identification method. It first identifies miRNA and mRNA clusters and then estimates association between the two types of clusters. A group of miRNAs and mRNAs are denoted as a module if the miRNAs in that group regulate the mRNAs in the same group. The clustering of miRNAs or mRNAs requires two input parameters, *i.e.*, a measure of dissimilarity between miRNAs or mRNAs, and the number of clusters. Clustering is performed using the partition around medoids algorithm [239]. The clustering method is based on multivariate random forest (MRF) [240] and requires two input matrices, *i.e.*, a miRNA-mRNA map matrix and a matrix of mRNA expression values. The miRNA and mRNA clusters with statistically significant associations represent the potentially regulatory miRNA-mRNA modules. The method can be used to identify miRNA-mRNA pairs that have a negative or positive association. The method converts massive amount of information into a small number of potentially regulatory miRNA-mRNA modules, thereby, enabling biologists to identify the miRNAs and mRNAs for further experimentation.

Semantic Similarity-Integrated approach for Modularization (SSIM) integrates various gene-gene pairwise similarity values, including information obtained from gene expression, protein-protein interactions and GO annotations, in the construction of modules using affinity propagation clustering [194]. Affinity propagation is a clustering method that does not use

pre-selected centers for clustering. Instead, it generates exemplars that best represent a group of data points by considering all similarities between pairs of data points and testing all data points as potential exemplars. A group of data points that have the same exemplar can then be considered to be in a cluster. Performance of SSIM has been demonstrated using data from the osmotic shock response in *S. cerevisiae* and the prion-induced pathogenic mouse model. SSIM approach reveals the hierarchical structure of gene modules to gain a broader functional view of the biological system. It can facilitate comprehensive and in-depth analysis of high throughput experimental data at the gene network level. It can quickly infer gene modules with coherent biological meaning and thereby accelerate systems biology studies in complex diseases.

Sridharan et al. [193] have identified modules from biochemical networks based on shortest retroactive distances. They have focused on cyclical interactions based on the rationale that reactions that mutually influence each other belong to the same group. They have used "Shortest Retroactive Distance (ShReD)", to characterize the connectivity between two vertices that interact retroactively. The retroactive interaction represents a mechanism for mutual feedback, and thus expresses interdependence. Modules have been created by a modified version of Newman's community finding algorithm [187]. The authors have identified closely interacting groups of biochemical reactions as a module by recognizing the modular hierarchy inherent in biochemical networks. "conserved modules across organisms (COMODO)[26]" uses an objective selection criterion to identify conserved expression modules between two species [197]. The method uses as input microarray data and a gene homology map and provides as output pairs of conserved modules and searches for the pair of modules for which the number of sharing homologs is statistically most significant relative to the size of the linked modules.

Iterative Network Partition (iNP) has created modules in yeast protein complex network and breast cancer gene co-expression network [198]. A

[26]http://homes.esat.kuleuven.be/~kmarchal/Supplementary_Information_Zarrineh_2010/comodo/index.html

Bayesian partition method "comCIPHER[27]" has identified drug-gene-disease co-modules [184]. The method comCIPHER first partitions the genes in the closeness profiles into different gene modules. Then, in each gene module, it partitions the drugs and diseases into two categories, *i.e.*, those that are associated with the gene module, and those that are not. Drugs and diseases associated with the same gene module as well as those genes themselves form a co-module. The authors have applied comCIPHER to a set consisting of 723 drugs, 275 diseases and 1442 genes that lead to detection of 86 significantly enriched drug-gene-disease co-modules. Among them, 24 co-modules have showcased new drug-disease associations. comCIPHER is a novel way of presenting drug-gene-disease relationships. But the method has some shortcomings. It can focus only on small subset of genes, though this may lose some useful information. Large dimensions have led to computational difficulties. Moreover, when the predefined module number is chosen large, comCIPHER tends to split the data into smaller blocks and increases the precision at the cost of increased computational load.

Zhu et al. [189] have found subnetworks in the human protein-protein interaction network within which all proteins have evolved at similar rates since the human and mouse split. Identified at a given co-evolving level, the subnetworks with non-randomly large sizes have been defined as co-evolving modules. The proteins within the detected modules tend to be conserved, evolutionarily old and enriched with housekeeping genes, while proteins outside modules tend to be less-conserved, evolutionarily younger and enriched with genes expressed in specific tissues. The overall conservation of cancer genes has been mainly attributed to the cancer proteins enriched in the conserved modules. One of the demerits of the greedy search algorithm is that some modules have been found to share proteins and their boundaries could not be clearly defined.

MOfinder[28] has detected overlapping modules in yeast and human protein-protein interaction (PPI) networks [199]. MOfinder converts the PPI file into a sparse matrix and performs a global approximate minimum degree order-

[27]http://bioinfo.au.tsinghua.edu.cn/comCIPHER/
[28]http://bsb.kiz.ac.cn/mofinder/

ing (AMD) of the sparse matrix. In global AMD the densely connected elements (module) get clustered along the diagonal followed by a local AMD to give local approximate minimum degree ordering. MOfinder uses a sliding window along the diagonal to fetch the local sparse matrix and make the local AMD. If the clustering coefficient value of the submatrix in the sliding window is not less than the cut-off, MOfinder saves the submatrix as a module. The sliding window then moves one step along the diagonal to find new modules, and the iteration process gets repeated until the sliding window reaches the end. Lastly, MOfinder removes redundant modules (if module A is included in module B, A is removed) and saves the results. MOfinder allows flexibility and user customization with adjustable parameters. It is fast in practice for large networks. With less runtime requirement, it can meet the need of biological analysis. However, Mofinder detects small-sized modules. The reason is its stringent clustering coefficient cut-off value. Setting the clustering coefficient cut-off value to a small value can increase the number of detected modules especially loosely connected modules. In that case, establishing their biological significance would be difficult.

Hendrix et al. [200] have introduced a fast and theoretically guaranteed method, called DENSE[29] (Dense and ENriched Subgraph Enumeration). Given a phenotype-expressing organism, the DENSE algorithm tackles the problem of identifying genes that are functionally associated with a set of known phenotype-related proteins by enumerating the "dense and enriched" subgraphs in genome-scale networks of functionally associated or interacting proteins. DENSE requires the user input of two parameters, *i.e.*, the enrichment (μ) and the density (γ). A higher value of γ will produce more connected (clique-like) subgraphs. A higher value of the enrichment ($\mu \geq 0.5$) will produce subgraphs that are primarily composed of the "query" vertices, whereas a very low value ($\mu \leq 0.001$) will result in enumeration of all the subgraphs that satisfy the γ threshold and contain at least one query vertex. A "dense" subgraph is defined as one in which every vertex is adjacent to at least some γ percentage of the other vertices in the subgraph where $\gamma > 50\%$, which corresponds to a set of genes with many strong pairwise protein

[29]http://www.freescience.org/cs/DENSE/

functional associations. The researchers' prior knowledge is incorporated by introducing the concept of an "enriched" dense subgraph in which at least μ percentage of the vertices are contained in the knowledge prior query set. Genes contained in such dense and enriched subgraphs, or μ-enriched, γ-dense quasi-cliques, have strong functional relationships with the previously identified genes, and so are likely to perform a related task. When applied to the protein functional association network of *C. acetobutylicum* ATCC 824, obtained from STRING[30] [241] database, DENSE has been able to predict known and novel relationships, including those containing regulatory, signaling, and uncharacterized proteins. It is able to efficiently calculate dense and enriched subgraphs on large and sparse graphs with a power-law degree distribution.

Dense Module Searching (DMS) method [201] identifies candidate subnetworks or genes for complex diseases by integrating the association signal from genome wide association study (GWAS) datasets into the human PPI network. A module is defined as a subgraph within the whole network with a locally maximum proportion of low-P-value genes. DMS extensively searches for subnetworks enriched with low P-value genes in GWAS datasets. Compared with pathway-based approaches, this method introduces flexibility in defining a gene set and can effectively utilize local PPI information. DMS has successfully identified a set of significant modules and candidate genes, including some well-studied genes not detected in the single-marker analysis of GWA studies in two GWAS datasets for complex diseases, *i.e.*, breast cancer and pancreatic cancer. However, the proposed strategy has been more specifically designed for GWAS, thus making it a more direct application possible for geneticists, but not for all kinds of biochemical pathways.

Konietzny et al. [202] have developed a new Bayesian method, based on a probabilistic topic model, for directly identifying functional modules of gene or protein families using co-occurrence patterns in a collection of annotated sequence samples. In particular, they have used a Bayesian method known as Latent Dirichlet Allocation (LDA). Topic models have been used in text mining applications to reveal statistical relationships among the words in

[30]http://string-db.org/

collections of text documents, because it has been observed that strong relationships usually correlate well with semantic agreement of words. They have defined a functional module as a set of proteins that jointly participate in a biological process. The method has explored co-occurrence patterns of protein families across a collection of sequence samples to infer a probabilistic model of arbitrarily-sized functional modules. It is capable of simultaneously processing a large number of genome or metagenome annotations. On the other hand, it is a well known fact that prediction methods for functional relationships that rely on conserved genomic context are prone to false positive predictions if pseudo-genes are involved. Hence, LDA, like other cooccurrence techniques, could principally be misguided in cases where multiple copies of an orthologous group reside in the same genome by chance without being retained by selection for a certain functionality, *e.g.* due to repeats of a genomic sequence.

BinTree Seeking (BTS)[31] [203] mines bi-sparse[32] and cohesive modules in protein interaction networks based on Edge Density of Module (EDM) and matrix theory. BTS detects modules by depicting links and nodes rather than nodes alone, and its derivation procedure is totally performed on adjacency matrix of networks. It automatically determines the number of modules in a protein interaction network. However, there is a need to decrease the computational complexity of BTS. For analyzing a protein interaction network with more than 10,000 nodes, it could take several days for execution at present state, depending on the configurations of computation platform.

While there is a general agreement that a biochemical module should represent a group of connected network components, there is less consensus on the criteria that should be used to systematically extract biologically meaningful modules. They can be derived by the aforesaid algorithms and methods. In addition, they can be mathematically derived from structure of the whole network under consideration [242]. They can be created based on clearly defined input, output, pathway and cellular compartments [243], spa-

[31] http://www.csbio.sjtu.edu.cn/bioinf/BTS/

[32] A bi-sparse module is sparsely connected internally and densely connected with other bi-sparse or cohesive modules.

tial locations in cytoplasm, as defined by protein scaffolds and anchors [244], absence of retroactivity [245] and classification of Petri net t-variants [246]. Sometimes modules are also based on functional input-output relationships and operational boundaries that may not always correspond to conventional cell boundaries. Thus division of a biological reaction network into smaller units highly facilitates its investigation.

1.4 Scope of the thesis

The proposed thesis comprises seven chapters including four contributory ones. Chapter 1 introduces the thesis while Chapter 7 concludes it along with a brief note on scopes for further research. We now describe in brief the content of the remaining chapters. Chapter 2 includes an extensive review on MAPK, Ca^{2+} and Wnt STPs to maintain fluidity of the succeeding chapters.

1.4.1 Modularization Algorithm [1–4]

In Chapter 3, we describe an algorithm for modularization of MAPK, Ca^{2+} and Wnt STPs [1, 2, 4]. Modularization is a process which divides a network into smaller units for better understanding and analysis of the original network. There is no single definition available for a module. We define a module as a subset of the original biochemical network, which tends to be self-sufficient in terms of biological function and have minimal dependency on the remaining part of the network. Unlike studying a signaling pathway as a whole, this enables one to study the individual modules (less complex smaller units) easily and enables a better study of the entire pathway. The justification for dividing a network into a number of modules lies in the fact that the complexity of each module is much less than that of the entire pathway and is an easier means of studying the network by parts. Thus analyzing all the modules separately generated from a pathway, one can have a better operational view of the whole network [1, 2].

1.4.2 Comparison of Some Partitioning Algorithms [1, 5]

We partition the human Wnt STP, in Chapter 4, into multiple partitions or modules by five algorithms, inspired from different concepts, for finding the best partitioning algorithm among them. Greedy [247], Farhat's [248],and Kernighan-Lin's [249] algorithms are graph partitioning techniques. Newman's community finding algorithm [187] is dedicated towards finding communities in networks. Modularization algorithm [1] detects functional modules in biological networks. A comparative study has been done among partitions created by these algorithms by considering, *viz.*, 'valid attribute' and 'functional enrichment' scores (Subsection 4.4). Based on these scores, the Modularization algorithm has been found to create good partitions from the human Wnt STP. Later modules of 31 species-specific Wnt STPs are studied and compared for detection of conserved modules.

1.4.3 Deriving Phylogenetic Trees from Modules [6, 7]

In Chapter 5, we determine phylogenetic properties of 48 species-specific Wnt STPs. A phylogenetic comparison among species-specific Wnt STPs, taking their various factors into account, is expected to throw light on their evolution which is not explicitly known yet. For this purpose, we create alternative phylogenetic trees from topology and modules. They are compared with two reference phylogenetic trees (NCBI taxonomy[33] taxonomy tree and 18S rRNA tree) to find the level of similarity between evolution of a pathway and general course of evolution in multiple species. The tree bearing maximum similarity with the reference trees will naturally be the best tree to represent Wnt STP phylogeny. By comparison, it is found that the module tree have maximum similarity with the reference phylogenetic trees. Hence, considering modules alone or together with other factors will prove to be beneficial while creating phylogenetic trees from biological pathways and studying their evolutionary patterns. The module tree shows conservation among species belonging to

[33]http://www.ncbi.nlm.nih.gov/Taxonomy/CommonTree/wwwcmt.cgi

phylum Chordata as well as Arthropoda when compared with salient features of Wnt evolution in major phyla.

1.4.4 A Wnt Diseasome: Modules in the Diseasome [8]

Chapter 6 emphasizes on creating a human Wnt STP related diseasome. A diseasome is a combined set of all known disorders or gene associations in a species. It is created by linking the complete set of genetic disorders with the complete list of disease genes [250]. Disease maps are potential knowledge bases that throw light on multiple disease related complicacies. The more connected a disease is to others, higher is its prevalence and associated mortality rate (otherwise known as comorbidity) [251]. For example, certain diseases like diabetes, obesity, Gaucher disease and Parkinson's disease often co-occur in the same individual. Diseasome-wise studies are needed to understand such situations. In this chapter, we create a disease map from human Wnt STP by linking diseases having common causative genes. We associate 112 diseases to 48 genes of human Wnt STP to create a gene-disease network and derive a diseasome of 823 disease associations. The diseasome portrays comorbidity among human diseases associated with Wnt STP gene(s).

1.5 Conclusive remarks

Detecting meaningful modules from a biochemical pathway is an important task in Bioinformatics. In this thesis, we propose a new partitioning algorithm, *i.e.*, Modularization algorithm in Chapter 3. We apply this algorithm to some real life pathway data, *i.e.*, MAPK, Ca^{2+} and Wnt STPs, and analyze their modules. Review of these pathways are furnished in Chapter 2. Superior performance of the Modularization algorithm is established in Chapter 4 by comparing it with some traditional graph partitioning and community finding algorithms. As an application of this algorithm, pathway phylogeny is established from species-specific modules in Chapter 5. In Chapter 6, we create a human Wnt Diseasome.

We start the next chapter with a review on MAPK, Ca^{2+} and Wnt

STPs. A detailed description regarding discovery, pathway structure, biological function and association with diseases of these pathways is also furnished.

Chapter 2

A Brief Review of MAPK, Ca^{2+} and Wnt STPs

2.1 Introduction

In this chapter, we review MAPK, Ca^{2+} and Wnt STPs discovered in the late 20th century. It may be mentioned here that MAPK, Ca^{2+} and Wnt STPs are considered as pathway data in Chapter 3; while all the other chapters (Chapters 4, 5 and 6) deal with Wnt STP for analyzing the methodology developed in those chapters. The review includes their discovery, pathway structure and biological functions in cellular environment and their role/association in causing human diseases and disorders.

2.2 MAPK STPs

The first available MAPK genes are KSS1 and FUS3 [252] in budding yeast, and ERK1, ERK2 and ERK3 in mammals [253]. They were named so because of Myelin Basic Protein (MBP) [254] and Microtubule-Associated Protein-2 (MAP2) [255]. These proteins were used then to measure ERK1 and ERK2 activity. The MAP acronym was retained later, but with a different meaning. The name Mitogen-Activated Protein Kinase (MAPK) was assigned to these enzymes to acknowledge the fact that they had first been detected as mitogen-stimulated tyrosine phosphoproteins in the early 1980s [256]. Later, the name MAPK evolved into the family name for a growing number of related kinases [257]. MAP kinases are proline-directed serine/threonine kinases that are activated by dual phosphorylation in response to diverse extracellular stimuli [72].

2.2.1 MAPK pathway structure

MAPK pathway (Figure 2.1) is one of the most ubiquitous signal transduction systems [71]. It is characterized by the general path, "Stimulus > MAPKKK > MAPKK > MAPK > Response", where MAPKK is the kinase of MAPK and MAPKKK is the kinase of MAPKK. The symbol "$A > B$" stands for "A stimulating B" [72]. In most of the cases, MAPKKK is activated by small G proteins such as Ras and Rap1 [26, 45]. The sequential activation

of the MAPK cascade eventually results in the activation of transcription factors, phospholipases or cytoskeletal proteins, microtubule-associated proteins and the expression of specific sets of genes in response to environmental stimuli. MAPK pathway is conserved in all eukaryotes and plays a key role in regulation of gene expression as well as cytoplasmic activities. It is even active in plants. For example, the Arabidopsis genome consists of 23 MAPKs, 12 of which are ERK type and the others are plant specific [258].

In general MAPK STPs comprise the classical 'MAPK', 'JNK and p38', and 'ERK5' pathways [259]. The JNK family of MAPKs, also known as the stress-activated protein kinase 1 (SAPK1) family, includes the widely expressed $p46^{JNK1}$ and $p54^{JNK2}$ as well as the brain-specific $p49^{JNK3}$. JNKs undergo MKK4- or MKK7-mediated dual phosphorylation at their Thr-Pro-Tyr motif to achieve activation. The p38 MAPK family (also known as SAPK2 family) includes four isoenzymes (α, β, γ and ρ), which are primarily activated by MKK3- or MKK6- mediated dual phosphorylation of their Thr-Gly-Tyr motif. ERKs are activated by dual phosphorylation of a Thr-Glu-Tyr motif, and consist of $p44^{ERK1}$, $p42^{ERK2}$, $p110^{ERK5}$ and $p60^{ERK7}$ MAPKs [80].

2.2.2 Some biological functions of MAPK STPs

MAPK STPs promote cell growth, differentiation, stress response, T-cell development, inflammation and apoptosis in mammals. Signaling by MAPKs affects activity/localization of individual proteins, transcription of genes, increased cell cycle entry, and promotes changes that orchestrate complex processes such as embryogenesis and differentiation. MAPKs, including JNK, p38 and Erk, play crucial roles in cell migration [260]. These enzymes mediate acute responses to hormones such as changes in membrane permeability, cell motility and transcription of immediate early genes; homeostatic responses of intermediate duration such as stimulus-induced long term potentiation in neurons; and sequenced programs required for animal development [75]. They transduce a large variety of external signals; leading to a wide range of cellular responses, including mating, filamentation, high osmo-

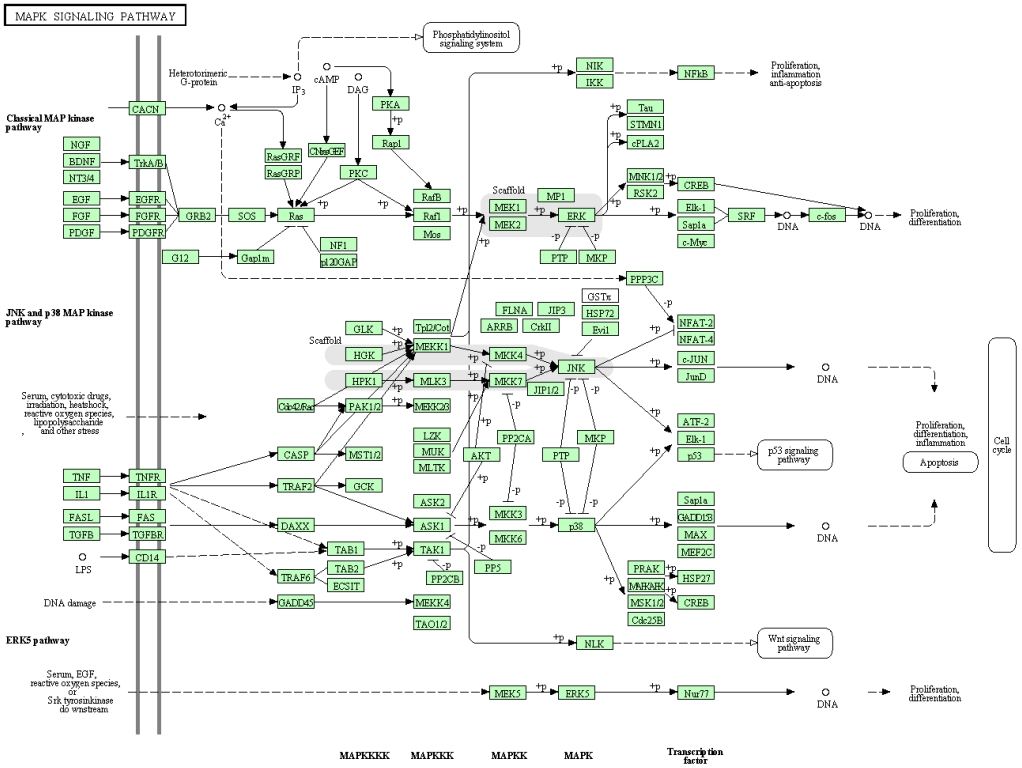


Figure 2.1: Human MAPK STP as given in KEGG
The Mitogen-Activated Protein Kinase (MAPK) cascade is a highly conserved module that is involved in various cellular functions, including cell proliferation, differentiation, migration, division and death.

larity responses, cell wall remodeling and sporulation in *S. cerevisiae* [261]; morphogenesis and spatial patterning of *D. amoebae*; eye development in *D. melanogaster*, vulva induction in *C. elegans* [45].

2.2.3 Role of MAPK STPs in some diseases and disorders

The same components of MAPK STPs act differentially in the pathogenic mechanisms of many human diseases. MAPK STPs are thought to contribute to Alzheimer's disease pathogenesis [72] through various mechanisms including induction of neuronal apoptosis, transcriptional and enzymatic activation of β and γ-secretases, phosphorylation and stabilization of amyloid precursor protein by JNK. These pathways also contribute to neuroinflammatory

responses and neuronal death in the pathogenesis of Parkinson's disease [72]. Aberrant expression and activation of p38 MAPK in motor neurons and microglia are thought to be important for Amyotrophic lateral sclerosis progression [72]. In breast cancer development, upregulation of the Ras-MAPK signaling can occur through multiple facets that activate downstream chromatin targets to activate the Ras-MAPK pathway. The activation of the Ras-MAPK pathway generates a plethora of responses in breast cancer tumors and cell lines [79].

Noonan syndrome, an autosomal dominant disorder that features short stature, facial dysmorphia, pulmonary valve stenosis, pectus deformities and webbed or short neck, is caused by dysregulated RAS-MAPK signaling [78]. Noonan syndrome, Costello syndrome and Cardio-Facio-Cutaneous (CFC) syndrome are autosomal dominant multiple congenital anomaly syndromes characterized by a distinctive facial appearance, heart defects, musculocutaneous abnormalities, and mental retardation. They are comprehensively termed as "the RAS/MAPK syndromes" including LEOPARD syndrome for their occurrence due to disorders in mutations of molecules of the RAS/MAPK pathway [73]. LEOPARD syndrome is characterized by multiple lentigines, electrocardiographic conduction abnormalities, ocular hypertelorism, pulmonary stenosis, abnormal genitalia, retardation of growth, and sensorineural deafness. Later, these disorders along with hereditary gingival fibromatosis, neurofibromatosis type 1, arteriovenous malformation, autoimmune lymphoproliferative syndrome, Legius syndrome are termed as "RASopathies'; Ras/MAPK pathway dysregulation being their causative agent [74].

2.3 Ca^{2+} STP

The enormous importance of Ca^{2+} ions in a cell was discovered in 1883 [262]. Then the inositol 1,4,5-trisphosphate ($InsP_3$) was discovered as a Ca^{2+}-releasing agent [263] followed by Ca^{2+} oscillations in non-excitable cells [264]. Since then the study of Ca^{2+} STP has acquired the status of an important field of research [265]. Calcium STPs are very peculiar in nature. When there is an extracellular change, cells get the message either by introduc-

tion of Ca^{2+} ions into cytoplasm or by its evacuation to outside through ion channels. This mechanism is aided by organellar, and nuclear storage of Ca^{2+} ions. Normal intracellular Ca^{2+} level is 10^{-7}M, much lower than the extracellular concentration of 10^{-3}M, and lower than that of individual organellar concentrations. Cytoplasmic Ca^{2+} ion concentration must be maintained at low levels as it precipitates phosphate, which is the standard energy currency of cells. Prolonged high intracellular calcium levels even lead to cell death. Hence, cells first evolved techniques for decreasing effect of free calcium ions towards cytosol which later is used as well for signal transduction across and inside the cell [28] as seen in Figure 2.2.

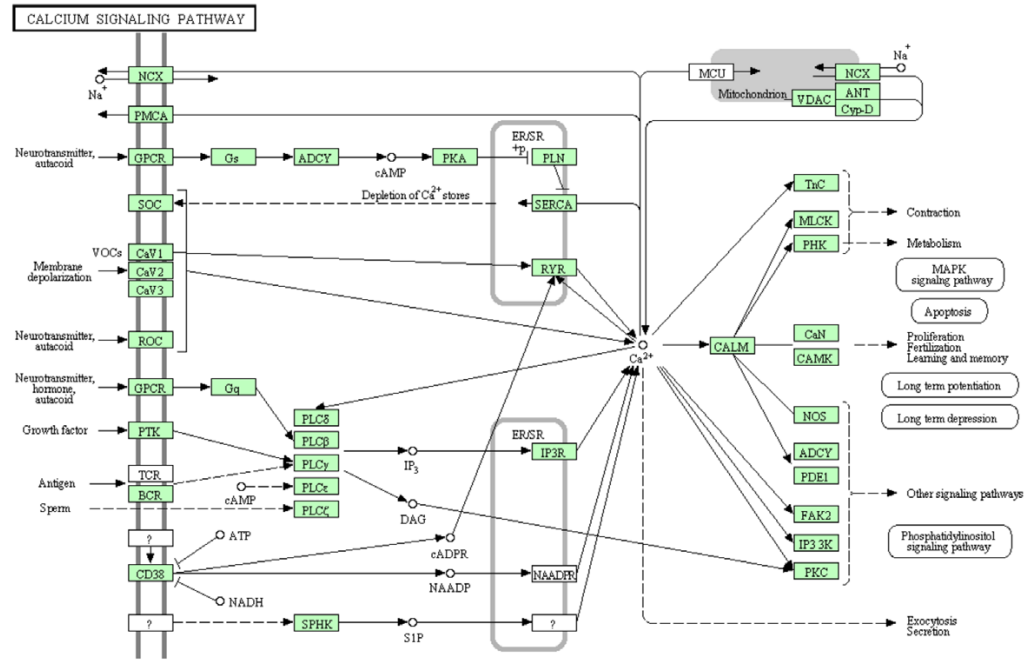


Figure 2.2: Human Ca^{2+} STP as given in KEGG
The influx of Ca^{2+} ions from the environment or release from internal stores causes a very rapid and dramatic increase in cytoplasmic Ca^{2+} ion concentration, which has been widely exploited for signal transduction.

2.3.1 Ca^{2+} pathway structure

Intracellular free calcium ion concentration gets maintained by calcium binding proteins and calcium pumps (Figure 2.3) present in unit membranes. Calcium binding proteins (buffer and trigger proteins) bind with free calcium ions when present in abundance, and maintain its low level in cytoplasm. Buffer proteins bind with free calcium ion(s) as its concentration goes up in intracellular environment (*e.g.*, Calsequestrin). On the other hand, trigger proteins bind with free calcium ion(s) and change their confirmation to modulate enzymes and ion channels, *e.g.*, Calmodulin [266, 267]. But these proteins serve for temporary solution. Calcium pumps are necessary to maintain low cytosolic calcium level against the highly calcium rich extracellular environment [268]. Different kinds of calcium pumps are present in plasma membrane, mitochondrial membrane and smooth endoplasmic reticular membrane [28, 269, 270]. Many intracellular STPs rely on elevated Ca^{2+} ion concentration as an important indicator [271]. In addition, intercellular Ca^{2+} wave propagation may act to coordinate the response of nearby cells in the tissue, leading to the concerted response of the tissue [272]. Most interestingly Ca^{2+} STP can have local as well as global effects depending on the concentration and mode of Ca^{2+} entry inside a cell. When signal travels within submicromolar range, the concerned STP is essentially local. Information transferred in these pathways does not effect global changes like gene transcription, and it terminates in the vicinity of signal origin [273]. Nuclear and cytosolic Ca^{2+} signals can have effects that are independent of one another [274].

Behavior of calcium ions can be studied by calcium ion sensitive fluorescent indicators (aequorin/fura-2) inside a cell. Local openings of individual (or small groups of) calcium release channels in endoplasmic reticulum represent small and localized signal, which is seen in one or more discrete regions of cell. They are called as calcium blips, quarks, puffs or sparks, and represent elementary calcium signaling units. When extracellular change is strong and persistent, this localized signal can propagate as a regenerative calcium wave through cytosol. It is known as calcium spike. One such spike may

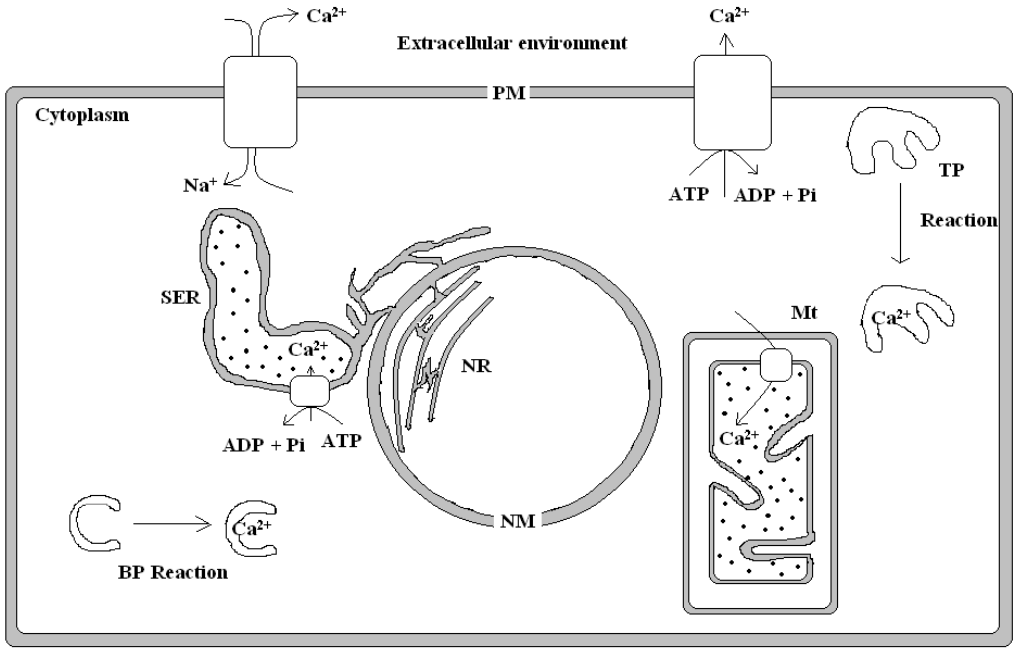


Figure 2.3: Diagrammatic view of various mechanisms for $[Ca^{2+}]_i$ balance in a cell

Calcium ion influx-outflux is a common signaling mechanism. Once it enters the cytoplasm it exerts allosteric regulatory effects on many enzymes and proteins. Ca^{2+} ions can transduce signal after influx resulting from activation of ion channels or as a second messenger. The influx of Ca^{2+} ions from the environment or release from internal stores causes a very rapid and dramatic increase in cytoplasmic calcium concentration, which has been widely exploited for signal transduction. [M - Plasma Membrane, SER - Smooth Endoplasmic Reticulum, Mt - Mitochondria, TP - Trigger Protein, BP - Buffer Protein, NM - Nuclear membrane, NR - Nucleoplasmic reticulum]

be followed by a series of further spikes (each one lasting for some seconds). These oscillations can persist as long as receptors are activated at the cell surface. Ca^{2+} behavior is peculiar in the sense that "released calcium ions initially stimulate more self-release, a process known as calcium-induced calcium release, but as its concentration gets high, inhibition of further Ca^{2+} release prevails" [269].

2.3.2 Some biological functions of Ca^{2+} STP

Ca^{2+} gradients within cells have been found to initiate cell migration, exocytosis, lymphocyte killer cell activity, acid secretion, transcellular ion transport, neurotransmitter release, gap junction regulation and numerous other

functions [275]. Calcium is essential for cell growth and survival. It influences the cell cycle in more than one way. That is, depletion of the $InsP_3$ receptor gated Ca^{2+} pool results in cell cycle arrest at G_0/G_1 and S phases. Ca^{2+} is necessary and sufficient for resumption of meiosis in marine eggs. A spike of Ca^{2+} triggers completion of meiosis and initiation of mitosis [276, 277]. It has also been found that the process of gene transcription depends on how Ca^{2+} enters into a cell. Ca^{2+} entry through voltage-dependent L type Ca^{2+} channels and N-methyl-D-aspartic acid (NMDA) receptors initiates gene transcription through distinct DNA-regulatory elements [278]. Cellular Ca^{2+} levels quantitatively correlate with level of expression of transcription factors in single cells [279]. Calreticulin, a molecule previously thought to be a buffer protein, appears to regulate the glucocorticoid nuclear hormone receptor [280]. Increase in intranuclear Ca^{2+} initiates gene expression and cell cycle progression, but also can activate degradative processes in programmed cell death or apoptosis. Prolonged existence of high Ca^{2+} ion concentration activates nucleases that cleave DNA and degrade cell chromatin. Ca^{2+} dependent proteases, phosphatases and phospholipases break DNA, resulting in a loss of chromatin structural integrity [281]. A variety of $InsP_3$-dependent Ca^{2+} signals coded in time, amplitude and space occur during the process of cardiac cell differentiation and heart development [282]. With such a versatile function list, we have deemed this pathway fit as a pathway data.

2.3.3 Role of Ca^{2+} STP in some diseases and disorders

Perturbed cellular calcium homeostasis plays a prominent role in the pathogenesis of Alzheimer's disease [87]. Recent evidences indicate that neuronal Ca^{2+} STP is abnormal in not only Alzheimer's disease, but also in a wide range of neurodegenerative disorders, such as Parkinson's disease, amyotrophic lateral sclerosis, Huntington's disease and spinocerebellar ataxia [88]. Alterations in calcium homeostasis are also associated with heart failure. A complex array of mechanisms regulate intracellular free calcium levels in the heart and vasculature, and a failure in these systems to maintain normal Ca^{2+} homeostasis has been linked with both heart failure and hyper-

tension [81]. Ca^{2+} signals in congenital immunodeficiency syndromes. In the immune system, impaired Ca^{2+} STP in T and B cells has been linked with several inherited immunodeficiency diseases, *i.e.*, Severe Combined ImmunoDeficiency (SCID), X-linked Gammaglobulinaemia (XLA), Common Variable ImmunoDeficiency (CVID) and WiskottAldrich Syndrome (WAS) [89].

A leak in the Ca^{2+}-releasing ryanodine receptor of sarcoplasmic reticulum explains symptoms of malignant hyperthermia. Malignant hyperthermia is a genetic disease in which inhalational anesthetics induce skeletal muscle rigidity and extreme hyperthermia (among other symptoms). Nearly 30 mutations linked with the RYR1 gene are associated with this defect. Central core disease (a myopathy usually associated with malignant hyperthermia, characterized by hypotonia, proximal muscle weakness, and lack of oxidative or phosphorylase activity in the central "core" of type I and type II muscle fibers) is also linked with mutations in the RYR1 gene. Brody's disease is characterized by muscle cramping and exercise-induced impairment of relaxation. It is inherited as an autosomal recessive trait that is linked with defects in the gene SERCA1 in some of the cases [90].

Localized increase in $[Ca^{2+}]_i$ can modulate a variety of subcellular processes. Some of these processes are targeted fusion of specialized cytosolic vesicles with the plasma membrane, or activation of kinase cascades leading to cell type specific gene expression. Either or both of these processes could mediate the function of the polycystin pathway. Hence, it is believed that defects in this STP can form the basis of cyst formation in polycystic kidney disease [85]. Defects in PMCA pump are identified in the form of a hereditary deafness. The defect concerns the PMCA2 pump, which is abundantly expressed in the brain, particularly in the Purkinje cell of the cerebellum and in the hair cells of the corti organ of the inner ear [82, 283].

2.4 Wnt STPs

The name 'Wnt' was initially coined after two homologous genes found in fruitfly and mouse, namely, 'Wingless' and 'INT1' [284]. Wingless is a *D. melanogaster* mutant lacking wings [285]. Recessive mutation in the wingless

gene affects wing and haltere (either of the rudimentary hind wings used for maintaining equilibrium during flight) development in *D. melanogaster* [286]. The 'INT' genes were originally identified as vertebrate genes near several integration sites of Mouse Mammary Tumor Virus (MMTV) [287, 288]. The wingless gene was subsequently characterized as segment polarity gene in fruitfly, which functions during embryogenesis [289] and adult limb formation [290]. But, now it represents a family of several genes belonging to different species, including 19 in human.

2.4.1 Wnt pathway structure

Wnt molecules are secreted cysteine-rich, lipid-modified glycoproteins. They bind to FZD receptors along with co-receptor LRPs and initiate the downstream steps, those altogether are known as Wnt STP [29] as seen in Figure 2.4. Initially it was known that Wnt proteins bind to FZDs. But by 2001, it has become evident that LRPs also function as Wnt co-receptors along with the FZDs [29, 291]. Wnt STP was then recognized as diversified into 4 branches: (i) the β-catenin (canonical Wnt) pathway, (ii) the planar cell polarity pathway, (iii) Wnt/Ca^{2+} pathway and (iv) a pathway that regulates spindle orientation and asymmetric cell division [292–294]. But later, the 4th branch was found to be a part of the planar cell polarity pathway [295]. Ultimately, it is established that Wnt STP proceeds through three separate generalized pathways, namely, the canonical (β-catenin) and noncanonical (Ca^{2+} and planar cell polarity) pathways. In noncanonical signaling, Wnts bind to FZD receptors to activate DVL, but the downstream pathways activated by this binding do not involve GSK-3β or β-catenin [296].

Normally cytosolic β-catenin is degraded by a complex containing AXIN, Adenomatosis Polyposis Coli (APC) and Glycogen Synthase Kinase 3 (GSK3) in proteasomes. The canonical Wnt STP leads to stabilization of β-catenin by preventing its degradation [29]. As the amount of β-catenin rises, it accumulates in the nucleus, where it interacts with specific transcription factors, *i.e.*, LEF leading to regulation and expression of Wnt-target genes [295, 297–299].

Planar cell polarity pathway activates JNK, and directs asymmetric cy-

toskeletal organization and coordinated polarization of cell morphology within the plane of epithelial sheets [300]. The coordination of cellular polarization is an important feature of development and critical for organ function. Epithelial apical-basolateral polarity enables organs and tissues to perform vectorial functions, including transport of fluid or directed secretion of specialized components. In addition, most epithelial tissues require a second axis of polarity, commonly referred to as "planar cell polarity (PCP)", which is within the plane of an epithelium. This type of polarity is, however, not restricted to epithelial tissues, but is also found in mesenchymal cell types throughout animal development [301–303].

On the other hand, Wnt/Ca^{2+} pathway leads to release of intracellular calcium, possibly mediated by G proteins. It involves activation of Phospho-Lipase C (PLC), Protein Kinase C (PKC) and Calmodulin-dependent Kinase II (CamKII) [300].

2.4.2 Some biological functions of Wnt STPs

Wnt STPs are involved in regulation of cell fate determination, proliferation, differentiation, migration, apoptosis [304, 305] and regulation of bone mass [306] among others. It enables cells to influence behavior of their neighboring cells during development [30, 111, 307]. In matured organisms, Wnts are implicated in maintaining stem cell-like fates in the intestinal epithelium [308], skin [309] and hematopoietic cells [310]. Wnt STPs play a crucial role in the evolution of axial differentiation in all multicellular animals. However, its inhibition is important at later stages of body plan formation in vertebrates. It controls initial formation of the neural plate and many subsequent patterning decisions in the embryonic nervous system, including formation of the neural crest. It regulates the anatomy of the neuronal cytoskeleton and the differentiation of synapses in the cerebellum [311]. Wnts have a crucial role in synaptic physiology, as they modulate the synaptic vesicle cycle, the trafficking of neurotransmitter receptors and the interaction of these receptors with scaffold proteins in postsynaptic regions. In addition, Wnts participate in adult neurogenesis and protect excitatory synaptic ter-

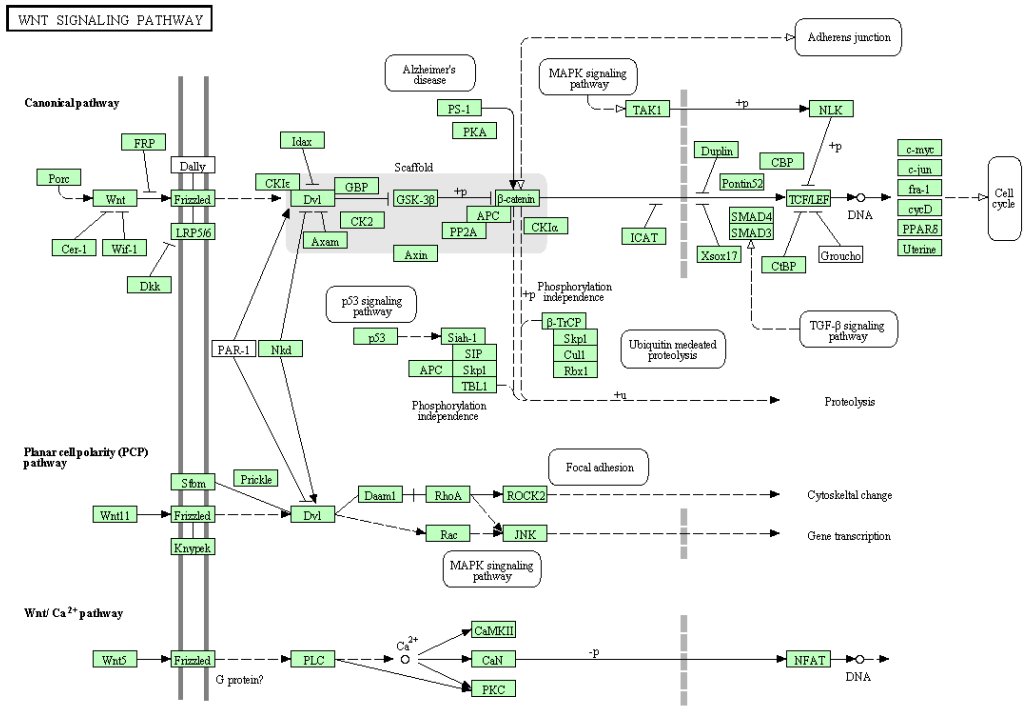


Figure 2.4: Human Wnt STP as given in KEGG

Wnt proteins are secreted morphogens that are required for basic developmental processes, such as cell-fate specification, progenitor-cell proliferation and the control of asymmetric cell division, in many different species and organs. The Wnt STP describes a network of proteins that are most well known for their role in embryogenesis and cancer. These proteins are also involved in normal physiological processes in adult animals.

minals from amyloid-β oligomer toxicity [312]. In the developing vertebrate limb, Wnt STP has been shown to have important functions during limb bud initiation, limb outgrowth, early limb patterning, and later limb morphogenesis events [313].

In absence of β-catenin, skin stem cells fail to initiate follicular morphogenesis. β-catenin is necessary for fate decisions of stem cells to form follicular keratinocytes; but it is not the single factor for inducing follicular differentiation of keratinocytes, as additional mesenchymal signals may be required [314, 315]. Several Wnt and TCF genes are coordinately required for hippocampal development [316]. Wnt10b gene maintains preadipocytes in undifferentiated state. Its disruption causes transdifferentiation of myoblasts into adipocytes. Thus the signaling mechanism is involved in both

adipocyte differentiation and mesodermal cell fate determination [317]. Wnt STPs also have essential role in limb initiation and induction of the apical ectodermal ridge in vertebrates [318]. Wnt proteins are involved in chondrocyte differentiation in early mesenchymal condensations, affecting skeletal maturation and bone growth [132]. They may also play major role in regulating hematopoietic cell fate. Inhibition or stimulation of Wnt STP is sufficient to shift natural differentiation program of bone marrow cells [319]. Recently, Wnt STPs have been found to take part in mammalian metabolism, *i.e.*, regulation of hepatic metabolism. This finding raises the possibility that Wnt STPs may play a similar role in the metabolic regulation of other tissues [320]. These observations have generated a strong urgency to understand the Wnt cascade whose multi-faceted functionality is still not properly known as stated by Cadigan and Peifer, 2009 [321].

2.4.3 Role of Wnt STPs in some diseases and disorders

Inappropriate activation and dysregulation of Wnt STPs contribute to numerous human tumors [91–93] and cancers (prostate, head and neck, breast, thyroid, colorectal, colon, stomach, lung, leukemia, ovary, kidney, liver, urinary bladder, prostate, neuroblastoma, mesothelioma, hepatoblastoma, and astrocytoma among others) [94, 97–99]. Reduced activation of Wnt STPs have been implicated in osteoporosis [126, 127]. Abnormal expression of Wnt family members and their regulatory molecules during embryogenesis can disturb Wnt-controlled balance that results in skeletal malformation [132]. Canonical Wnt STP requires the presence of LRP5/6 at the cell surface of the target cells. Gain or loss of functions in LRP5 has been linked to dramatic changes in bone mass, eliciting major interest in this receptor. Loss of LRP5 gene function results in OsteoPorosis-PseudoGlioma (OPPG) syndrome [103, 104]. High Bone Mass (HBM) trait is also linked with the same locus of gene (11q12-13) that controls OPPG syndrome [130]. Hence, the same gene was getting affected in opposite manners in both the above cases [322]. Abnormal β-catenin signaling may have role in causation of desmoid tumors. Post-translational modifications of β-catenin, which are

associated with its cytoplasmic accumulation, are frequently observed in fibroproliferative diseases with characteristics of dysregulated wound healing. These diseases include hypertrophic scar formation, aggressive fibromatoses, and Lederhose disease [120]. Wnts are also implicated in osteoarthritis [100]. Activated Wnt STP is involved in leukemia [117]. Dysregulated Wnt signaling is connected with the development of cystic kidney diseases [323, 324]. Changes in Wnt ligands and pathway components are associated with many kidney diseases, including acute renal failure, diabetic nephropathy and ischaemic injury [116]. Wnt STP is also implicated in the pathophysiology of rheumatoid arthritis [102].

Recent review on human diseases associated with mutations of the Wnt STPs components including [94] shows that WNT3 gene mutations are identified in tetra-amelia [101]. WNT10A mutation is detected in odonto-onchyodermal hypoplasia [119]. WNT10B mutations are detected in split-hand/foot malformation [118] and obesity [123]. A partial loss of WNT7A gene function causes Fuhrmann syndrome, whereas more severe limb truncation phenotypes are observed in Al-Awadi/Raas-Rothschild/Schinzel phocomelia syndrome result from null mutations [125]. WNT5A and other Wnt pathway genes are differentially expressed in psoriatic plaques [124]. WNT4 has role in the pathogenesis of renal fibrosis [325]. Loss of SFRP1 gene function is an probable early, aberrant molecular event in renal cell carcinogenesis [326]. FZD10 gene is found to be over-expressed in case of synovial sarcomas [129]. Missense mutations in LRP5 gene lead to several other diseases like Van Buchem disease, autosomal dominant sclerosteosis and Osteoporosis type I syndrome [131]. Mutations in either LRP5 or FZD4, two distinct genes within the EVR1 locus, can cause familial exudative vitreoretinopathy [105]. Mutations in LRP5 may lead to the development of diabetes and obesity [114]. A missense mutation in LRP6 is identified to be associated with coronary artery disease [121]. Altered GSK3-β activity and increased Wnt1 expression are associated with schizophrenia [133, 134].

Changes in AXIN1 and CTNNB1 genes have role in neuroepithelial brain tumors [112]. AXIN2 gene is involved in non-syndromic tooth agenesis [106, 107]. Aberrant Wnt signaling mediated by APC or β-catenin mutations

initiates the majority of human colorectal cancers and drives tumorigenesis through the activation of specific genes such as MYC [110]. About 90% of colorectal cancer cases originate from the constitutive activation of the canonical Wnt STP [109]. Mutations in the APC and β-catenin genes are implicated in sporadic hepatoblastomas and familial adenomatous polyposis [113]. β-catenin is also found to be an active player in hepatoblastomas, benign liver neoplasms, hepatocellular carcinoma and cholangiocarcinomas [122]. Later it was discovered that both the canonical and noncanonical Wnt STPs have complementary roles in hepatocellular carcinoma, where the canonical signaling contributes to tumor initiation, and noncanonical signaling to tumor progression [327]. Mutations in CREBBP and EP300 genes are implicated in RubinsteinTaybi syndrome. At present, more than 100 pathogenic mutations are known for the two genes together [108]. The Wnt pathway effector gene TCF7L2 is associated with type II diabetes [114, 115]. WNT5B gene is found to be overexpressed in leiomyoma cells [128].

2.5 Conclusive remarks

MAPK STPs are generalized processes taking part in most of the household events of cells. Calcium STP is important as gene expression events are effected by concentration as well as mode of entry of Ca^{2+} ions inside a cell. Wnt STPs are involved in multiple biological processes. Perturbations of these pathways are associated with various human pathologies including cancers. Thus these STPs, especially Wnt STP, form the pathway data for demonstrating effectiveness of the methodologies developed in the succeeding chapters. All the three pathways are considered in Chapter 3 for creation of modules. Human Wnt STP is used for performance comparison of some partitioning algorithms in Chapter 4, and creation of a Wnt diseasome in Chapter 6. Species-specific Wnt STPs are used for deriving a phylogenetic tree from modules in Chapter 5. Wnt STP genes and their associated diseases are manually curated for construction of a human Wnt diseasome in Chapter 6.

Chapter 3

Modularization Algorithm

3.1 Introduction

Modularization is a process which divides a network into smaller units for better understanding and analysis of the original network. There is no single definition available for a module. We define a module as a subset of the original biochemical network, which tends to be self-sufficient in terms of biological function and has minimal dependency on the rest part of the network. Unlike studying a signaling pathway as a whole, this enables one to study the individual modules (less complex smaller units) easily and enables a better study of the entire pathway. Thus analyzing all the modules generated from a pathway separately, one can have a better operational view of the whole network [1, 2]. We have already described some methods for partitioning a biochemical pathway in Chapter 1.

In this Chapter, we will develop a modularization algorithm [1]. For dividing an STP, creation of modules starts with a member having maximum number of relations in a given network. The module grows in size by including neighbors of the starting member in successive steps. The neighbors are either included into or excluded from the module depending on the number of their relations being present its inside or outside. That is, if a member has less than or equal to a certain number of relations (known as complexity level c) outside the module, it gets included in the module. The term, "c", needs to be specified and can be varied by the user. We have to select an appropriate c-value for each pathway. The effectiveness of the algorithm is demonstrated on human MAPK, Ca^{2+} and Wnt STPs.

MAPK pathway is one of the most ubiquitous signal transduction systems [71]. It is characterized by the general path, "$Stimulus > MAPKKK > MAPKK > MAPK > Response$", where MAPKK is the kinase of MAPK and MAPKKK is the kinase of MAPKK. The symbol "$A > B$" stands for "A stimulating B" [72]. In most of the cases, MAPKKK is activated by small G proteins such as Ras and Rap1 [26, 45]. The sequential activation of the MAPK cascade eventually results in the activation of transcription factors, phospholipases or cytoskeletal proteins, microtubule-associated proteins and the expression of specific sets of genes in response to environmental stimuli.

MAPK pathway is conserved in all the eukaryotes, and plays a key role in regulation of gene expression as well as cytoplasmic activities. It is even active in plants, *e.g.*, the Arabidopsis genome consists of 23 MAPKs, twelve of which are ERK type and the others are plant specific [258]. Detailed description of the MAPK STP is furnished in Chapter 2.

Ca^{2+} STP are very peculiar in nature. Normal intracellular Ca^{2+} level ($10^{-7}M$) is much lower from the extracellular concentration of $10^{-3}M$. Ca^{2+} ions precipitate phosphate of the established energy currency of cells. Moreover, high concentration of intracellular Ca^{2+} ions lead to cell death. This is the reason why Ca^{2+} ion concentration must be maintained at low level in cytoplasm. Hence cells have evolved techniques for free Ca^{2+} ion binding to reduce its effect towards cytosol, which later is used as well for signal transduction across and inside the cell [28]. Ca^{2+} gradients within cells have been observed, which is used to initiate cell migration, exocytosis, lymphocyte killer cell activity, acid secretion, transcellular ion transport, neurotransmitter release, gap junction regulation and numerous other functions [275]. Details of this pathway are described in Chapter 2.

Wnt molecules are secreted cysteine-rich, lipid-modified glycoproteins. They bind to FZD receptors along with co-receptor LRPs and initiate the downstream steps, those altogether are known as Wnt STP [29]. Wnt STP proceeds through three separate generalized pathways, namely the canonical (β-catenin) and noncanonical (Ca^{2+} and planar cell polarity) pathways. Wnt STPs are involved in regulation of cell fate determination, proliferation, differentiation, migration, apoptosis [304,305] and regulation of bone mass [306] among others. It enables cells to influence behavior of their neighboring cells during development [30, 111, 307]. In matured organisms, Wnts are implicated in maintaining stem cell-like fates in the intestinal epithelium [308], skin [309] and hematopoietic cells [310]. In Chapter 2, the Wnt STP is described in detail.

3.2 The proposed modularization algorithm

In order to decompose a network into several modules, we have proposed an algorithm which is described in this section. The algorithm views an entire biochemical pathway as a graph having gene products and chemical compounds as vertices, and edges being different kinds of interactions among them. An edge can be a gene-protein interaction, protein-protein interaction or protein-compound interaction. First, we define some useful terms for describing the algorithm.

3.2.1 Some Useful terms

- E: Set of all nodes (representing gene products and chemical compounds) present in a network (pathway), excluding isolated nodes
- M: Set of nodes present in a module (a part of network)
- c: A user defined parameter used for generating modules
- k: *Extension* index (stage of inclusion of immediate neighbors of nodes in a module)
- M^k: Set of nodes present in a module after k^{th} *extension*
- N_S: Set of immediate succeeding nodes of a given node
- n_s: An individual member of N_S
- N_P: Set of immediate preceding nodes of a given node
- n_p: An individual member of N_P
- r: type of interaction that exists between n_p and n_s; $r = a$ depicts a relation of activation, $r = b$ depicts a relation of binding or association, $r = i$ depicts a relation of inhibition, $r = d$ depicts a relation of indirect effect between nodes n_s and n_p
- N_R: Set of relations

- $n_r = (n_p, n_s, r)$: An individual member of N_R
- R_{np}: Total number of relations that exist with n as the preceding node
- R_{ns}: Total number of relations that exist with n as the succeeding node
- M_P: Set of permanent nodes (nodes having all their relations inside a module)

The total number of relations with n as either a preceding or succeeding node is given by

$$T_n = R_{np} + R_{ns} \tag{3.1}$$

Since R_{np} and R_{ns} are outdegree and indegree, respectively of a node n, T_n is the total degree of the node n. T_n represents the total number of relations associated with node n. T_{n^k} stands for the total number of relations of node n that gets included in a module during k^{th} *extension*. Likewise, T_M represents a set, comprising T_n-values, where $n \in M$.

3.2.2 Description of the algorithm

The algorithm starts with detection of a node n having maximum number of relations in the node pool E for a given network. By default, we named modules by their starting node's name. Considering the detected node as the starting point (the starting member is always a permanent member), an initial module is created for relations r, where n is either a *predecessor* or a *successor*. Thus the module is *extended* by including these nodes. Here an eventuality can arise where more than one node may have maximum number of relations. Then any one of the nodes (having maximum number of relations) that is encountered first by the algorithm is taken as the start point by default (followed by the others).

Once a module is initialized, the total number of relations (T_n) of every individual node is considered. For a node in a module, if the number of relations lying inside the module is equal to the total number of relations

associated with it, the node is considered to be *permanent*. If a node in a module has more than c relations that lie outside the module, it gets *excluded* from the module along with decreasing the previous *non-permanent* nodes' total relations T_n by one. The *extension* and *exclusion* processes continue till there is no new node to be considered for its membership in any module or a node is present under consideration. That is, all the nodes of a module are declared as *permanent*. It is to be mentioned here that once a node is declared *permanent*, it is deleted from the node pool E. Hence a single node can not be included in more than one module. Also, if a node appears more than once in a network, its positional significance is taken into account. That is, if a node X is present four times in a network, it will be considered four times as X_1, X_2, X_3 and X_4. After successful creation of a module, the algorithm searches for another starting point and repeats the above processes to create another module. The algorithm continues to run till exhaustion of all the nodes present in the node pool E. Pseudocode of the algorithm is furnished here as Algorithm 1. Now we describe the function of the algorithm with an example network.

3.2.3 Modularization of an example network

The example network contains 26 nodes and 26 relations (Figure 3.1). The relations are of different kinds, *i.e.*, activation (a), inhibition (i), binding (b), and indirect effect (d). Chemical compounds are denoted with a 'circle', activation with '$\rightarrow$', binding with '$-$', and inhibition with '$-|$'. Each member of the set given below shows the relation between a node and its following node. The term (A,F,a) denotes existence of activation relation from node A to node F. Thus we have a set N_R of members

$N_R = \{$(A,F,a), (B,G,a), (C,H,a), (D,I,a), (F,K,b), (G,K,b), (H,K,b), (I,K,b), (K,L,b), (L,P,a), (Q,P,i), (T,P,i), (P,Z,a), (E,J,a), (J,N,a), (M,J,a), (O,P,a), (R,S,a), (S,P,a), (R,W,a), (W,X,a), (X,Y,a), (X,Y,i), (U,V,a), (V,P,a), (V,Z,a)$\}$

We have to calculate total number of relations for all the nodes present in the

1 Pseudo code for Modularization Algorithm

Ensure: $E \neq \phi$

1: Find start/central node

if $(T_n \leftarrow Max\{T_M\})$ **then**

$n \leftarrow$ start point/central node

$M_P \leftarrow M_P \cup \{n\}, E \leftarrow E - \{n\}$

$k \leftarrow 0$

end if

2: Extend module

for $(k \leftarrow k + 1)$ **do**

select nodes from N_S and N_P of n and put in M^k

end for

3: Check permanency of nodes

if $N_S \cup N_P \subset M^k$ for a node n^k **then**

$E \leftarrow E - \{n^k\}, M_P \leftarrow M_P \cup \{n^k\}$

end if

4: Exclude node

if $[T_n^k -$ number of nodes in M^k related to $n^k] > c$ **then**

$M^k \leftarrow M^k - \{n^k\}$

for $(n^{(k-1)} \notin M_P)$ **do**

$T_{n^{(k-1)}} \leftarrow [(T_{n^{(k-1)}}) - 1]$

end for

end if

5: Build a complete module

repeat

Step 2-4

until $M^k \subset M_P$

6: Create next module

repeat

Step 1-5

until $E = \phi$

network in order to choose the node with maximum number of relations as the starting point of an originating module. Total number of relations (T) of a node can be calculated by summing relations of a node first as predecessor and then successor.

$T_A = 1+0 = 1$, $T_B = 1+0 = 1$, $T_C = 1+0 = 1$, $T_D = 1+0 = 1$, $T_E = 1+0 = 1$, $T_F = 1+1 = 2$, $T_G = 1+1 = 2$, $T_H = 1+1 = 2$, $T_I = 1+1 = 2$, T_J

= 1+2 = 3, T_K = 1+4 = 5, T_L = 1+1 = 2, T_M = 1+0 = 1, T_N = 1+1 = 2, T_O = 1+0 = 1, T_P = 1+6 = 7, T_Q = 1+0 = 1, T_R = 2+0 = 2, T_S = 1+1 = 2, T_T = 1+0 = 1, T_U = 1+0 = 1, T_V = 2+1 = 3, T_W = 1+1 = 2, T_X = 2+1 = 3, T_Y = 2+0 = 2, T_Z = 0+2 = 2.

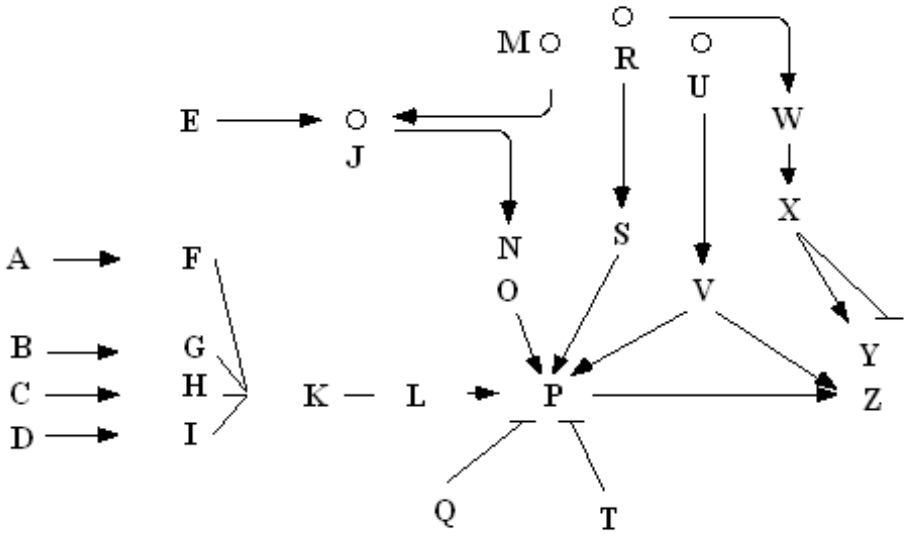


Figure 3.1: The example Network
The relations are of different kinds, *i.e.*, activation (a), inhibition (i), binding (b), and indirect effect (d). Chemical compounds are denoted with a 'circle', activation with '→', binding with '−', and inhibition with '−|'.

Here node P is having the highest number of relations. So P becomes the starting point of the module. After the first extension, with immediate neighbors, the module resembles Figure 3.2(a). Now we describe below the steps for determining the modules of the network. Here c-value is 2.

1. All relations of Q, T and O lie in the created module. So they became permanent members as shown in Figure 3.2(b).

2. After second extension (Figure 3.2(c)), respective relations of L, S and Z lie in the module. Hence they were also considered as permanent members

of the module. Node K has more than 2 relations that lie outside the present module. So K cannot be a member of module P (Figure 3.2(d)).

3. After third extension as shown in Figure 3.2(e), for V, R and U, their corresponding relations are inside the module. So except W and X, every member present in the module became permanent, and they were excluded from the node pool of the network.

4. Fourth extension made W permanent. Like wise after fifth and sixth extensions, all the members are permanent. Hence creation of the first module was complete. Module P (Figure 3.2(f)) contains 13 permanent members.

5. The whole process is again repeated taking K, *i.e.*, the node with maximum relations from the rest nodes present in the node pool.

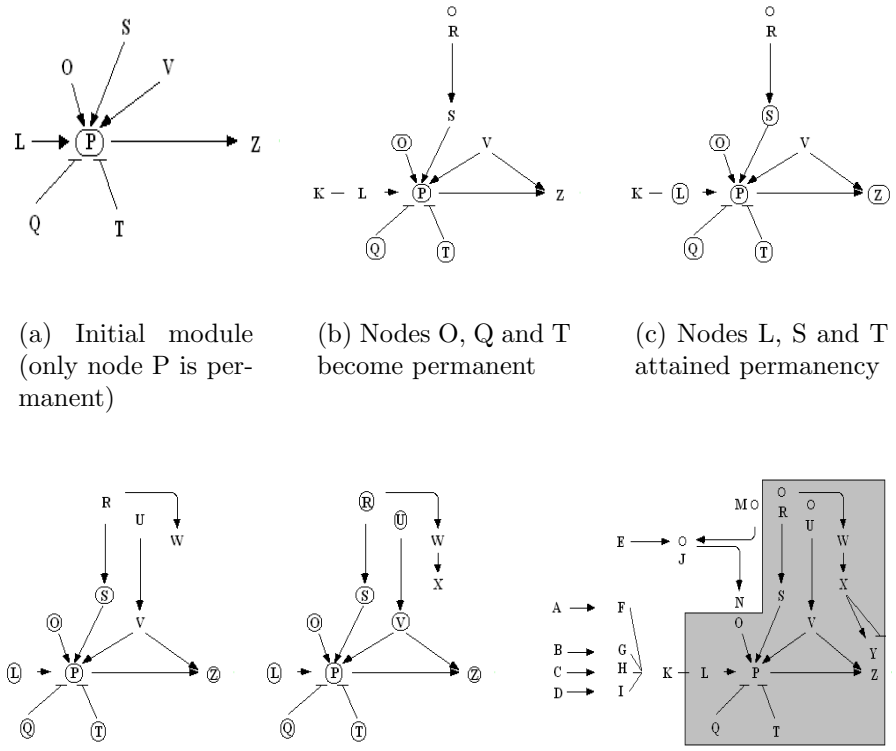
6. The node pool became a null set, after creation of three modules namely P, K and J. The modularized entire network is given in Figure 3.3.

3.3 Results

Now we describe modules obtained from MAPK, Ca^{2+} and Wnt STPs. All these pathways have been obtained from KEGG/Pathway database.

3.3.1 Modularization of MAPK STP

Human MAPK STP is a complex network of 135 nodes and 182 relations ($|N_R| = 182$). TAO1/2 is the only isolated node present in this network, *i.e.*, $|E| = 135 - 1 = 134$. Modules are created from the same pathway at complexity level of 1, 2, 3, 4 and 5. A list of modules created for each c value is given in Table 3.1.



(a) Initial module (only node P is permanent)

(b) Nodes O, Q and T become permanent

(c) Nodes L, S and T attained permanency

(d) Node K is excluded from the module along with addition of two nodes

(e) Nodes R, U and V are declared permanent

(f) Module P (marked with gray color) in the whole network

Figure 3.2: Various stages in construction of module P and the modularized example network

Modularization for $c = 1$

For $c = 1$, human MAPK STP gets divided into 38 modules (Figure 3.4). Twelve of them are isolated modules. Among the rest modules only 6 seem to give any biological significant, *i.e.*, modules *JNK*, *p38*, *ERK*, *Ras*, *MEKK1*, and *GRB2*. Module *ERK* along with modules *TrkA/B*, *GRB2* and *Ras* roughly represented the conventional MAPK pathway. Modules *JNK* and *MEKK1* are part of the JNK pathway. p38 pathway is represented by a

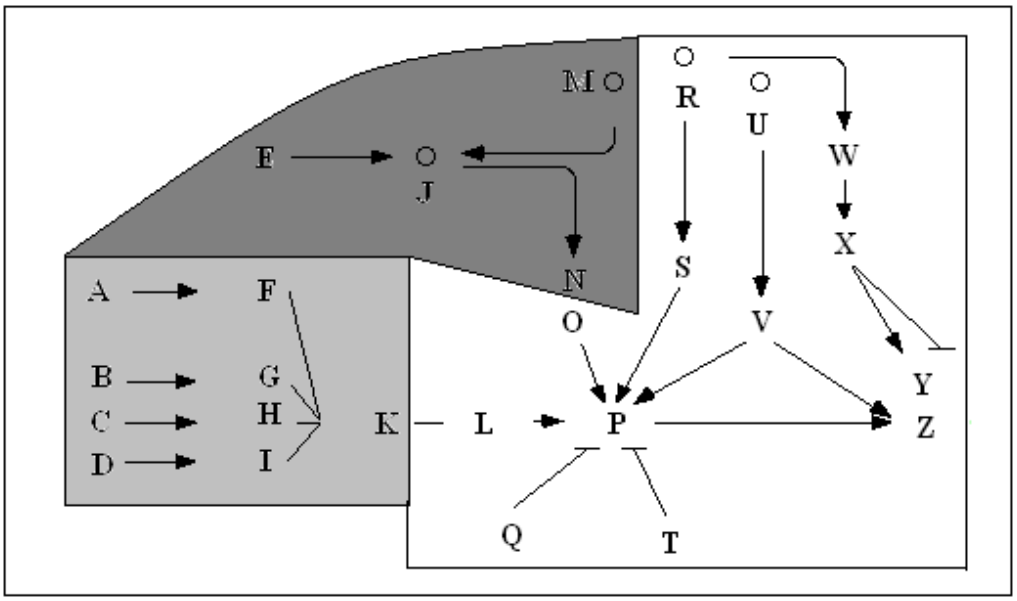


Figure 3.3: Modularized example network for $c = 2$
Dark gray part - module J, white part - module P and light gray part - module K.

module named *p38*. The partly known ERK pathway is represented by module *ERK5*. Here c value is very low. So the network is facing over-splitting problem. May be with an increase in c value more significant modules will be found.

Modularization for $c = 2$

For $c = 2$, 32 modules are obtained from human MAPK STP (Figure 3.5). Modules *JNK*, *p38*, *ERK*, *MEKK1*, *GRB2*, and *ERK5* remain unchanged except increase by a single node or relation. Module *(PKC*, *PKA* and *Mos* merged with module *Ras* making the combined module more meaningful. In addition *ASK1* emerges as a major module here. Still 9 isolated modules and 14 small to medium modules are left without any satisfactory explanation. So we switch to modularization for $c = 3$.

Table 3.1: List of modules of human MAPK STP for different c value

		$c = 1$		$c = 2$		$c = 3$		$c = 4$		$c = 5$	
Sl. no.	module name	node	rel	node	rel	node	rel	node	rel	node	rel
01	JNK	14	15	14	15	21	23	22	25	22	25
02	p38	13	14	13	15	13	15	40	47	40	47
03	ERK	16	18	17	19	17	19	17	19	17	19
04	Ras	10	09	15	15	16	18	27	29	33	42
05	MEKK1	07	06	07	06	13	14	11	12	07	06
06	TAK1	04	03	04	03	14	16	03	02	03	02
07	MKK4	03	02	03	02	04	03	02	01	02	01
08	MKK7	04	03	04	03	04	03	04	03	04	03
09	MEK1	01	Nil	01	Nil	01	Nil	01	Nil		
10	MEK2	01	Nil	01	Nil	01	Nil	01	Nil		
11	ASK1	02	01	07	06	07	06				
12	TNFR	02	01	02	01	02	01				
13	GRB2	07	06	07	06	11	10				
14	JIP3	01	Nil	01	Nil	01	Nil				
15	MKK3	02	01	02	01	02	01				
16	MKK6	01	Nil	01	Nil	01	Nil				
17	TrkA/B	04	03	04	03						
18	IL1R	02	01	02	01						
19	Ca^{2+}	03	02	03	02						
20	CASP	03	02	03	02						
21	TRAF2	02	01	02	01						
22	TRAF6	02	01	02	01						
23	TAB1	03	02	03	02						
24	RafB	02	01	01	Nil						
25	Raf1	01	Nil	01	Nil						
26	Tp12/cot	01	Nil	01	Nil						
27	MLK3	02	01	03	02						
28	NIK	02	01	02	01						
29	IKK	01	Nil	01	Nil						
30	JIP1/2	01	Nil								
31	PP2CA	01	Nil	01	Nil						
32	DAXX	05	04								
33	PKC	02	01								
34	PKA	01	Nil								
35	Mos	01	Nil								
36	MP1	01	Nil								
37	ERK5	04	03	04	03	04	03	04	03	04	03
38	GADD45	02	01	02	01	02	01	02	01	02	01

The details of modules obtained from human MAPK STP for c value of 1, 2, 3, 4 and 5 are given in this table. The column *node* indicates number of nodes and column *rel* gives number of relations present in a module.

Modularization for $c = 3$

We are getting 18 modules for $c = 3$ as seen in Figure 3.6. Modules *GRB2*, *Ras* and *ERK* divide effectively the classic MAPK STP into 3 parts. JNK pathway is divided into 4 parts namely modules *MEKK1*, *MKK4*, *MKK7*, and module *JNK*. Modules *p38*, *ASK1* and *TAK1* counter for p38 pathway except a few small modules. Here the problem of over-splitting is minimized with only 4 isolated modules and 2 small modules.

Modularization for $c = 4, 5$

For $c = 4$, the network is separated into 12 modules. Two singleton modules are present. But, some modules like *p53* are getting much larger and complex

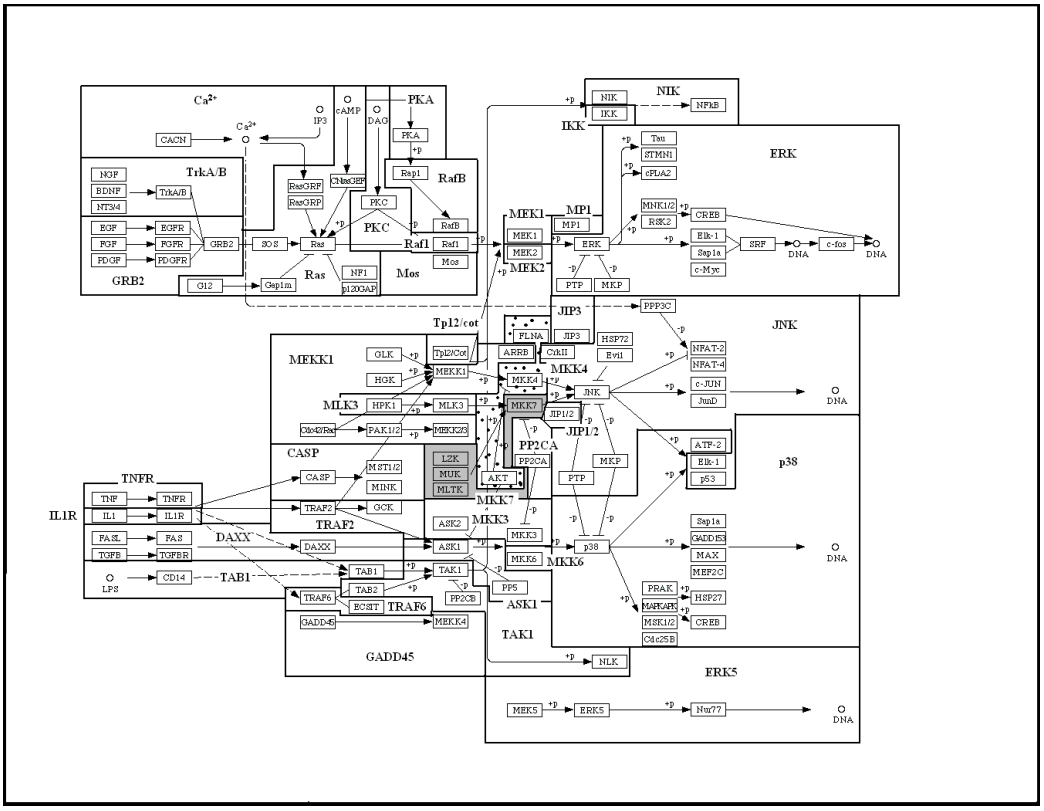


Figure 3.4: Modules of human MAPK STP for $c = 1$

Here the network is divided into 38 modules. Twelve of them are isolated and very few were eligible for consideration of their biological significance.

in size. The scenario become difficult for c value of 5 as modules become more large in size. The modularized networks of human MAPK STP for $c = 4$ and $c = 5$ are given in Figure 3.7 and Figure 3.8 respectively. For higher values of c, the modules become even more complex.

3.3.2 The best set of modules of MAPK STP

The best set of 18 modules is obtained for $c = 3$. Module *GRB2*, *Ras* and *ERK* divide effectively the classic MAPK pathway into 3 parts. JNK pathway is divided into 4 parts namely modules *MEKK1*, *MKK4*, *MKK7* and *JNK*. Modules *p38*, *ASK1* and *TAK1* are parts of p38 pathway. Here the problem of over splitting is a lot minimized with only 4 isolated modules and

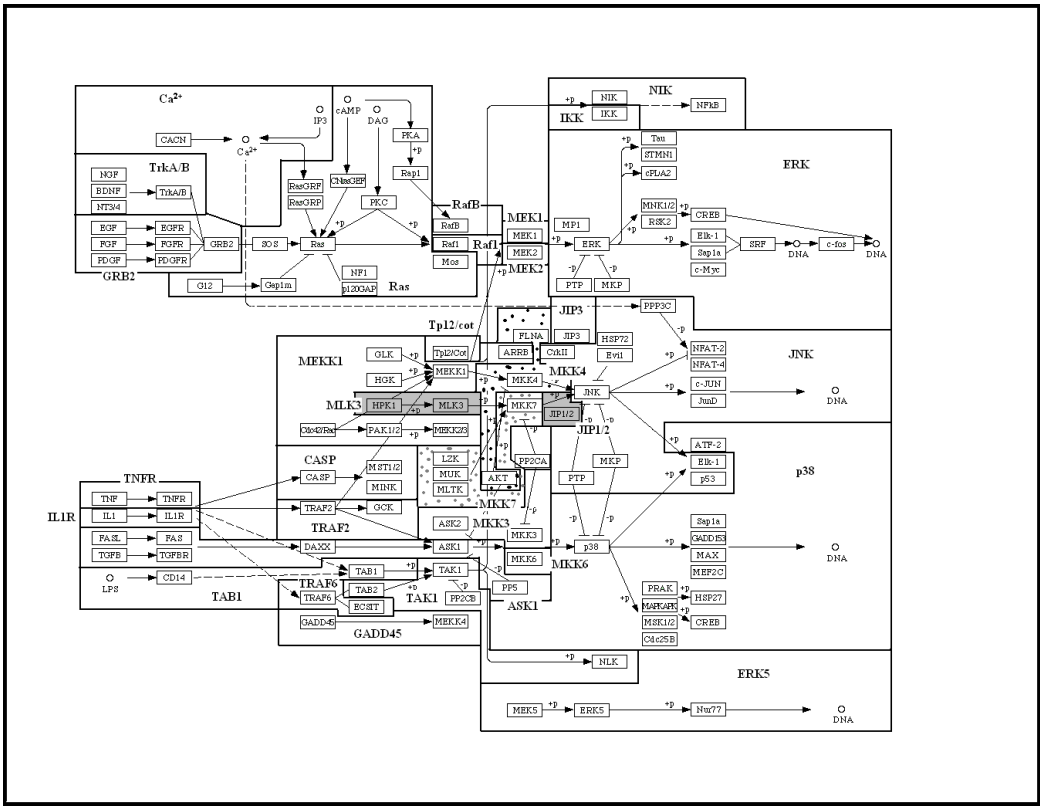


Figure 3.5: Modules of human MAPK STP for $c = 2$
For c-value of 2, the network splits into 32 modules.

2 small modules. The details are shown in Figure 3.6 [1].

3.3.3 Comparative study on modules of MAPK STP of 9 species

We have done a comparison among MAPK STP modules of the taken set of 9 species for $c = 3$, keeping in mind complexity of a module and over splitting of a network. MAPK STPs of *D. melanogaster* and *S. cerevisiae* are very simple and different in layout from MAPK STPs of the rest species. So we have modularized them for $c < 3$. They have been divide into three independent modules each. The remaining 7 species have been modularized for $c = 3$. MAPK STPs of *H. sapiens* and *M. musculus* are almost identical. So their modules have maximum resemblance with each other followed by

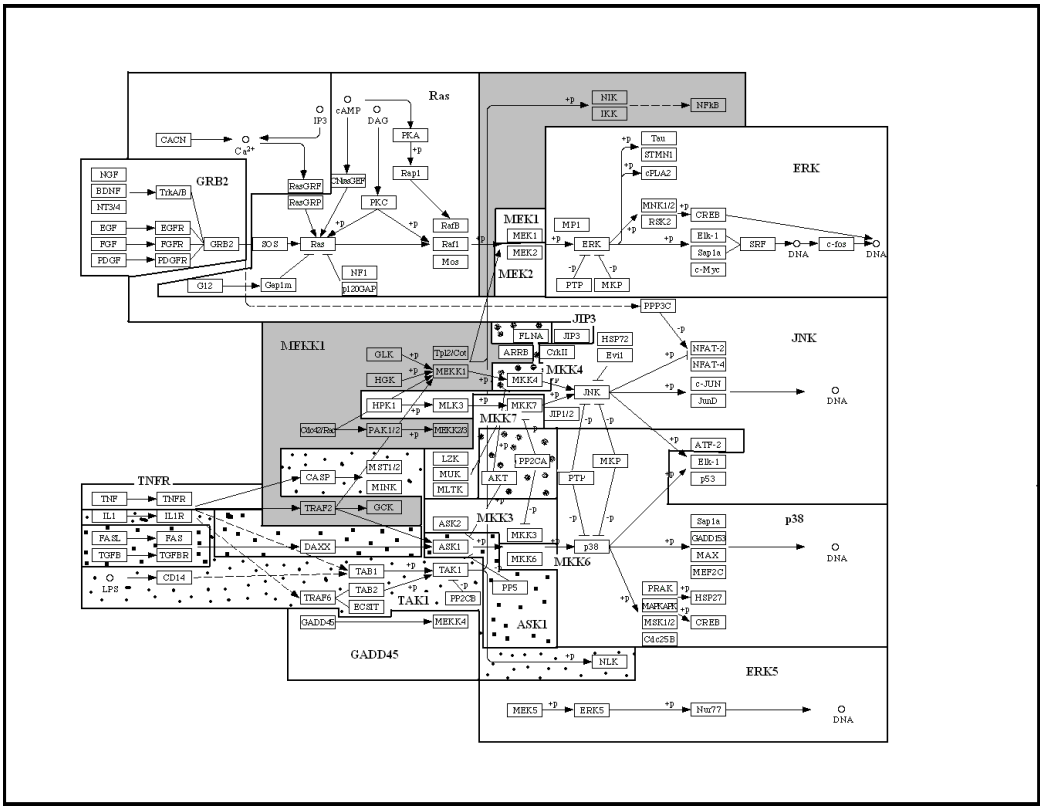


Figure 3.6: Modules of human MAPK STP for $c = 3$

Human MAPK STP got divided into 18 modules for $c = 3$. Here the problem of over splitting of the network is minimized.

the modules of *R. norvegicus*. In *B. taurus*, maximum modules found in *H. sapiens* are in elementary stage and exist in separate modules. MAPK STP of *C. familiaris*, *S. scrofa* and *P. troglodytes* are least developed. Hence, we have found a gradual development of the pathway from *C. familiaris* to *H. sapiens* [1]. Details of the species-specific modules are given in Table 3.2.

3.3.4 Modularization of Ca^{2+} STP

Human Ca^{2+} STP contains 55 nodes. One node *(C13050)* is isolated. So $|E|$ = 55 - 1 = 54. These 54 nodes are having 59 relations among them. Modules are created from the same pathway at complexity level of 1, 2, 3, 4, 5, 6, and 7. For $c = 7$, the whole network emerges into a single module, so there is no

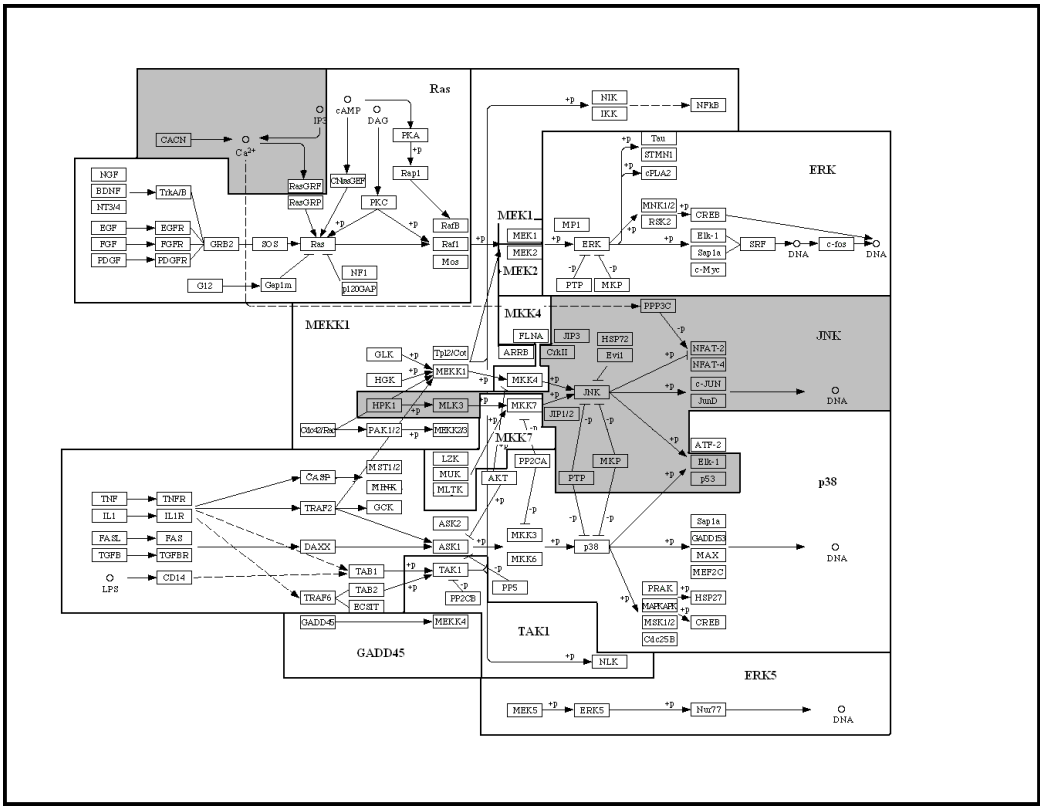


Figure 3.7: Modules of human MAPK STP for $c = 4$

Modularization for $c = 4$ separates human MAPK pathway into 12 modules. Most of them are eligible for consideration of their biological significance.

need to consider the value of c beyond 7. The results are given in Table 3.3.

Modularization for $c = 1$

Eleven modules are obtained for $c = 1$. Module *(C00076)2* emphasizes role of plasma membrane, endoplasmic reticulum and mitochondria in Ca^{2+} ion balance of cells. But Ryanodine receptors present in endoplasmic reticular membrane are not included in this module. Module *CALML6* represents role of calmodulin like proteins (calcium binding proteins) that upon binding with free Ca^{2+} ions change confirmation and trigger other enzymes and ion channels. *(C00076)1* module contains Ca^{2+} channels present in plasma membrane for import purpose. Module *BST1* deals with Ca^{2+} ion flow from outside to inside of bone marrow cells but how its intracellular balance is

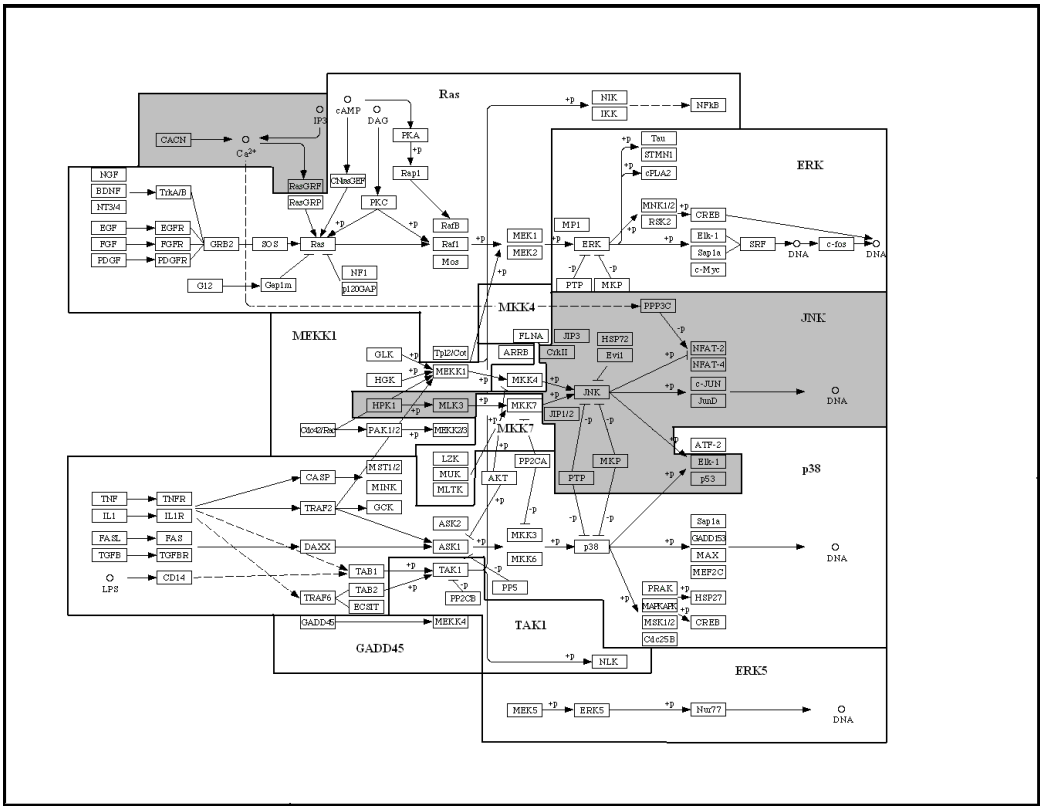


Figure 3.8: Modules of human MAPK STP for $c = 5$
For $c = 5$, we get only 10 large modules that defy our basic purpose of simplifying a network.

maintained is not clear in the module. For $c = 1$, the network is splitting profusely. Excessive splitting is giving rise to a lot of small modules as shown in Figure 3.9. This is the reason why we are unable to assign any biological significance to the rest 7 modules. So modularization of the same network is done for $c = 2$.

Modularization for $c = 2$

Six modules are obtained for $c = 2$ as shown in Figure 3.10. Modules *CALML6* and *BST1* remain unchanged. In module *(C00076)2*, ryanodine receptors are included, covering the complete process of Ca^{2+} ion flow and balance in a cell. *(C00076)1* is increased by one node PLCD3. The changed module shows plasma membrane based Ca^{2+} import channels and interac-

Table 3.2: Modules obtained from MAPK STPs of 7 different species for c = 3

Human and Mouse		Cow		Rat		Pig		Chimpanzee		Dog	
name	N	name	N	name	N	name	N	name	N	name	N
JNK	21	Ca^2+	3	JNK	5	Ca^2+	4	Ca^2+	14		
p38	13	ERK	13	p38	17	C-jun	2				
ERK	17	c-fos	2	ERK	14						
Ras	16	G12	2	Ras	25			Ras		Ras	7
MEKK1	13	IKK	2	MEKK1	14			MEKK1	12		
TAK1	14	CD14	2	LPS	2			CASP	2		
MKK4	4	MKK4	5							MKK4	2
MKK7	4	CDC42/Rac	2	MKK7	5					CDC42/Rac	3
MEK1	1	FASL	2	MEK1	1						
MEK2	1					FASL	3				
ASK1	7	TGFB	2	TGFBR	5	TGFB	2	TGFB	2		
TNFR	2	TNFR	3			TNFR	3				
GRB2	11	EGF	2	GRB2	11	EGF	2	EGF	2		
JIP3	1	FGF	2								
MKK3	2	TrkA/B	2								
MKK6	1										
ERK5	4	Nur77	2	ERK5	4					Nur77	2
GADD45	2										

This table contains information about modules obtained from MAPK STPs of *H. sapiens* (human), *R. norvegicus* (rat), *M. musculus* (mouse), *B. taurus* (cow), *S. scrofa* (pig), *C. familiaris* (dog)and *P. troglodytes* (chimpanzee). Column *N* is giving number of nodes present in the modules.

Table 3.3: List of modules of Ca^{2+} STP for different c value

		c = 1		c = 2		c = 3,4		c = 5		c = 6		c = 7	
Sl. no.	module name	node	rel	node	rel	node	rel	node	rel	node	rel	node	rel
01	(C00076)2	24	23	25	24	29	28	40	43	46	51	54	59
02	CALML6	08	07	08	07	08	07	08	07	08	07		
03	(C00076)1	06	05	07	06	07	06	06	05				
04	C01245	02	01	09	08	10	12						
05	C00165	01	Nil	01	Nil								
06	BST1	04	03	04	03								
07	PLCE1	02	01										
08	PLCG1	02	01										
09	PLCB1	03	02										
10	PLCD3	01	Nil										
11	RYR1	01	Nil										

The details of modules obtained from Human Ca^{2+} STP for c value of 1, 2, 3, 4, 5, 6, and 7 are given in this table. The column *node* indicates number of nodes and column *rel* gives number of relations present in a module.

tion of the imported Ca^{2+} ions with one of the PLC group. *C01245* module includes proteins belonging to PLC family and their relation with C01245. From prior knowledge we know that PLC group members break into C01245 as a result of activation. C01245 molecule is a ligand for ITPR1 (Inositol 1,4,5-TriPhosphate Receptor, type 1) present in endoplasmic reticular membrane. This module is resulted due to convergence of modules *C01245*, *PLCB1*, *PLCG1* and *PLCE1* found for $c = 1$. So the problem of over splitting noticed for $c = 1$ is reduced here. Still we are left with the problem

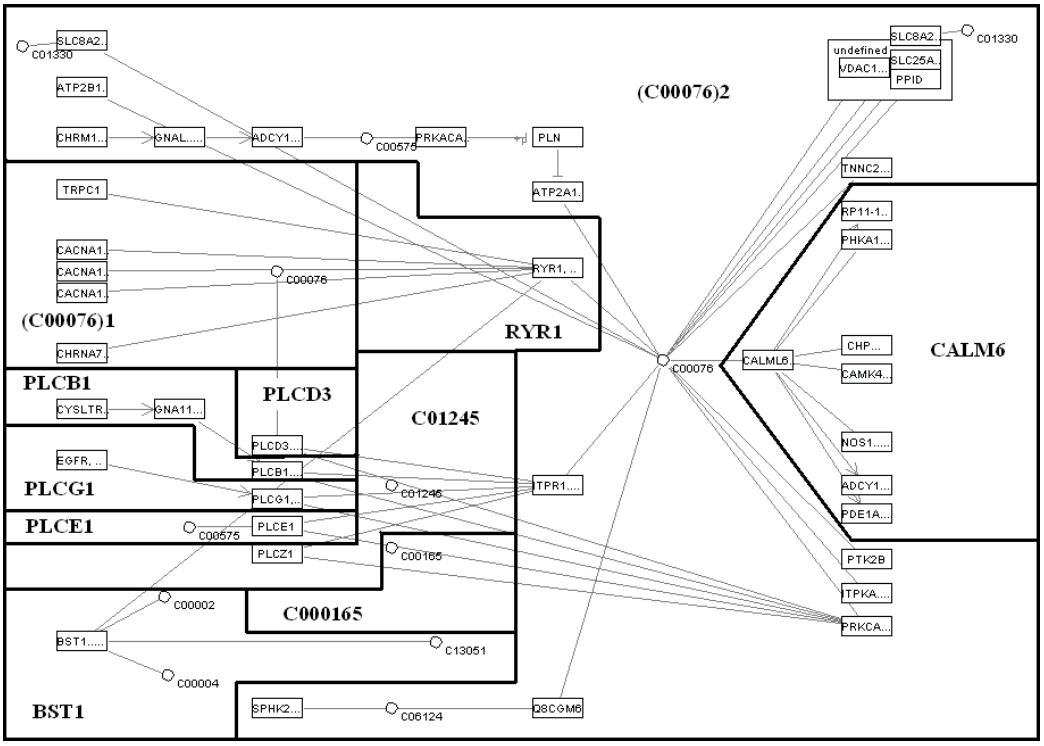


Figure 3.9: Modules of human Ca^{2+} STP for $c = 1$

For $c = 1$, human Ca^{2+} STP is divided into 11 modules. Here black lines separate the modules from each other. The figure shows many small modules resulted due to low complexity level that led to over splitting of the network. Each module is named after its starting node and marked with bold faced letters.

of prior-described module *BST1* and a small module *C00165* comprising of a single node to describe. Its difficult to analyze such small modules. This lead to modularization of the network for $c = 3$.

Modularization for $c = 3$

Four modules are created for $c = 3$ as shown in Figure 3.11. Module *BST1* for $c = 2$ is merged with module *(C00076)2* that gives a complete explanation of Ca^{2+} ion balance in bone marrow cells through ryanodine receptors present in endoplasmic reticular membrane. Module *C00165* for $c = 2$ is merged with module *C01245*. C00165 is an byproduct when PLC group members break to C01245. Like C01245, it is not a ligand for ITPR1. It binds with PKC that takes part in controlling plasma membrane based Ca^{2+} ion channels.

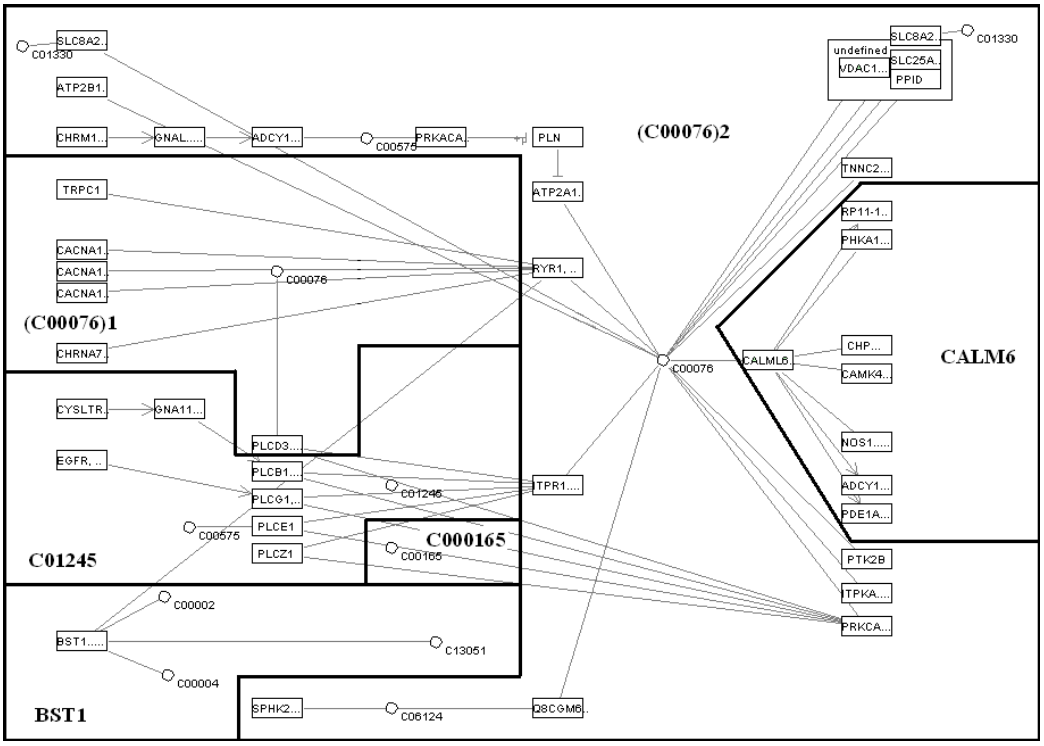


Figure 3.10: Modules of human Ca^{2+} STP for $c = 2$

We obtained 6 modules from human Ca^{2+} STP for $c = 2$. These modules are separated from each other by black lines in the figure. Here we are getting less number of modules than Figure 3.9.

But we were able to decipher it's role clearly only after its emergence with module *C01245*. For $c = 2$, where C00165 is included in another module, it is confusing to decipher and understand this information. We are getting exactly similar modules for $c = 4$.

Fixing the c value

Now question arises that once biologically significant modules are found for some value of c, whether modularization should continue for higher values of c to get more meaningful modules or stop the process. To get a logical answer to this question, we have obtained modules for $c = 5$ and 6. Three modules are found for $c = 5$. Module *(C00076)2* is increased by several nodes and relations, that make it large and complex, hence our primary objective of dividing a complex network to simpler units is failed here. *(C00076)1*

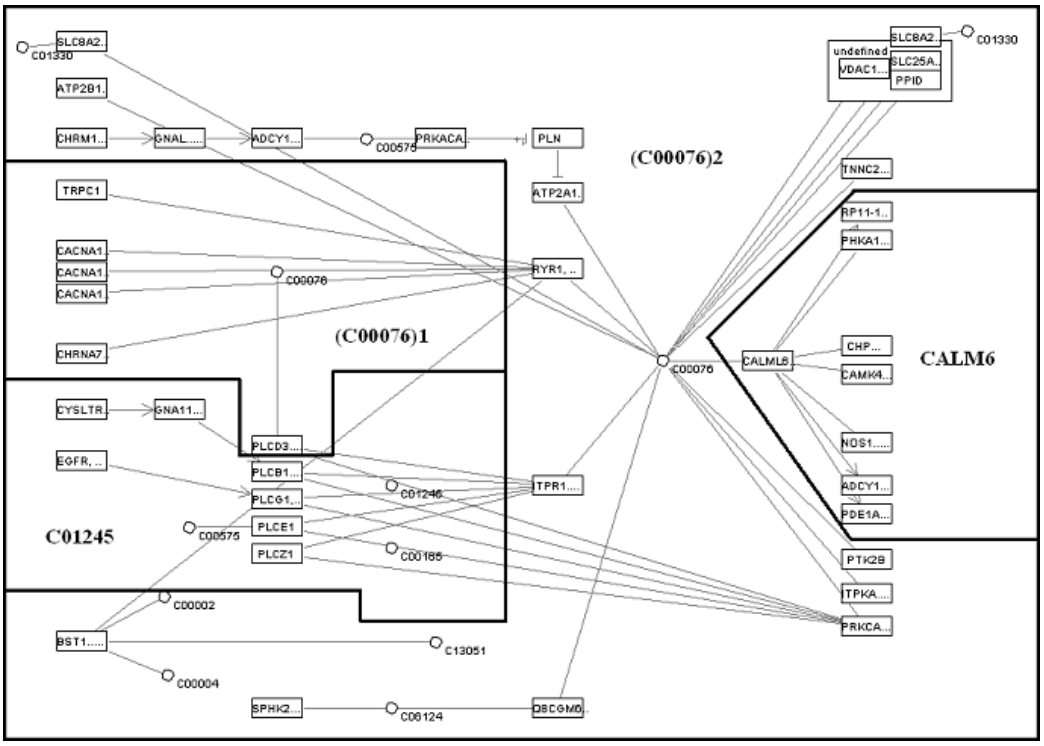


Figure 3.11: Modules of human Ca^{2+} STP for $c = 3$ and 4

Here the network is divided into 4 parts due to modularization. For $c = 3$ and 4 we are getting identical modules.

module is decreased by one node and one relation that again gives rise to the already discussed problem of Ca^{2+} ion balance inside the module. Module *CALML6* is identical to that of obtained for lower values of c. For $c = 6$, we have obtained 2 large modules. The whole network is divided into 2 parts, *i.e.*, the unchanged *CALML6* module and *(C00076)2* module comprising the rest part of the network. For $c = 7$, the whole network rounds up to a single module.

Hence it is found that after a certain level, modularization with increasing c-value will yield similar results with that of the previous complexity level or the modules will be enough large making their study and analysis difficult. As our objective is simplified study of a network and we are getting biological significant modules for $c = 3$, we fixed $c = 3$ for modularization of Ca^{2+} STP. This value of c is used later for analyzing Ca^{2+} STPs belonging to different

species.

3.3.5 The best set of modules of Ca^{2+} STP

The best set of modules was obtained for $c = 3$ (Figure 3.11). Here, module *BST1* for $c = 2$ is merged with the module *(C00076)2* that provides a complete explanation of Ca^{2+} ion balance in bone marrow cells through ryanodine receptors present in endoplasmic reticular membrane. Module *C00165* for $c = 2$ is merged with module *C01245*. C00165 is a byproduct of PLC group members when they break into C01245. Like C01245, it is not a ligand for ITPR1. It binds with PKC that takes part in controlling plasma membrane based Ca^{2+} ion channels. But we are able to decipher its role clearly only after its merging with module *C01245*. Thus it appears that we can associate biological significance to these modules [2].

3.3.6 Comparative study on modules of Calcium STP of 7 species

Modularization algorithm is applied to Ca^{2+} STP of 7 different species (bta, cfa, hsa, mmu, ptr, rno, and ssc) for $c = 3$. Now we discuss about the species-specific modules one after another for these species. In *P. troglodytes (Chimpanzee)*, *(C00076)2* module is under developed. Role of endoplasmic reticulum in Ca^{2+} ion balance is negligible. However, mitochondria plays significant role in maintaining Ca^{2+} ion balance. The module exists in 2 parts (*(C00076)2* and *469986*. Absence of coordination between plasma membrane based ion channels, endoplasmic reticular receptors and mitochondria indicates negligible role of Ca^{2+} ion in signal transduction for *P. troglodytes*. The rest modules that we found in *H. sapiens* are not present. In case of *C. familiaris (dog)*, module *(C00076)2* shows role of Ca^{2+} ions in muscle contraction, a fact not shown in the same module for chimpanzee. In addition, part of module *C01245* is detected. In *S. scrofa (pig)* for the first time, plasma membrane, endoplasmic reticulum and mitochondria are coordinating with each other to maintain Ca^{2+} balance in module *(C00076)2*. Still it is a far cry from Ca^{2+} signaling mechanism of *H. sapiens*. Here 2 members

of module *(C00076)1* are also detected. For Ca^{2+} STP of *B. taurus (cow)*, we get 5 modules. The original *(C00076)2* module exists in 3 parts (module *(C00076)2*, module *GNAS1* and module *CD38*) with several members missing. Module *C01245* is fully developed but contains 2 members of the less developed *(C00076)1* module. Partly developed module *CALM2* shows role of Ca^{2+} binding proteins. As a whole Ca^{2+} STP of *B. taurus* shows highest similarity with that of *H. sapiens* among all. As Ca^{2+} STP of *H. sapiens*, *R. norvegicus* and *M. musculus* are identical, we get similar modules for these three species. Table 3.4 gives data about modules present in Ca^{2+} STP of these 7 different species. Analysis of these modules obtained from different species lead us to a conclusion that *H. sapiens*, *R. norvegicus* and *M. musculus* have highly developed Ca^{2+} signaling mechanisms, *B. taurus* and *S. scrofa* lie as intermediates. But that of *C. familiaris* and *P. troglodytes* is very much under developed [2].

Table 3.4: Modules obtained from Ca^{2+} STP of different species

hsa,rno,mmu		bta		ssc		cfa		ptr	
name	nodes	name	nodes	name	nodes	name	nodes	name	nodes
(C00076)2	29	(C00076)2	13	(C00076)2	9	(C00076)2	5	(C00076)2	4
CALML6	08	C01245	10	CHRM1	2	GNAS	3	469986	3
(C00076)1	07	CALM2	5	(C00076)1	2	PTGER3	2		
C01245	10	CD38	4						
		GNAS1	5						

This table contains information about modules obtained from Ca^{2+} STP of *H. sapiens* (human), *R. norvegicus* (rat), *M. musculus* (mouse), *B. taurus* (cow), *S. scrofa* (pig), *C. familiaris* (dog)and *P. troglodytes* (chimpanzee).

3.3.7 Modularization of Wnt STP

Human Wnt STP (Figure 2.4) contains 60 nodes crisscrossed among themselves by 70 relations. Here, a discussion is pursued to find the ideal c-value for partitioning the human Wnt STP. We found varied number of modules for different c-values [1-13], c being an integer. The set of modules obtained for $c = 13$ remains unaltered for further higher values of c. While scrutinizing the modules, we concisely tried to shrink this range of c-value to get an ideal single value of c. Detailed description about the modules is given in Table 3.5.

Table 3.5: List of modules obtained from human Wnt STP for different c-values

Module Name	c=1	c=2	c=3	c=4	c=5	c=6	c=7 & 8	c=9, 10 & 11	c=12 & 13
LEF1	11	14	14	14	14	14	14	14	51
CTNNB1	5	7	8	9	11	11	14	37	-
(DVL1)1	4	6	7	11	13	26	23	-	-
(DVL1)2	2	8	10	10	10	-	-	-	-
WNT16	4	5	8	4	-	-	-	-	-
AXIN1	4	4	4	3	3	-	-	-	-
PLCB1	3	7	7	7	7	7	7	7	7
RHOA	3	2	-	-	-	-	-	-	-
(FZD10)1	2	3	-	-	-	-	-	-	-
GSK3B	2	2	-	-	-	-	-	-	-
PSEN1	2	-	-	-	-	-	-	-	-
MAP3K7	2	-	-	-	-	-	-	-	-
Tp53	2	2	2	2	2	2	2	2	2
(FZD10)2	2	-	-	-	-	-	-	-	-
RAC1	2	-	-	-	-	-	-	-	-
(FZD10)3	2	-	-	-	-	-	-	-	-
DKK1	2	-	-	-	-	-	-	-	-
APC2	2	-	-	-	-	-	-	-	-
CHP	2	-	-	-	-	-	-	-	-
VANGL2	2	-	-	-	-	-	-	-	-

This table contains information about the modules obtained from human Wnt STP for different c-values. The first column contains the names of the modules while the remaining columns provide number of nodes present in those modules for all the considered c-values [1-13]. For c-values of 7 and 8, 9, 10 and 11, and 12 and 13, there is no change in the number of modules as well as their size and topology.

Modules $LEF1$, $PLCB1$ and $Tp53$

Modularization starts with the node having maximum relations in human Wnt STP, *i.e.*, LEF1. For $c = 1$, module $LEF1$ is made up of 11 nodes, that roughly includes the nuclear part of the canonical Wnt STP. This initial module is devoid of NEMO-Like Kinase (NLK), a target of TAK1 (a MAP kinase kinase kinase), and the biomolecules that are related to LEF1. For $c = 1$, module $LEF1$ cannot be explained from biological significance point of view. So, c-value of 1 cannot be considered as ideal. The module size increases to 14 for $c = 2$ and remains stagnant up to c-value of 11. For $c = 12$ and 13, $LEF1$ contains 51 nodes, including the whole canonical and planar call polarity STPs. Such a big module is difficult to validate. So, modules for $c = 12$ and 13 cannot be considered as ideal. Hence, the ideal c-value should range in [2-11].

Modules $PLCB1$ contains 3 nodes for $c = 1$ and then remains unchanged in size (7 nodes) for the rest of the c-values (2-13). It represents the Wnt/Ca^{2+} STP, that leads to release of intracellular Ca^{2+} ions, via G-proteins. Module $Tp53$ (2 nodes) remains unaltered for all the c-values (1-13).

Module $CTNNB1$ and $(DVL1)1$

The module $CTNNB1$ starts with CTNNB1 molecule and its 5 neighboring nodes. This module's size increases up to the c-value of 6, as given in Table 3.5. For $c = 7$ and 8, the module's size (14 nodes) remained the same. Similar is the case for $c = 9$, 10 and 11, where it's size increases to 37 nodes. For $c = 13$, it merges with another module creating another bigger module. Such large modules are unwanted as it becomes tricky to establish their biological significance. So the ideal range for c-value shrinks to [2-8].

Module $(DVL)1$ is a part of the canonical Wnt STP, which is present in the protein phosphatase 2 protein scaffold. Its size varies from 4 to 26 nodes with c-value in [1-8]. For c-values higher than 8, this module merges with some other module resulting in a large module (Table 3.5). In fact, for c-values of 6, 7 and 8, module $(DVL)1$, becomes large which is not desirable. So, we restrict the c-value to be in [2-5].

Module $(DVL1)2$, $AXIN1$ and $WNT16$

Module $(DVL1)2$ presents the planar cell polarity pathway. The module starts with 2 nodes ($c = 1$) and increases up to 10 nodes ($c = 3$, 4 and 5) as given in Table 3.5. A higher value of c, compels this subpathway to merge with the canonical Wnt STP. We made a conscious effort not to mix the three subpathways of Wnt STP while creating the modules. So, the range for ideal c-value should lie in [2-5].

Module $AXIN1$ is also a part of the canonical Wnt STP (the protein scaffold region). Its size (4 nodes) remains unaltered for $c = 1$, 2 and 3. For $c = 4$ and 5, its size decreases to 3 nodes, a rare case where the size decreases with increase in c-value. However, this fact does not help in finding the ideal c-value.

Module $WNT16$ comprises the starting ligand-receptor binding steps of the canonical Wnt STP. For $c = 5$, this module is not present. So, the ideal c-value must be either 2 or 3 or 4. For c-value of 4, the size of module $AXIN1$' decreases from 4 nodes to 3 nodes, while more members are present in reality in this process. Hence, the ideal c-value should be 2 or 3. For $c =$

2, the problem of over-splitting arises with 11 small modules. So the ideal c-value is 3 for partitioning the human Wnt STP.

3.3.8 The best set of modules of Wnt STP

The human Wnt STP is modularized into 8 modules for $c = 3$ [3]. The separate modules can be seen in Figure 3.12.

1. The canonical STP around cell membrane is included in module $WNT16$.
2. The steps of the canonical Wnt STP taking place around the PP2AC protein scaffold is divided into modules $(DVL1)1$, $AXIN1$ and $CTNNB1$.
3. The part of canonical Wnt STP around and inside nucleus is enclosed in module $LEF1$.
4. Module $(DVL1)2$ represents the planar cell polarity pathway.
5. Module $PLCB1$ corresponds to the Wnt/Ca^{2+} pathway.
6. We could not associate any biological significance to the very small module $Tp53$.

3.3.9 Module conservation among Wnt STPs of 31 species

Here, we have compared modules of Wnt STPs of 31 different species (aag, aga, ame, api, bfo, bmy, bta, cbr, cel, cin, cfa, dan, dme, dpo, dre, ecb, gga, hsa, mcc, mdo, mmu, nve, oaa, ptr, rno, ssc, spu, tad, tca, xla and xtr). Module details are given in Table 3.6. Comparison of the modules brought forward functional conservation among the taken species. Modules Wnt and β-catenin are found to be conserved in 9 species (hsa, mmu, rno, bta, cfa, ptr, mcc, mdo and gga). Module TCF is conserved in 5 species (hsa, mmu, rno, bta and cfa). Module $Tp53$ is conserved in 12 species (hsa, mmu, rno, bta, cfa, ptr, mcc, gga, dre, xla, xtr and ecb). Module $(DVL)2$ is conserved in 11 species (hsa, mmu, rno, bta, cfa, ptr, mdo, gga, dre, spu and dme) and module PLC is the most conserved module, being present in a maximum

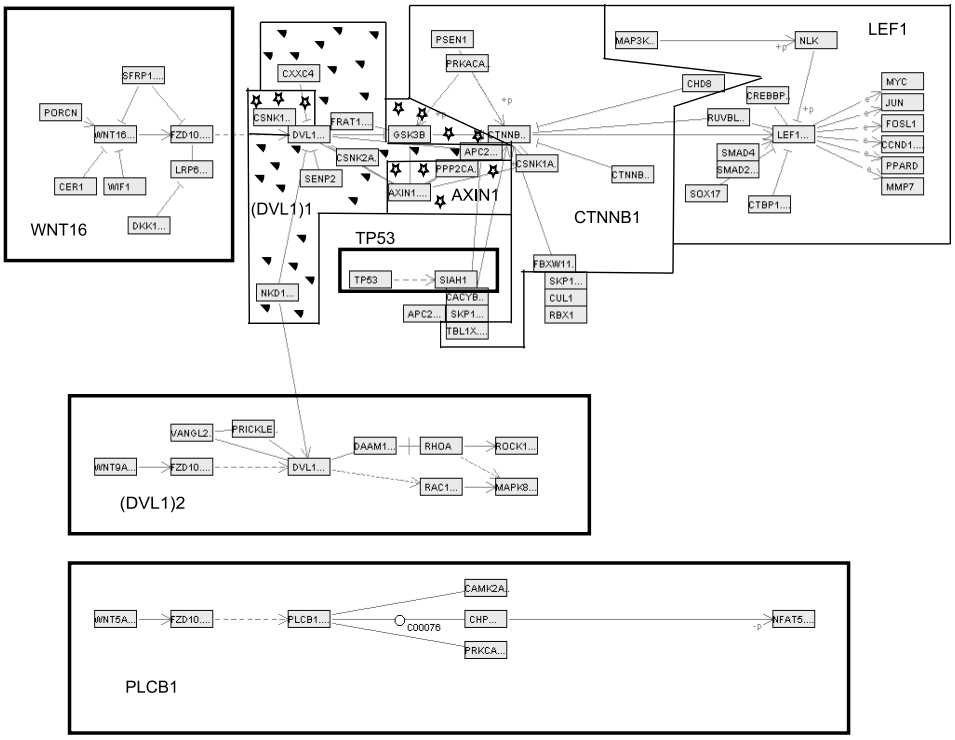


Figure 3.12: Modules of human Wnt STP for $c = 3$

The pathway is divided into 8 modules after modularization. The cell membrane potion of the canonical Wnt STP is included in module $WNT16$. The steps of the canonical Wnt STP taking place around the PP2AC protein scaffold is divided into modules $(DVL1)1$, $AXIN1$ and $CTNNB1$. The part of canonical Wnt STP around and inside nucleus is enclosed in module $LEF1$. Module $(DVL1)2$ is made up of the planar cell polarity pathway. Module $PLCB1$ comprises of the Wnt/Ca^{2+} pathway. We cannot associate any biological significance to the very small $Tp53$ module.

number of 17 species (hsa, mmu, rno, bta, cfa, ptr, mcc, mdo, gga, dre, xla, spu, xtr, dme, ecb, nve and ame) [3].

3.4 Conclusive remarks

We have presented here a novel partitioning algorithm, called Modularization algorithm. Modularization algorithm is dedicated towards finding biologically significant modules from signal transduction networks. The algorithm

Table 3.6: Module information of species-specific Wnt STPs

sp.	n	r	t	WNT	(DVL)1	Axin	β-catenin	TCF	p53	(DVL)2	PLC
hsa	60	70	8	WNT [8]	(DVL)1 [7]	Axin [4]	β-catenin [8]	TCF [14]	p53 [2]	(DVL)2 [10]	PLC [7]
mmu	60	70	8	WNT [8]	(DVL)1 [7]	Axin [4]	β-catenin [8]	TCF [14]	p53 [2]	(DVL)2 [10]	PLC [7]
rno	59	69	8	WNT [7]	(DVL)1 [7]	Axin [4]	β-catenin [8]	TCF [14]	p53 [2]	(DVL)2 [10]	PLC [7]
bta	58	68	8	WNT [7]	(DVL)1 [6]	Axin [4]	β-catenin [8]	TCF [14]	P53 [2]	(DVL)2 [10]	PLC [7]
cfa	58	68	8	WNT [8]	(DVL)1 [7]	Axin [4]	β-catenin [7]	TCF [13]	p53 [2]	(DVL)2 [10]	PLC [7]
ptr	58	67	8	WNT [8]	(DVL)1 [7]	Axin [4]	β-catenin [8]	TCF [13]	p53 [2]	(DVL)2[10]	PLC [6]
mcc	55	63	8	WNT [7]	(DVL)1 [6]	Axin [4]	β-catenin [8]	TCF [13]	p53 [2]	(DVL)2 [8]	PLC [7]
mdo	54	64	7	WNT [8]	(DVL)1 [7]	Axin[2]	β-catenin [9]	TCF [11]	-	(DVL)2 [10]	PLC [7]
gga	54	63	8	WNT [7]	(DVL)1 [6]	Axin [3]	β-catenin [8]	TCF [11]	p53 [2]	(DVL)2 [10]	PLC [7]
dre	52	60	7	WNT [8]	-	Axin [4]	β-catenin [7]	TCF [13]	p53 [2]	(DVL)2 [11]	PLC [7]
xla	43	45	6	WNT [7]	-	-	β-catenin [8]	TCF [11]	p53 [2]	(DVL)2 [8]	PLC [7]
spu	39	45	6	-	(DVL)1 [7]	Axin [2]	β-catenin [5]	TCF [10]	-	(DVL)2 [9]	PLC [6]
xtr	37	36	6	WNT [3]	-	-	β-catenin [7]	TCF [6]	p53 [2]	(DVL)2 [12]	PLC [7]
dme	36	42	7	WNT [6]	(DVL)1 [5]	Axin [2]	β-catenin [6]	TAK1 [2]	-	(DVL)2 [9]	PLC [6]
ecb	36	38	7	(Frizzled)1 [5]	(DVL)1 [3]	-	β-catenin [9]	TAK1 [2]	p53 [2]	(DVL)2 [8]	PLC [7]
nve	32	33	6	(Frizzled)1 [5]	-	Axin [5]	β-catenin [6]	TAK1 [2]	-	(DVL)2 [7]	PLC [7]
ame	30	32	5	-	(DVL)1 [4]	-	β-catenin [9]	TAK1 [2]	-	(DVL)2 [8]	PLC [7]
dpo	28	30	4	(Frizzled)1 [8]	-	-	β-catenin [7]	-	-	(DVL)2 [8]	PLC [5]
tca	26	27	4	(Frizzled)1 [7]	-	-	β-catenin [7]	-	-	(DVL)2 [6]	PLC [6]
aag	24	22	4	(Frizzled)1 [4]	-	-	β-catenin [4]	-	-	(DVL)2 [10]	PLC [6]
oaa	22	22	4	WNT [2]	-	-	β-catenin [7]	-	-	(DVL)2 [8]	PLC [5]
cel	22	20	3	-	-	-	β-catenin [10]	-	-	RhoA [6]	PLC [6]
aga	20	18	3	(Frizzled)1 [11]	-	-	β-catenin [4]	-	-	-	PLC [5]
ssc	19	16	4	FRP [2]	-	-	β-catenin [5]	TCF [7]	-	-	PLC [5]
bfo	18	16	3	-	-	-	β-catenin [9]	-	-	(DVL)2 [5]	PLC [4]
cin	17	14	3	-	(DVL)1 [7]	-	-	-	-	(DVL)2 [5]	PLC [5]
dan	16	12	4	-	(DVL)1 [2]	-	β-catenin [4]	-	-	(DVL)2 [5]	PLC [5]
bmy	13	11	3	-	(DVL)1 [4]	-	-	-	-	(DVL)2 [5]	PLC [4]
api	13	10	3	-	(DVL)1 [4]	-	-	-	-	(DVL)2 [5]	PLC [4]
tad	6	4	2	-	-	-	-	-	-	Rac [2]	PLC [4]
cbr	4	3	1	-	(DVL)1 [4]	-	-	-	-	-	-

The modules are created for $c = 3$. Each module's size in terms of nodes is given with it in parentheses. The table throws light on the developmental trend of Wnt STPs among the set of species under consideration. (sp.: three lettered species code, n: number of connected nodes in a species-specific pathway; r: number of relations present the connected component of a species-specific pathway; t: total number of modules created from a species-specific pathway)

has successfully identified functional modules from MAPK, Ca^{2+} and Wnt STPs by centralizing the mostly connected nodes. Some significant modules have been obtained from the human MAPK STP for c-value of 3. The comparative study of MAPK STPs among the taken 7 species have shown gradual development of the pathway from *P. troglodytes* to *H. sapiens* via *S. scrofa*, *P. troglodytes*, *B. taurus* and *R. norvegicus*. We have successfully conferred biological significance to the modules obtained from human Ca^{2+} STP for complexity level (c-value) of 3. The comparative study indicates gradual increase in development of Ca^{2+} STPs starting from *P. troglodytes* to *H. sapiens* via *C. familiaris*, *S. scrofa*, *B. taurus*, *M. musculus* and *R. norvegicus*. Eight modules have been obtained from the human Wnt STP for $c = 3$. Modules of 31 species-specific Wnt STPs have been compared to find conservation among them for the same c-value. Module PLC has been found to be the most conserved module, being present in a maximum number of 17 species.

When a pathway is difficult to analyze as a whole in certain species, or certain module(s) of it is(are) only functional for a set of species, modularized study is quite helpful. The algorithm does not require any predefined cut-size like the graph partitioning techniques. Unlike community finding algorithms, it creates modules even if there is no naturally existing partitions in a biological network. It can also be applied to other kinds of biochemical pathways with appropriate modifications, if needed.

Moreover, a signal transduction pathway can be perceived as a black box operating with many layers, where the input and output of each layer is known through laboratory experiments. But what exactly happens inside the black box and the way these layers co-ordinate among themselves is difficult to grasp. Probably the task will be easier if we try to understand the mechanism of the black box layer by layer and try to trace a particular input through various layers of the black box till we reach the output. Here our created modules are equivalent to layers of the black box. Now it may happen that a particular input may or may not be involved with all the nodes of intermediate layers of the black box. Likewise, in a signal transduction pathway, the input signal may not involve all the nodes present in all the

modules. These ideas may lead to a better design of an artificial system that can successfully mimic biological pathways. Performance of the Modularization algorithm in comparison with some traditional graph partitioning and community finding algorithms will be discussed in Chapter 4.

Chapter 4

Comparison of Some Partitioning Algorithms

4.1 Introduction

In the last Chapter, we have developed an algorithm for modularization of a signal transduction pathway. The present Chapter will describe an extensive comparative analysis of the Modularization algorithm with some existing traditional partitioning algorithms. For this comparative analysis, we have considered human Wnt STP. The comparison will be made using functional enrichment scores described in Section 4.4.

There exist various approaches for partitioning networks/pathways. They include hierarchical clustering techniques [176–178], graph partitioning techniques [179, 180], community finding algorithms [186–188, 212, 217, 224, 226], block modeling methods [181], differential equation based methods [182] and cartographic representations [183]. Among them, approaches based on graph partitioning [179, 180, 185] and community structure detection methods [186–188] are popular. Algorithms based on these two concepts have been vastly used to divide, study and analyze networks. Some of these methods have also been applied to biological networks. Here, we have considered a few graph partitioning and a community finding algorithms for comparison.

Newman et al. have proposed a series of algorithms [186–188, 212, 217, 224, 226] to find communities in various kinds of networks. These algorithms optimize a network's divisions based on the properties of the network itself. We have compared our method with a community finding algorithm of Newman [187], which has already been applied to metabolic pathways along with other kinds of networks. Newman's algorithm optimizes a quality function known as "modularity" over possible divisions of a given network. Modularity score is directly dependent on the network architecture in terms of adjacency matrix and eigenvalues of a symmetric matrix calculated from the adjacency matrix. Positive value of modularity indicates possible presence of modules in a network. One important aspect of the algorithm is that it refuses to divide a network if no good division exists. In other words, a negative value of modularity indicates no possible division of the given network. Throughout this chapter we have referred this algorithm as Newman's community finding algorithm.

In this Chapter, we have compared performance of five partitioning algorithms including our proposed modularization algorithm for creating modules from human Wnt STP (Figure 2.4). For this purpose, we have formulated a scoring method based of Gene Ontology (GO) [328] attributes. The scoring method emphasizes on goodness of attributes rather than number of valid attributes. An attribute is a GO term that emphasizes particular characteristics of a gene, *i.e.*, process, position inside cell and function. We have used the Biological Networks Gene Ontology tool (BINGO) [329] for comparing performance among Greedy [247], Farhat's [248], Modularization [1], Newman's community finding [187] and Kernighan-Lin's [249] algorithms for partitioning the human Wnt STP. Implementation of these algorithms has been done in C and Matlab (Version 7.0.4).

4.2 Human Wnt STP

The Wnt STP is an ancient and evolutionarily conserved pathway. As described earlier, Wnt molecules are secreted cysteine-rich, lipid-modified glycoproteins. They bind to FZD receptors along with co-receptor LRPs and initiate the downstream steps, those altogether are known as Wnt STP [29]. They comprise a large family of nineteen proteins in humans hinting to a daunting complexity of signaling regulation, function and biological output [330]. Wnt STP proceeds through three separate generalized pathways, namely, the β-catenin, Ca^{2+} and planar cell polarity pathways. Wnt STPs are involved in regulation of cell fate determination, proliferation, differentiation, migration, apoptosis [304, 305] and regulation of bone mass [306] among others. It enables cells to influence behavior of their neighboring cells during embryonal development [30, 111, 307]. In matured organisms, Wnts are implicated in maintaining stem cell-like fates in the intestinal epithelium [308], skin [309] and hematopoietic cells [310]. In Chapter 2, the Wnt STP is described in detail.

Human Wnt STP (Figure 2.4) contains 80 nodes in total, as given in the KEGG database [167]. Each node corresponds to a biomolecule taking part in the overall process of the pathway. The connected component of the

pathway is found to be a network of 60 nodes crisscrossed among themselves by 70 relations. Each of these relations depict association among two of the aforementioned biomolecules. This connected component is used as data.

4.3 Algorithms

We have used the Biological Networks Gene Ontology tool (BINGO) [329] for comparing performance among Modularization [1], Newman's community finding [187], Greedy [247], Farhat's [248], and Kernighan-Lin's [249] algorithms. Implementation of these algorithms is done in C and Matlab (Version 7.0.4). A description of the Modularization algorithm with its pseudocode is already furnished in Chapter 3. Brief descriptions of the remaining algorithms along with their pseudocodes are provided in upcoming subsections.

4.3.1 Newman's community finding algorithm

Newman's community finding algorithm [187] is based on the fact that "a good division of a network into communities is not merely one in which there are few edges between communities; it is one in which there are fewer than expected edges between communities". It optimizes a quality function known as modularity over the possible divisions of a network. Newman showed that modularity can be expressed in terms of the eigenvectors of the characteristic matrix of a network, known as the 'modularity matrix'. Based on this matrix, Newman's community finding algorithm generates results of demonstrably higher quality in shorter running times.

4.3.2 Greedy algorithm

Greedy algorithm [247] tries to minimize the cut-size[1]. It chooses one starting vertex first. Then the next vertex that increases the cut-size by the least amount gets added while growing the partition. The procedure continues till the partition is big enough to permit creation of user defined number of

[1]In graph theory, a cut is a partition of the vertices of a graph into two disjoint subsets. Hence cut-size is the number of edges ignored to create a partition.

2 Pseudo code for Newman's Community Finding Algorithm

STEP 1:
Construct Modularity Matrix
Find leading eigenvalue and corresponding eigenvector of the matrix
Divide the network into two partitions according to the signs of the elements of this vector
repeat
  STEP 1 for each of the partition
until The proposed split makes zero or negative contribution to the total modularity

partitions. Here, a user has to predefine the number of partitions wanted from the network. Then the whole process gets repeated on the rest part of the graph to create further partitions.

3 Pseudo code for Greedy Algorithm

Unmark all vertices
Choose a pseudo-peripheral vertex, mark it and add it to current partition
FOR the desired number of partitions DO
repeat
  Among all the unmarked vertices adjacent to the current partition, choose the one with least number of unmarked neighbors
  Mark it and add it to the current partition
until the current partition is big enough
IF there are unmarked vertices left THEN
Choose the one adjacent to the current partition with least number of unmarked neighbors as starting vertex for next partition

4.3.3 Farhat's algorithm

Farhat's algorithm [248] chooses the starting vertices of each partition in a greedy way. In comparison with greedy algorithm, it is fast and avoids recursive bisection. It is also able to divide a network into any number of partitions when the desired number of partition is not a power of 2. But, quality of the partitions (measured in terms of cut-size) is sometimes questionable. Also the algorithm tends to be sensitive to the choice of the starting vertex.

4 Pseudo code for Farhat's Algorithm

```
Unmark all vertices
FOR the desired number of partitions DO
Among all the vertices chosen for the last partition, choose the one with the
smallest non-zero number of unmarked neighbors, mark all these neighbors
and add them to current partition (Pseudo-peripheral vertex is taken as
starting vertex)
repeat
  FOR each vertex v in the current partition DO
  FOR each unmarked neighbor x of v DO
  Add x to the current partition
  Mark x
until the current partition is big enough
```

4.3.4 Kernighan-Lin's algorithm

Kernighan-Lin's algorithm [249] is one of the earliest graph partitioning algorithms. It takes random partitions of a network as input. Then the algorithm improves the partitions iteratively by minimizing the edge cut-size. It first calculates the gain in the reduction of edge-cut for each vertex when it is moved from one partition of the graph to the other. Then it moves the unlocked vertex which has the highest gain, from the partition in surplus to the partition in deficit. This vertex is then locked and the gains are updated. This procedure is repeated till all the vertices are locked, even if the highest gain may be negative. The last few moves that had negative gains are then undone and the bisection is reverted to the one with the smallest edge-cut so far. The process continues iteratively and terminates finally, when any reduction cannot be achieved in the edge-cut. It is a local optimization algorithm, with a limited capability for getting out of local minima by the way of allowing moves with negative gain. It delivers reasonable results for small networks. But, it is quite inefficient for large networks. This drawback limits the algorithm's applicability towards large biological networks.

5 Pseudo code for Kernighan-Lin's algorithm

```
Given two sub-partitions
Compute the difference value all vertices
Unmark all the vertices
Take a random initial cut-size
Among all unmarked vertices, find the pair of nodes with the biggest gain
(which might be negative)
Mark the nodes
Update difference value as if the nodes had been swapped
Find the number (j) for which the next cut-size will be the minimum of
initial cut-size
Swap the first j node pairs
Continue until no further cut-size improvement can be achieved
```

4.4 Scoring Method

BINGO [329] is an open source java tool to determine the Gene Ontology (GO) terms that are significantly over-represented in a set of genes. GO [328] is a public consortium of databases that provides a controlled vocabulary of terms aiming at a gene's or a cluster of genes' biological annotations. It consists of three hierarchically structured sets of vocabularies that describe gene products in terms of their associated 'Biological Process (BP)', 'Molecular Function (MF)' and 'Cellular Component (CC)' information; 'Go Full (GF)' being the superset of these sets. BINGO runs as a plug-in to Cytoscape [331] and retrieves the relevant GO annotations and propagates them upward through the GO hierarchy, *i.e.*, any gene annotated to a certain GO category is also explicitly included in all parental categories. It tries to answer the basic question, "While sampling X genes (test set) out of N genes (reference set), what is the probability that x or more of these genes belong to a functional category C shared by n of the N genes in the reference set?" Hypergeometric test answers this question in the form of a P-value. P-values depict a created partition's capability to lie in one category of biological function. If a particular partition created by a partitioning algorithm returns more number of valid GO terms with lower P-values than the others, the algorithm is believed as a better algorithm for creating partitions. Based

on this assumption, we have designed the 'valid attribute score' [5].

Valid attribute-wise analysis takes into consideration the number of valid GO attributes that the algorithm in consideration gets as result from a query with respect to a background database. Here, we have considered GO attributes obtained with P-value of the order of 10^{-5} or smaller as valid. The threshold P-value is fixed in such a manner that some valid attributes from majority of the partitions can be collected. Counting the number of valid attributes that a partition is found to be associated with, is a well established way of determining the biological validity of that partition. Many clustering algorithms follow it as a comparative measure to establish their superiority among the others [332]. Here, we have considered three background databases, namely 'BP', 'CC' and 'GF'.

P-values provide a good indication about the prominence of a certain functional category. But, no index of validity exists among the valid GO terms with lower P-values. Are they all equally valid or some of them are more valid than the others? Does such an index effect comparative results? By devising a validity index ('functional enrichment score'), we have showcased the change in results. Functional enrichment score-wise analysis takes into account functional enrichment scores of a set of partitioning algorithms. The functional enrichment score S_A of an algorithm A is defined as the mean of enrichment scores of the partitions it has created, and is given by

$$S_A = \frac{1}{p}\sum_{i=1}^{p} S_{P_i} \tag{4.1}$$

where p is the number of partitions obtained by algorithm A. In turn the enrichment score S_{P_i} of a partition P_i is the average of the individual enrichment scores ($S_{T_{ij}}$s) of associated individual (j^{th}) attributes (T_{ij}s). Thus S_{P_i} is given by

$$S_{P_i} = \frac{1}{q}\sum_{j=1}^{q} S_{T_{ij}} \tag{4.2}$$

where q is the number of attributes in partition P_i. Enrichment score $S_{T_{ij}}$

of an individual attribute T_{ij} is calculated by comparing the performance of algorithm A with the performance of a background database in detecting over-expressed gene categories associated with the attribute T_{ij}. $S_{T_{ij}}$ depicts the efficiency of the partitioning algorithm in placing nodes having a common attribute in a partition with respect to a background database. Let x be the number of nodes associated with an attribute T_{ij}, which lies in a partition P_i, and X ($\geq x$) be the number of nodes present in partition P_i. Then x/X is the ability of an algorithm for placing nodes in a partition that are associated with attribute T_{ij}. Let y be the number of nodes associated with an attribute T_{ij} in a background database, and Y ($\geq y$) be the number of attributes in that database. Then y/Y is the ability of the background database to associate genes to attribute T_{ij}. Thus $S_{T_{ij}}$ may be defined as

$$S_{T_{ij}} = \frac{x}{X} / \frac{y}{Y} \tag{4.3}$$

Higher the value of S_A, better is the algorithm for creating significant partitions.

In other words, we have taken the ratio of the performance of an algorithm with respect to the performance of a background database in assigning an attribute to a partition. Functional enrichment score is a measure to quantify the level of performance of an algorithm in creating biologically significant partitions. While comparing a few algorithms, higher the value of S_A, better is the algorithm for creating significant partitions. We have created three sets of enrichment scores, corresponding to three background databases, for each algorithm (Modularization, Newman's community finding, Greedy, Farhat's and Kernighan-Lin's) to get a better comparison.

4.5 Results

The best set of partitions created by each of the aforementioned algorithms are needed for the purpose of comparison. Hence, multiple sets of partitions are obtained by Modularization, Greedy and Farhat's algorithms. Every

individual set of partitions is evaluated by calculating their average functional enrichment score of associated valid attributes. The set of partitions having the highest functional enrichment score is deemed the best and used for comparison. Partitions obtained by the modularization algorithm (Figure 4.1) from human Wnt STP are already described in Chapter 3. Partitions of the rest algorithms are described here.

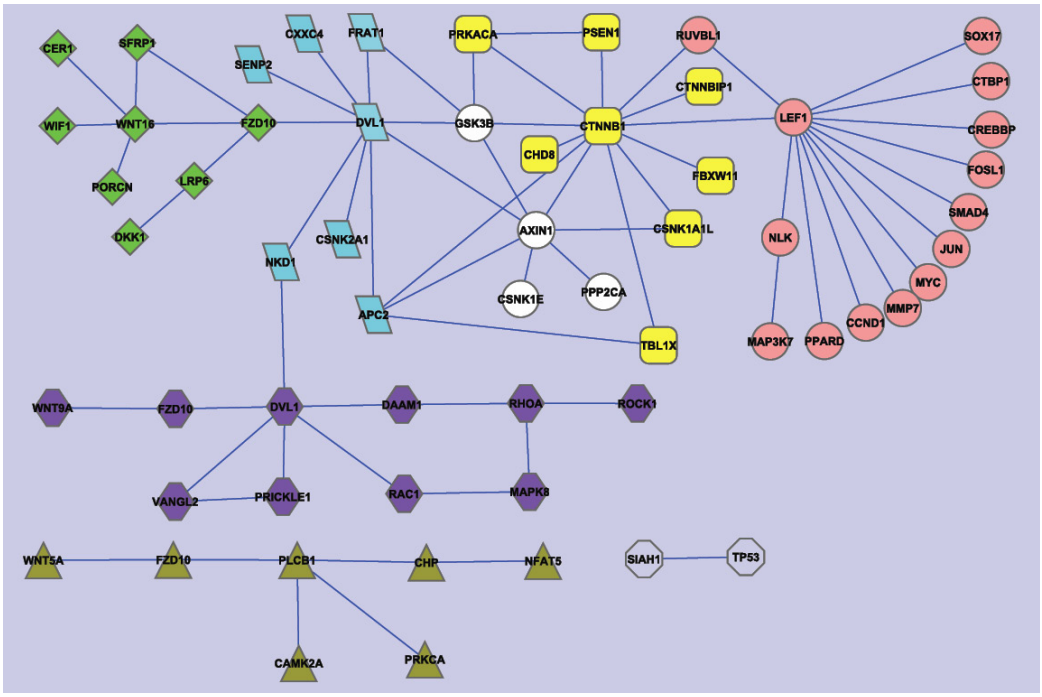


Figure 4.1: 8 partitions made by the Modularization algorithm
P: partition, P1: pink circles, P2: yellow round rectangles, P3: sulphate parallelograms, P4: green diamonds, P5: violet hexagons, P6: white circles, P7: dark green triangles and P8: blue octagons.

4.5.1 Partitions obtained by Newman's community finding algorithm

Newman's community finding algorithm [187] has created 8 partitions from human Wnt STP for modularity (Q) value of 0.6599 and ΔQ value of 1.0470e^{-017}. ΔQ is the user specified threshold to create partitions. A negative or zero ΔQ value prompts Newman's community finding algorithm to stop creating par-

titions as it does not improve the value of modularity (Q). Thus, the limiting condition for the algorithm to continue partitioning is $\Delta Q \geq 0$. However, we have found that the algorithm is unable to converge for a very small ΔQ value of 1.0470×10^{-18}. Hence, we have adjusted the threshold accordingly at a higher value of $1.0470\text{e} \times 10^{-17}$. The partitioned human Wnt STP is shown in Figure 4.2. The node details are given in Table 4.1. Partition 1 represents the planar cell polarity pathway, partition 2 comprises ligands of the canonical Wnt STP; partition 3 is a singleton partition comprising node DKK1; partition 4 represents the isolated relation between SIAH1 and Tp53, partition 5 is the Wnt/Ca^{2+} pathway; partition 6 embodies the part of canonical Wnt STP inside nucleus at gene level (LEF1 and its target genes, of which maximum double as proto-oncogenes), and partitions 7 and 8 includes steps of the canonical Wnt STP taking place around the PP2AC protein scaffold. When partitions obtained by Newman's community finding algorithm are compared with modules obtained by modularization algorithm, partition 1 is found to be the same as module 5, partitions 2 and 3 are subsets of module 4, partition 5 is the same as module 7, partition 6 is the same as module 1, partitions 7 and 8 are divided into modules 2, 3 and 6.

4.5.2 Partitions obtained by Greedy algorithm

A set of 9 partitions obtained by Greedy algorithm [247] has resulted in the highest average functional enrichment score. So, we have considered this set of partitions as the best set of partitions obtained by Greedy algorithm. The created partitions are shown in Figure 4.3 and the node details are given in Table 4.1. Partition 1 involves the mechanism in β-catenin binding complex. But many members like PPP2CA and CSNK1E do not lie in this partition. Ideally they should be in this partition as functionally they are part of the binding complex. Partition 2 is centered on DVL1 gene and its regulators and inhibitors. Wnts bind with FZD along with co-regulation of LRPs. So ideally both LRP6 and DKK1 should lie in this partition, which is not the case. Partition 3 is made up of some planar cell polarity pathway members. Other members of this subpathway are found to lie in 3 different

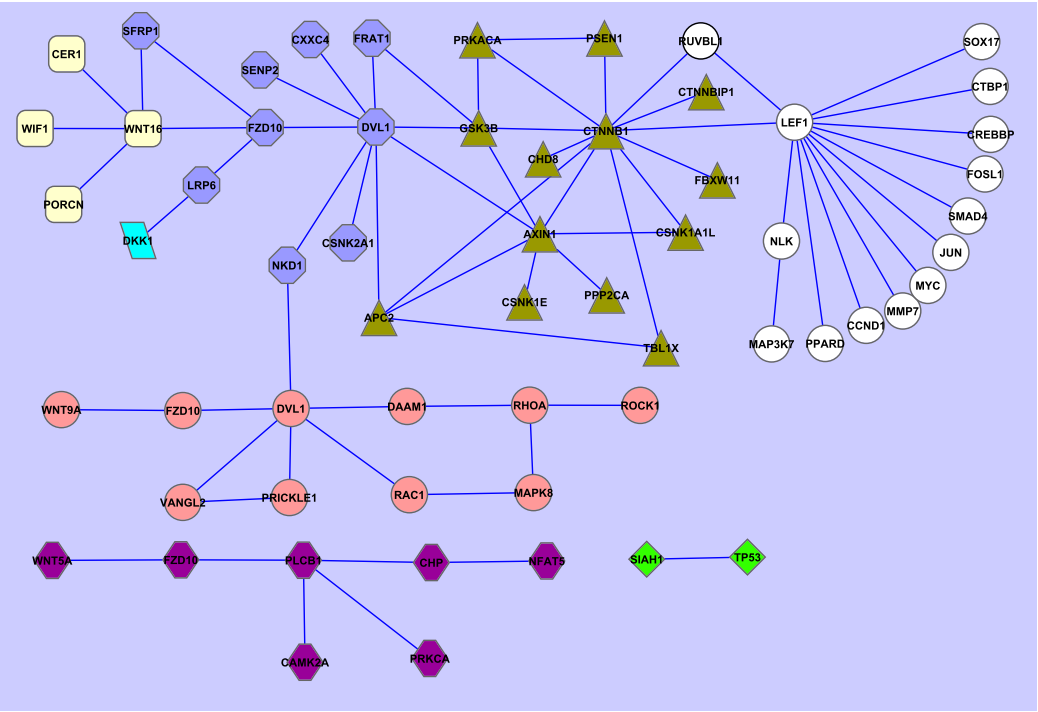


Figure 4.2: 8 partitions made by Newman's community finding algorithm
P: partition, P1: pink circles, P2: yellow round rectangles, P3: sulphate parallelograms, P4: green diamonds, P5: violet hexagons, P6: white circles, P7: dark green triangles and P8: blue octagons.

partitions indicating oversplitting of the network. Partition 4 is the weakest candidate to be considered as a module, as all of its nodes are randomly scattered in nucleus and cytoplasm. Partition 5 is effectively a cluster of genes, transcription factors, regulators and suppressors related to LEF1 that have implications towards oncological developments in humans in abnormal cases. But, some nodes like MAP3K7, CREBBP, CTBP1, FOSL1 are absent in this module. They are found to be members of Partition 6. Members of partition 7 are scattered all over the network and hence, the partition is associated with a little significance. Partition 8 is composed of the initiating steps of canonical Wnt STP, which starts with binding of stimulated Wnt ligand with FZD. Partition 9 is constituted by the members of Wnt/Ca^{2+} STP, except a few like RHOA, ROCK1 and MAPK37 that ideally should lie in other partitions.

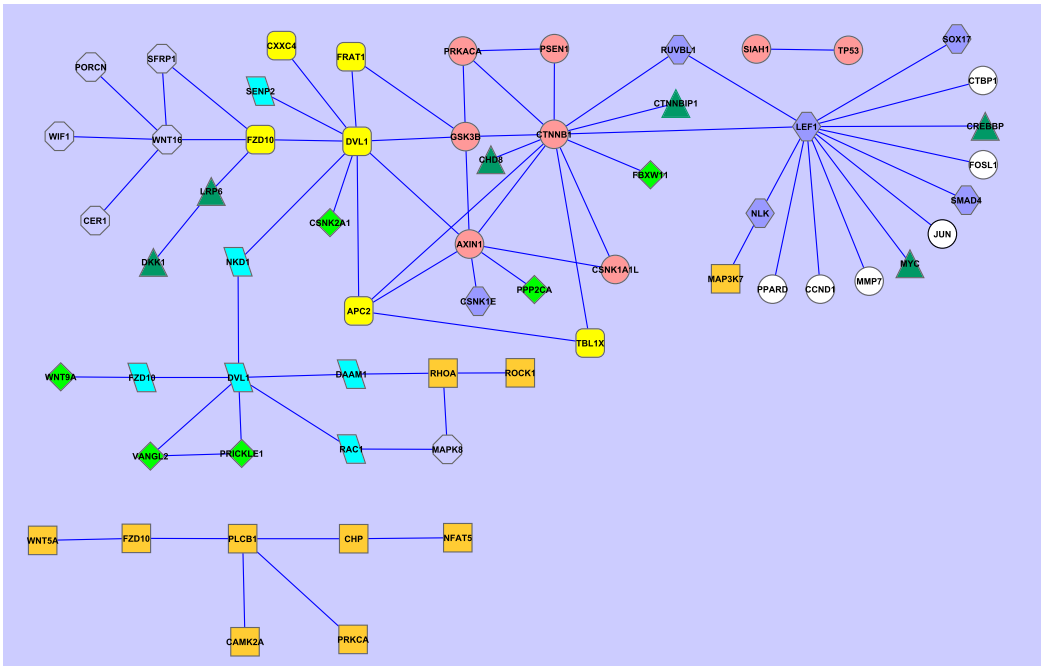


Figure 4.3: 9 partitions made by Greedy algorithm
P: partition, P1: pink circles, P2: yellow round rectangles, P3: sulphate parallelograms, P4: green diamonds, P5: violet hexagons, P6: white circles, P7: dark green triangles, P8: blue octagons and P9: orange rectangles.

4.5.3 Partitions obtained by Farhat's algorithm

A set of 11 partitions obtained by Farhat's algorithm [248] has resulted in the highest average functional enrichment score. So, we have considered this set of partitions as the best set of partitions. The partitions are shown in Figure 4.4 and node details are given in Table 4.1. Partition 1 depicts the β-catenin binding complex, but many nodes are missing. Partition 2 comprises some remaining nodes of the β-catenin binding complex. Partition 3 shows the initiating steps of canonical Wnt STP including LRP6. But, LRP should not be included in this partition as it regulates FZD and FZD is not a node of this module. Partitions 4, 6, 7 and 10 do not carry any biological significance, as they contain scattered nodes. Partition 5 depicts the activities of Wnt STP inside nucleus, yet the main node LEF1 is missing. Partition 8 is unbalanced due to its lack of functional connectedness. It contains some members of the

planar cell polarity pathway along with some from the canonical Wnt STP. Such kind of modules display poor structural integrity. Part of the planar cell polarity pathway lies in partition 9, but not all. The whole Wnt/Ca^{2+} pathway comes under partition 11, but on the other hand, it also contains 2 nodes of the canonical Wnt STP. Altogether, except partitions 4, 6, 7 and 10, the rest modules are nearer to being ideal modules. But the algorithm needs many improvements before it can be considered as a good partitioning algorithm for biological networks.

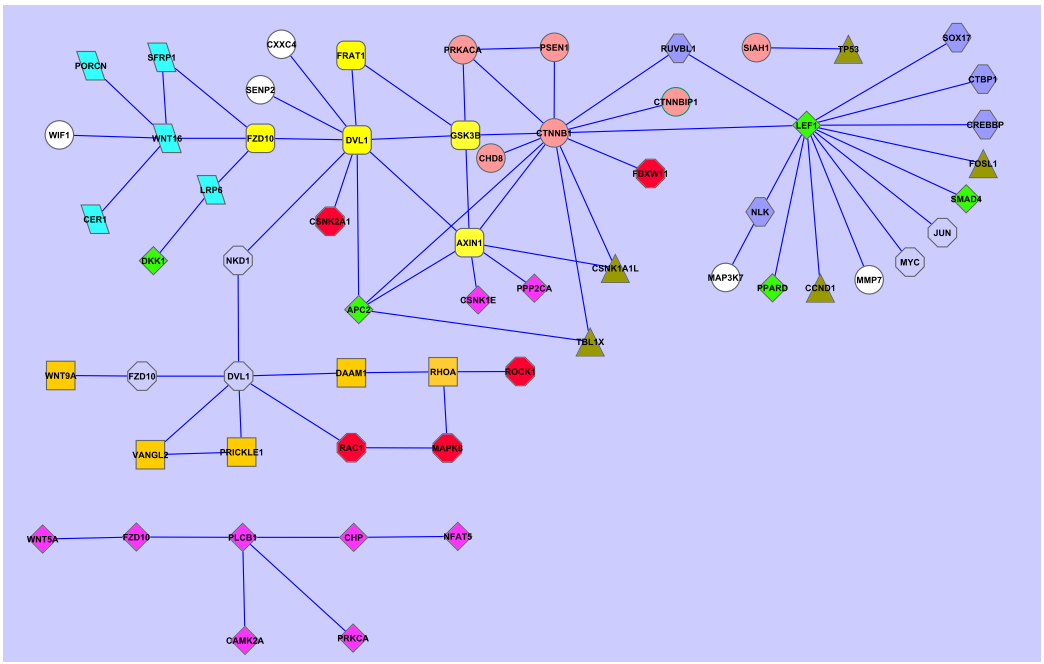


Figure 4.4: 11 partitions made by Farhat's algorithm
P: partition, P1: pink circles, P2: yellow round rectangles, P3: sulphate parallelograms, P4: green diamonds, P5: violet hexagons, P6: white circles, P7: dark green triangles, P8: blue octagons, P9: orange rectangles, P10: red octagons and P11: mauve diamonds.

4.5.4 Partitions obtained by Kernighan-Lin's algorithm

Here, we have used partitions obtained by greedy algorithm as input for Kernighan-Lin's algorithm [249]. Two partitions obtained from human Wnt STP by Kernighan-Lin's algorithm are shown in Figure 4.5, and the node

details are given in Table 4.1. Partition 1 contains some of the cell boundary members of canonical Wnt STP, the planar cell polarity pathway and Wnt/Ca^{2+} pathway. Partition 2 includes the β-catenin binding complex, and its down stream signal carriers that carry signal to the nucleus as well as to the pre-onco-complex components inside nucleus.

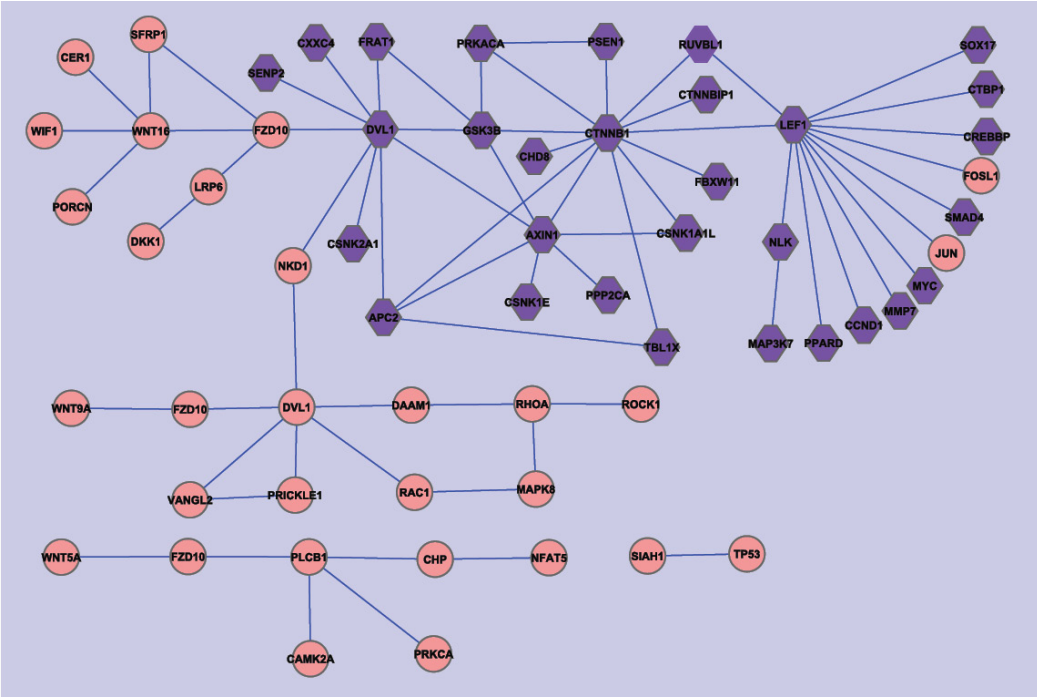


Figure 4.5: 2 partitions made by Kernighan-Lin's algorithm
P: partition, P1: pink circles and P2: violet hexagons.

When combined with other algorithms, Kernighan-Lin's algorithm may improve the original result. Here, initial partitions created by Greedy algorithm had a cut-size of 8 that got improved to 4 by this algorithm. But, the resulting partitions are huge. FOSL1 and JUN are nodes of partition 1, but ideally they should be in partition 2. It is not clear why the algorithm includes the disconnected edge between P53 and SIAH1 to partition 1.

However, we have compared the best set of modules obtained by the aforementioned algorithms by considering their association with GO attributes.

Table 4.1: The best sets of partitions created by different partitioning algorithms

Partition No.	Farhat	Greedy	Modularization	Newman	Kernighan-Lin
01	PSEN1, CTNNB1, PRKACA, CTNNBIP1, CHD8, SIAH1	PSEN1, CTNNB1, PRKACA, GSK3B, AXIN1, CSNK1A1L, SIAH1, TP53	LEF1, SMAD4, NLK, SOX17, CTBP1, CREBBP, RUVBL1, MYC, JUN, FOSL1, CCND1, PPARD, MMP7, MAP3K7	MAPK8, RAC1, ROCK1, RHOA, DAAM1, DVL1, PRICKLE1, FZD10, VANGL2, WNT9A	DKK1, PORCN, LRP6, CER1, WIF1, FZD10, WNT16, FZD10, SFRP1, WNT9A, VANGL2, PRICKLE1, WNT5A, PRKCA, DVL1, RAC1, DAAM1, FZD10, MAPK8, RHOA, ROCK1, PLCB1, CHP, CAMK2A, NFAT5, SIAH1, TP53, NKD1, JUN, FOSL1
02	GSK3B, DVL1, AXIN1, FRAT1, FZD10	APC2, DVL1, TBL1X, FRAT1, FZD10, CXXC4	CTNNB1, PSEN1, CTNNBIP1, CHD8, PRKACA, CSNK1A1L, FBXW11, TBL1X	WIF1, CER1, PORCN, WNT16	CCND1, MMP7, MYC, PRKACA, PSEN1, TBL1X, APC2, CTNNB1, GSK3B, AXIN1, FBXW11, DVL1, CSNK1E, PPP2CA, CTNNBIP1, CHD8, FRAT1, CXXC4, SENP2, CSNK2A1, CSNK1A1L, RUVBL1, MAP3K7, PPARD, NLK, SMAD4, LEF1, SOX17, CTBP1, CREBBP
03	WNT16, SFRP1, LRP6, PORCN, CER1	SENP2, NKD1, DVL1, FZD10, RAC1, DAAM1	DVL1, CXXC4, SENP2, CSNK2A1, FRAT1, APC2, NKD1	DKK1	–
04	DKK1, PPARD, APC2, SMAD4, LEF1	VANGL2, PRICKLE1, WNT9A, FBXW11, CSNK2A1, PPP2CA	WNT16, PORCN, FZD10, SFRP1, CER1, WIF1, LRP6, DKK1	SIAH1, TP53	–
05	NLK, SOX17, CTBP1, CREBBP, RUVBL1	CSNK1E, RUVBL1, LEF1, SMAD4, NLK, SOX17	DVL1, FZD10, RAC1, DAAM1, VANGL2, PRICKLE1, WNT9A, MAPK8, RHOA, ROCK1	NFAT5, PRKCA, CHP, CAMK2A, PLCB1, FZD10, WNT5A	–
06	MAP3K7, MMP7, WIF1, CXXC4, SENP2	CTBP1, MMP7, PPARD, CCND1, FOSL1, JUN	AXIN1, CSNK1E, GSK3B, PPP2CA	MMP7, PPARD, CCND1, FOSL1, JUN, MYC, CTBP1, SOX17, SMAD4, CREBBP, RUVBL1, NLK, MAP3K7, LEF1	–
07	APC2, TBL1X, CSNK1A1L, CCND1, FOSL1	MYC, CREBBP, CHD8, CTNNBIP1, LRP6, DKK1	PLCB1, FZD10, CAMK2A, CHP, PRKCA, WNT5A, NFAT5	CHD8, CTNNBIP1, CSNK1A1L, FBXW11, TBL1X, AXIN1, PPP2CA, APC2, CTNNB1, PRKACA, PSEN1, GSK3B, CSNK1E	–
08	JUN, MYC, NKD1, DVL1, FZD10	SFRP1, WNT16, PORCN, CER1, WIF1, MAPK8	TP53, SIAH1	NKD1, FRAT1, SENP2, DVL1, CSNK2A1, CXXC4, LRP6, FZD10, SFRP1	–
09	WNT9A, PRICKLE1, VANGL2, DAAM1, RHOA	RHOA, ROCK1, MAP3K7, WNT5A, FZD10, PLCB1, PRKCA, CHP, NFAT5, CAMK2A	–	–	–
10	MAPK8, ROCK1, RAC1, FBXW11, CSNK2A1	–	–	–	–
11	PPP2CA, CSNK1E, WNT5A, FZD10, PLCB1, PRKCA, CHP, NFAT5, CAMK2A	–	–	–	–

All the results are based on human Wnt STP data. Entries list the nodes in a partition. [Farhat's algorithm: 11 partitions; Greedy algorithm: 9 partitions; Modularization algorithm: $c = 3$, 8 modules; Newman's community finding algorithm: 8 partitions, $\Delta Q = 1.0470\text{e}^{-017}$; Kernighan-Lin's algorithm: 2 partitions, initial cut-size 8, final cut-size 4]

4.5.5 Using attributes

Here, the total number of valid attributes associated with the partitions obtained by an algorithm is taken as a measure of the partition's performance. Their overall performance is demonstrated in Figure 4.6. It shows that the Newman's community finding algorithm's partitions are returning maximum number of valid attributes (241, 25 and 343) with respect to all the three background databases namely 'BP', 'CC' and 'GF' followed by Kernighan-Lin's algorithm (65, 12 and 106). The next better performance is that of Modularization algorithm (37, 9 and 60) followed by Farhat's algorithm (31, 2 and 47). Greedy algorithm returns the least number of attributes (25, 3 and 39) among all. Newman's algorithm is appeared to be the best algorithm for creating partitions as they are found to be associated with the highest number of valid attributes [5]. Additional tables in support of this result are listed in Appendix.

But, a count of the valid attributes is not always enough. A deeper level study, as described in Figure 4.7, made us aware that small subsets of a large partition can be associated with many attributes. A large partition ensures presence of many subsets in it, which are associated with GO attributes; some of them being unique. Thus the corresponding P-values will be lower and they will be considered as valid. But only validity of an attribute is not sufficient for defining goodness of a module. Ideally, a valid attribute must be given more preference if it is associated with more number of nodes present in a partition than another one associated with less number of nodes in the same partition. In other words, we need to know the number of attributes that actually show some goodness (associated with more number of nodes) in justifying a partition. Hence, a functional enrichment score system has been defined to give weightage to valid attributes according to their goodness of performance.

4.5.6 Using Functional enrichment score

Functional enrichment score (S_A) depicts the efficiency of a partitioning algorithm in placing genes (having a common attribute) in a partition with

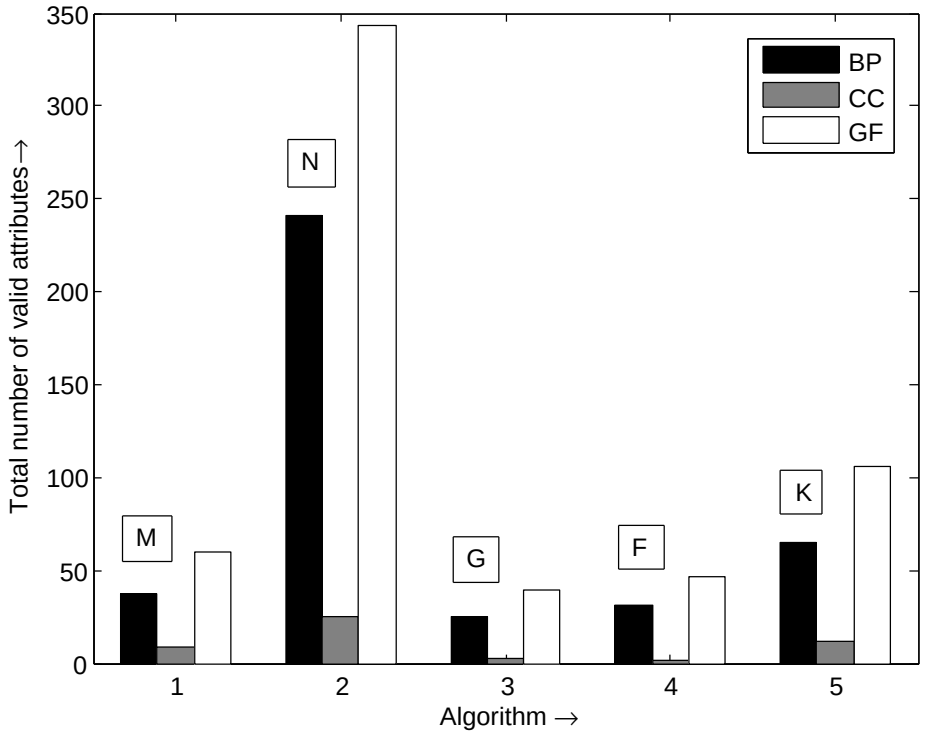


Figure 4.6: Comparison based on valid attribute score
BP- Biological Process, CC- Cellular Component, GF- GO Full, M- Modularization algorithm, N- Newman's algorithm, G- Greedy algorithm, F- Farhat's algorithm and K- Kernighan-Lin's algorithm.

respect to a background database. Higher the value of the score, better is the algorithm for creating partitions. The average enrichment scores (S_As) (Equation 4.1) of the different algorithms are shown in Figure 4.8. The figure depicted that the Modularization algorithm has performed the best among all the algorithms considered here. The algorithm has created partitions with average functional enrichment score of 163, 258, 274 approximately with respect to 'BP', 'CC' and 'GF' as background databases. Kernighan-Lin's algorithm has created partitions with the least average functional enrichment score preceded by Newman's algorithm. But, both the algorithms have created partitions for which many number of valid attributes are found to be associated (Figure 4.6). Hence, only counting valid attributes associated

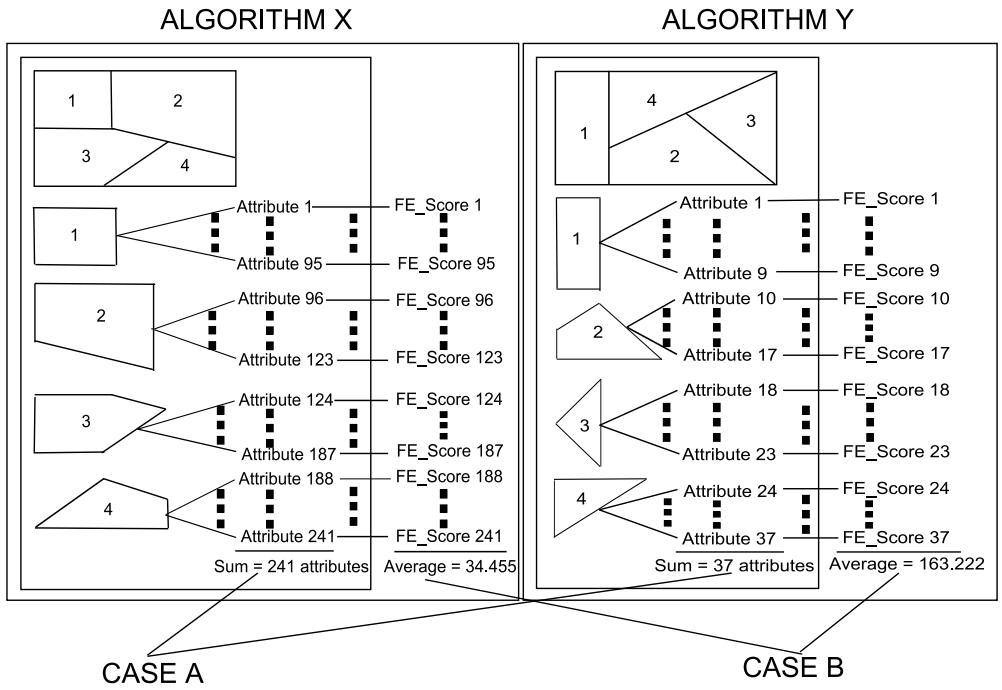


Figure 4.7: Methods of Algorithm Comparison
(CASE A) Comparison based on valid attribute score shows that algorithm X seems to be better than algorithm Y in creating partitions as the partitions are associated with more number of valid attributes. (CASE B) Comparison based on functional enrichment score of valid attributes shows that algorithm Y is better than algorithm X in creating partitions as the partitions are associated with some attributes, those have high association index (associated with more number of nodes in the partitions). Functional enrichment score is denoted as FE_Score .

with a partition is not a proper measure to deem that partition as good. Among the valid attributes, an association index must be established that can reflect goodness of a valid attribute. Functional enrichment scores reflect such an association index. Among Greedy and Farhat's algorithms, Greedy performed better for the background databases 'CC' and 'GF', while Farhat's algorithm created partitions that are found to be associated with more attributes of 'BP' database [5]. Additional tables in support of this result are listed in Appendix.

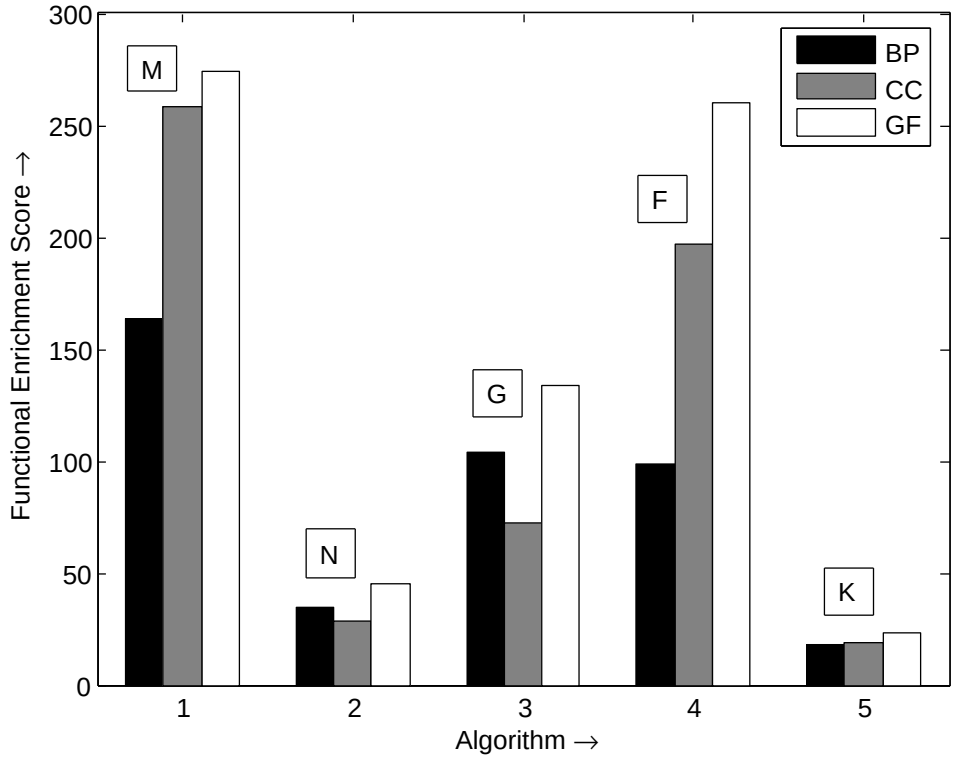


Figure 4.8: Comparison based on functional enrichment score of valid attributes
BP- Biological Process, CC- Cellular Component, GF- GO Full, M- Modularization algorithm, N- Newman's algorithm, G- Greedy algorithm, F- Farhat's algorithm and K- Kernighan-Lin's algorithm.

4.6 Conclusive remarks

In this Chapter, we have done a comparative analysis among five different partitioning algorithms. Three of them follow graph partitioning techniques while the rest two follow community finding and modularization techniques respectively. The partitions generated by these algorithms have been validated by comparing the level of extent to which they could associate to GO terms. A new GO attribute based score (Functional enrichment score) has been designed for validating these modules. The score establishes a validity index among GO attributes. It can be extended for performance mea-

surement of any kind of partitions/clusters/modules created from biological networks with existing ontology knowledge. Superior performance of the Modularization algorithm in comparison with some traditional graph partitioning and community finding algorithms has been reflected in this chapter. Modularization algorithm is a better algorithm to create modules from human Wnt STP than Greedy, Farhat's and Kernighan-Lin's graph partitioning algorithms, and Newman's community finding algorithm. A real life application of the modularization algorithm will be described in Chapter 5 by deriving a phylogenetic tree from the modules of species-specific Wnt STPs.

Chapter 5

Deriving Phylogenetic Trees from Modules

5.1 Introduction

As described earlier, a Signal Transduction Pathway (STP) includes a set of biomolecules. These biomolecules operate in a synchronous manner to create a cascade of reactions, ultimately generating a response to stimuli *in vivo*. Wnt STP is one of them. Wnt molecules are secreted cysteine-rich, lipid-modified glycoproteins. They bind to Frizzled seven-transmembrane-spanning receptors (FZDs) along with co-receptor LRPs (Lipoprotein Receptor related Proteins) and initiate a cascade of reactions, which altogether form the Wnt STP [29].

The pathway is involved in crucial cellular functions, *viz.*, regulation of fate determination, proliferation, differentiation, migration and apoptosis of cells [304, 305]. In matured organisms, Wnts are implicated in maintaining stem cell like fates in the intestinal epithelium [308], skin [309], and hematopoietic cells [310]. A number of studies has been dedicated to analyze the specific details of Wnt STPs in a few model organisms [333–335]. On the contrary, only a few investigations have been initiated to understand how this pathway itself has evolved [336].

Thus the present study deals with evolution of Wnt STPs. Here, we have considered Wnt STPs of 48 species (Table 5.1) ranging from placozoans to humans. The genes/proteins involved in this pathway can be considered as nodes and the interactions among them as edges. A pathway conceived in such a way is open to a wide range of network analysis techniques. Network comparison to uncover biological functions and phylogeny [337] is one of them. Networks derived from biological pathways (gene regulatory, metabolic, signal transduction, and protein-protein interaction networks) can be used for phylogenetic studies.

Networks can be compared by their size [338], similarity of the nucleotide sequences [339], amino acid sequences [340], enzyme sequences [341], and protein structural classification [342]. Some other ways of alignment are based on functional similarity of the enzymes [343], substrate-product relationships [344], proteins [345], and enzyme hierarchy, and gene ontology [346], chemical structures or compound similarity [347]. Presence of common topo-

logical structures, *i.e.*, graphlets or sub-networks [348], and presence or absence of pathways in an entire pathway repertoire [349] can also be utilized for pathway comparison.

A particular type of information used for pathway comparison can be termed as "factor". Similarity or distance scores computed by considering single and/or sets of related factors are widely used for phylogenetic tree construction [343]. In addition to such factors, we add another factor named 'modules' [1,3]. These factors encapsulate different categories of information about the pathways. One or all of the known factors can be ideally considered while creating a tree from a set of biological networks. Creating phylogenetic trees by considering different combination of factors and then selecting the best suitable tree is a common practice [350]. While selecting a combination of factors to create a tree, related factors must be given priority as they bring forward the complete picture of the dataset. Similarities among these factors, for a set of species, can be represented as a distance matrix. Phylogenetic trees can be constructed from such matrices by various existing softwares and tools [351, 352].

In this Chapter, we have created two such phylogenetic trees, *viz.*, the pathway tree and the module tree, by considering pathway topology and modules of Wnt STPs respectively. Three datasets corresponding to 48, 29 and 12 species have been considered; the later two sets being subsets of the former one. These phylogenetic trees represent evolution of the Wnt STP obtained at pathway and module level respectively. These trees have been analyzed separately to find out species arrangement with respect to common taxonomy. Then they have been compared with the NCBI taxonomy tree and 18S rRNA tree for assessment of their quality in representing evolution of Wnt STP. Finally, the best tree has been analyzed and associated with major changes found in Wnt STP evolutionary course. The phylogenetic trees have been created with MEGA version 4.0.2 [352].

5.2 Data

Species-specific Wnt STPs in KEGG/Pathway database [167] has been taken as raw data (Table 5.1). The database uses a unique three letter code for each species along with their biological and common names (wherever applicable), *viz.*, 'hsa' for *H. sapiens* (human). These three letter codes have been used extensively in this chapter.

Table 5.1: List of species and 18S rRNA reference ids

Sl. no.	KEGG code	Binomial nomenclature	Common name	18S rRNA sequence id
01	aag	*A. aegypti*	Yellow fever mosquito	U65375 [G]
02	aga	*A. gambiae*	Mosquito	AM157179 [G]
03	ame	*A. mellifera*	Honey bee	AY703484 [G-P]
04	aml	*A. melanoleuca*	Giant Panda	GL196163 [G]
05	api	*A. pisum*	Pea aphid	U27819 [G]
06	bfo	*B. floridae*	Florida lancelet	M97571 [G]
07	bmy	*B. malayi*	Filaria	AAQA01003643 [G-S]
08	bta	*B. taurus*	Cow	NR_036642 [G]
09	cbr	*C. briggsae*	-	FJ380929 [G]
10	cel	*C. elegans*	Nematode	EU196001 [G-P]
11	cfa	*C. familiaris*	Dog	AAEX02007663 [G-S]
12	cin	*C. intestinalis*	Sea squirt	AB013017 [G-P]
13	cqu	*C. quinquefasciatus*	Southern house mosquito	AAWU01013261 [G-S]
14	dan	*D. ananassae*	-	XR_046314 [G]
15	der	*D. erecta*	-	XR_046906 [G]
16	dgr	*D. grimshawi*	-	[S-E]
17	dme	*D. melanogaster*	Fruit fly	M21017 [G]
18	dmo	*D. mojavensis*	-	XR_047783 [G]
19	dpe	*D. persimilis*	-	XR_046906 [G]
20	dpo	*D. pseudoobscura pseudoobscura*	-	XR_053284 [G]
21	dre	*D. rerio*	Zebrafish	AC139725 [G-S]
22	dse	*D. sechellia*	-	XR_048770 [G]
23	dsi	*D. simulans*	-	AY037174 [G]
24	dvi	*D. virilis*	-	XR_049279 [G]
25	dwi	*D. willistoni*	-	XR_049811 [G]
26	dya	*D. yakuba*	-	XR_050457 [G]
27	ecb	*E. caballus*	Horse	AJ311673 [G-P]
28	gga	*G. gallus*	Chicken	M59389 [G]
29	hmg	*H. magnipapillata*	-	HQ392522 [G-P]
30	hsa	*H. sapiens*	Human	X03205 [G]
31	isc	*I. scapularis*	Black-legged tick	ABJB010180167 [G-S]
32	mcc	*M. mulatta*	Rhesus Monkey	FJ436026 [G-P]
33	mdo	*M. domestica*	Opossum	AJ311676 [G-P]
34	mmu	*M. musculus*	Mouse	X00686 [G]
35	nve	*N. vectensis*	Sea anemone	AF254382 [G]
36	nvi	*N. vitripennis*	Jewel wasp	GQ410677 [G-P]
37	oaa	*O. anatinus*	Platypus	AJ311679 [G-P]
38	phu	*P. humanus corporis*	Human body louse	FJ267399 [G-P]
39	ptr	*P. troglodytes*	Chimpanzee	AADA01268803 [G-S]
40	rno	*R. norvegicus*	Rat	X01117 [G]
41	smm	*S. mansoni*	-	U65657 [G]
42	spu	*S. purpuratus*	Purple sea urchin	L28055 [G]
43	ssc	*S. scrofa*	Pig	AY265350 [G]
44	tad	*T. adhaerens*	-	Z22783 [G]
45	tca	*T. castaneum*	Red flour beetle	HM156711 [G-P]
46	tgu	*T. guttata*	Zebra finch	ABQF01063677 [G-S]
47	xla	*X. laevis*	African clawed frog	X04025 [G]
48	xtr	*X. tropicalis*	Western clawed frog	AAMC01103672 [G-S]

[Notations in 5^{th} column: G: GenBank accession number (Complete sequence), G-P: GenBank accession number (Partial Sequence), G-S: GenBank accession number of predicted 18S rRNA sequence as given in SILVA database [353], S-E: Sequence taken from Stage and Eickbush [354].

5.2.1 18S rRNA sequence data

18S rRNA is a component of small eukaryotic ribosomal subunit (40S). 18S rRNA data is widely used in molecular analysis to reconstruct the evolutionary history and ancient divergences of organisms due to its slow evolutionary rate. Here, most of the 18S rRNA sequences have been taken from GenBank [355] for construction of the 18S rRNA tree. With a simple search dialogue of "— [organism] AND 18S ribosomal RNA [keyword] NOT (partial)", the sequence of interest can be extracted easily. If complete sequences are not available, the "NOT (partial)" dialogue can be omitted and a search for partial sequences can be done.

We have found 28 complete and 11 partial 18S rRNA nucleotide sequences for which GenBank accession numbers are listed in Table 5.1. Eight sequences have been taken from the SILVA comprehensive ribosomal RNA database (http://www.arb-silva.de/). It is an on-line resource for quality checked and aligned ribosomal RNA sequence data, which is free for academic use [353]. It provides regularly updated datasets of aligned small (16S/18S, SSU) and large subunit (23S/28S, LSU) rRNA sequences for all three domains of life (Bacteria, Archaea and Eukarya). We have taken sequences from the SSU r106 database whose respective GenBank accession numbers are given in Table 5.1. 18S rRNA sequence of *D. grimshawi* has been taken from Stage and Eickbush, 2007 [354] as it is not available in GenBank and SILVA.

5.3 Methodology

Here we describe the proposed methodology involved in creating phylogenetic trees from species specific Wnt STPs, taxonomy information from NCBI and 18S rRNA sequences. We have considered two different sets of factors to do system level evolutionary analysis of species-specific Wnt STPs.

A phylogenetic tree can be divided into a number of clades. A clade is a group of organisms that includes an ancestor and all descendants of that ancestor. The whole tree is also a clade as all the taxa are evolutionarily connected (either closely or distantly) among each other descending from a

common ancestor. So a clade must be chosen carefully so that it can convey inter-relatedness among a group of species, if not all the species of the tree. It is expected that species belonging to the same genus will be placed closer to each other than species belonging to different genera in the tree while considering evolution of species in general. Theoretically, the reason being each genus originating from a family or sub-family of living beings under different evolutionary pressure, and may be at different evolutionary time period.

On the other hand, while considering a particular STP's evolution, *e.g.*, Wnt STP of a set of species, deviation from general evolutionary rate can be expected. It is due to some stress factors like predators [356], habitat [357], rapid [358] as well as normal climate change [359,360] and mode of survival/allowance (in terms of tolerance, including parasitic [361, 362] and pathogenic [363, 364] mode of survival) of the species. Signal transduction pathways, being dependent on external stimuli for their action and response, must have been influenced by such factors. Some of these stress factors presumably map into genetic makeup of a species on the long run, and get reflected in the general evolutionary rate. However, some of them do not map into the genetic makeup. Instead, they bring some phenotypic changes, recognized as plastic component [365] or phenotypic plasticity [366]. Biological processes of a species get shaped by genetic makeup as well as phenotypic plasticity. Hence, the evolution of a biological-process is expected to deviate from the general evolutionary rate of species which is mainly based on genetic makeup.

Our aim is to know how much deviated is a phylogenetic tree (constructed from pathway topology/modules) from the standard evolutionary trees (18S rRNA and NCBI taxonomy tree). We have designed this work accordingly to test efficacy of alternative phylogenetic trees (*i.e.*, those based on pathway topology and module dissimilarity) that can later be used to establish Wnt evolution. Throughout this manuscript we have used some common notations while analyzing the phylogenetic trees. The notations are listed alphabetically in Table 5.2.

Table 5.2: List of notations used in Figures 5.2 and 5.3

Notation	Taxonomical rank	Name	Notation	Taxonomical rank	Name
P	Phylum	-	R	Order	Rodentia
A	Phylum	Arthropoda	Ar	Order	Artiodactyla
Cn	Phylum	Cnidaria	Pe	Order	Perissodactyla
N	Phylum	Nematoda	Pr	Order	Primates
Ch	Phylum	Chordata	Aa	Order	Anura
E	Phylum	Echinodermata	Di	Order	Didelmorphia
Pl	Phylum	Platyhelminthes	Cv	Order	Carnivora
Pz	Phylum	Placozoa	Rh	Order	Rhabditida
C	Class	-	F	Family	-
I	Class	Insecta	Dr	Family	Drosophilidae
Ah	Class	Arachnida	Mu	Family	Muridae
Av	Class	Aves	Pi	Family	Pipidae
Am	Class	Amphibia	Cd	Family	Canidae
Ac	Class	Actinopterygii	U	Family	Ursidae
M	Class	Mammalia	Cu	Family	Culicidae
As	Class	Ascidiacea	Ra	Family	Rhabditidae
L	Class	Leptocardii	G	Genus	-
S	Class	Secernentea	Do	Genus	Drosophila
O	Order	-	Ms	Genus	Mus
D	Order	Diptera	Rt	Genus	Rattus
Hy	Order	Hymenoptera	X	Genus	Xenopus
Mo	Order	Monotremata	Ae	Genus	Aedes
Co	Order	Coleoptera	Cl	Genus	Culex
Ph	Order	Phthiraptera	Al	Genus	Anophelinae
He	Order	Hemiptera	Cr	Genus	Caenorhabditis

5.3.1 Generation of the pathway tree

We have generated the pathway tree based on topological distances among species-specific pathways. Topological distance $D_P(x, y)$ between two pathways of species x and y can be defined as

$$D_P(x, y) = 1 - S_P(x, y) \tag{5.1}$$

where $S_P(x, y) \in [0, 1]$ is the topological similarity between the two species-specific pathways. $S_P(x, y)$ has been calculated by the GRAph ALigner algorithm (GRAAL) designed by Kuchaiev et al. 2010 [337] and implemented in the GraphCrunch2 software [367]. The algorithm uses topological information based on graphlets in order to perform network alignment. Given two networks, GRAAL finds an embedding of the smaller network into the larger one such that every node in the smaller network is aligned to exactly one node in the larger one. It aims at exposing as much topological similarity between the networks as possible. It is a seed-and-extend algorithm that greedily aligns nodes based on their signature similarities while traversing both networks simultaneously in a breadth-first manner.

GRAAL Algorithm [337]

Let us consider two graphs $G_1(U, E)$ and $G_2(V, F)$, where U and V are the sets of vertices in G_1 and G_2 respectively, while E and F are the sets of edges in G_1 and G_2 respectively. GRAAL aligns G_1 and G_2 by first computing a cost matrix C for aligning each node u in G_1 with each node v in G_2. The cost of aligning two nodes takes their signature similarity into account. The cost $C(u, v)$ of aligning nodes u and v is computed as

$$C(u,v) = 2 - ((1-\alpha) \times \frac{deg(u) + deg(v)}{max\_degG_1 + max\_degG_2} + \alpha \times S(u,v)) \quad (5.2)$$

where $\alpha \in [0,1]$ is the parameter that controls the contribution of node degrees to the cost function. As established earlier [337], $\alpha = 0.8$ yields maximum similarity scores. Hence, we have taken the default value of $\alpha = 0.8$ while computing similarities among species-specific Wnt STPs by the GraphCrunch2 software [367]. The terms deg(u) and deg(v) are degrees of nodes $u \in G_1$ and $v \in G_2$ respectively, while max_deg(G_1) and max_deg(G_2) are the maximum degrees of the nodes in G_1 and G_2 respectively. $S(u, v)$ is the signature similarity score of nodes u and v. Signature similarity of two nodes is their graphlet degree vector similarity [368].

In the cost matrix C, a cost of 0 corresponds to a pair of topologically identical nodes v and u, while a cost close to 2 corresponds to a pair of topologically different nodes. Once the cost matrix is created, GRAAL searches for an initial seed (u,v), (u,v) representing an edge between u and v, which has the lowest alignment cost in the matrix C. After seed formation, GRAAL builds spheres of all possible radii around nodes u and v in graphs G_1 and G_2, respectively. A sphere of radius r around node u in a network G is a set of nodes which are exactly at the distance r from node u, *i. e.*, $S_{G_1}(u,r) = \{w \in G$: $d(u, w) = r\}$, $d(u, w)$ being the distance between u and w. Spheres of the same radius in two networks are then greedily aligned together by searching for the pairs (u',v') such that $u' \in S_{G_1}(u,r)$ and $v' \in S_{G_2}(v,r)$, that are not already aligned but can be aligned with the minimal cost. After this process,

it is possible for some nodes in both the networks to remain unaligned as spheres of the same radius in different networks can have different number of nodes. GRAAL repeats itself by searching for a new seed and aligning the remaining components of two networks. It stops when each node from G_1 is aligned to exactly one node in G_2.

Using GRAAL, one can estimate Edge Correctness (EC), Node Correctness (NC) and Interaction Correctness (IC) while aligning two networks. Edge correctness is the percentage of edges in the first graph that are aligned to edges in the second graph. High edge correctness means that networks G_1 and G_2 share similar topologies. Node correctness is the percentage of nodes in network G_1 that are correctly aligned to nodes in G_2. Interaction Correctness is the percentage of interactions that are aligned correctly. An interaction u-w is correctly aligned if two connected nodes u and w from G_1 are correctly aligned to their partners in G_2, and if their partners interact in G_2. For NC and IC, the users need to know the correct node mappings, which is not always possible in real life cases. Thus EC measure is deemed the best way to estimate topological similarity between two networks. Edge correctness of an alignment g: $U \rightarrow V$ produced by the algorithm GRAAL can be defined as

$$EC = \frac{|\{(u,w) \in E : (g(u), g(w)) \in F\}|}{E} \times 100\% \tag{5.3}$$

where $g(u)$ and $g(w)$ are the partners of $u$and w respectively in F. We have calculated EC values among species-specific Wnt STPs. This is nothing but topological similarities $S_P(x, y)$s among these pathways as described in Equation 5.1. Their corresponding distances $D_P(x, y)$s have been used to create a distance matrix D_P, which in turn has been used to create a phylogenetic tree named as the pathway tree.

5.3.2 Generation of the module tree

The module tree has been generated solely based on one-to-one mapping of members present in modules of different species-specific Wnt STPs. A mod-

ule can be defined as a sub-pathway that tends to be self-sufficient by maintaining minimal dependency on the remaining part of the pathway. Modules have been created by the 'modularization algorithm' already developed in Chapter 3.

Table 5.3: Pathway size and number of modules of species-specific Wnt STPs

Serial Number	KEGG code	Binomial nomenclature	Common name	Pathway size	Number of modules
01	aag	*A. aegypti*	Yellow fever mosquito	34	6
02	aga	*A. gambiae*	Mosquito	31	6
03	ame	*A. mellifera*	Honey bee	38	7
04	aml	*A. melanoleuca*	Giant Panda	59	8
05	api	*A. pisum*	Pea aphid	32	6
06	bfo	*B. floridae*	Florida lancelet	45	7
07	bmy	*B. malayi*	Filaria	34	7
08	bta	*B. taurus*	Cow	59	8
09	cbr	*C. briggsae*	-	23	3
10	cel	*C. elegans*	Nematode	23	3
11	cfa	*C. familiaris*	Dog	58	8
12	cin	*C. intestinalis*	Sea squirt	42	7
13	cqu	*C. quinquefasciatus*	Southern house mosquito	35	7
14	dan	*D. ananassae*	-	37	7
15	der	*D. erecta*	-	37	7
16	dgr	*D. grimshawi*	-	37	7
17	dme	*D. melanogaster*	Fruit fly	37	7
18	dmo	*D. mojavensis*	-	37	7
19	dpe	*D. persimilis*	-	32	5
20	dpo	*D. pseudoobscura pseudoobscura*	-	32	6
21	dre	*D. rerio*	Zebrafish	59	8
22	dse	*D. sechellia*	-	37	7
23	dsi	*D. simulans*	-	22	5
24	dvi	*D. virilis*	-	38	7
25	dwi	*D. willistoni*	-	37	7
26	dya	*D. yakuba*	-	36	6
27	ecb	*E. caballus*	Horse	56	8
28	gga	*G. gallus*	Chicken	54	8
29	hmg	*H. magnipapillata*	-	31	6
30	hsa	*H. sapiens*	Human	60	8
31	isc	*I. scapularis*	Black-legged tick	30	6
32	mcc	*M. mulatta*	Rhesus Monkey	59	8
33	mdo	*M. domestica*	Opossum	55	7
34	mmu	*M. musculus*	Mouse	60	8
35	nve	*N. vectensis*	Sea anemone	33	6
36	nvi	*N. vitripennis*	Jewel wasp	38	7
37	oaa	*O. anatinus*	Platypus	47	7
38	phu	*P. humanus corporis*	Human body louse	37	7
39	ptr	*P. troglodytes*	Chimpanzee	58	8
40	rno	*R. norvegicus*	Rat	60	8
41	smm	*S. mansoni*	-	25	5
42	spu	*S. purpuratus*	Purple sea urchin	41	6
43	ssc	*S. scrofa*	Pig	21	4
44	tad	*T. adhaerens*	-	23	6
45	tca	*T. castaneum*	Red flour beetle	37	7
46	tgu	*T. guttata*	Zebra finch	47	7
47	xla	*X. laevis*	African clawed frog	51	8
48	xtr	*X. tropicalis*	Western clawed frog	57	8

Size of a pathway is given in terms of total number of members present in it.

Number of modules found in a pathway depends on its size and complexity. Hence, varying number of modules can be found for different species-specific Wnt STPs (Table 5.3). Presence, absence or modification (increase or decrease due to addition or deletion of nodes) found in modules of a pair of species-specific pathways represent distance between them. This approach

has been inspired by the Number of Common Enzymes (NCE) method described by Heymans and Singh, 2003 [343]. NCE method detects number of common enzymes between two pathways and tries to guess similarity based on that number. Here, rather than considering number of common enzymes, we have calculated number of common members present in the corresponding module of two species-specific Wnt STPs. The score has then been normalized by dividing it with the total number of non-redundant members present in both the species-specific modules. Hence, if the pathway of species x has m_1 modules and that of species y has m_2 modules, we have obtained a similarity matrix of order $m_1 \times m_2$. Each element of the matrix represents similarity score between two different modules belonging to two different species (Equation 5.5). Let M_1 be a module of species x, *i.e.*, M_1 is the set of all the nodes in the module. Similarly, in a species y, M_2 is a set of nodes that constitute a module. The score of similarity Sim(M_1,M_2) between module M_1 of species x and module M_2 of species y is defined as

$$Sim(M_1, M_2) = |M_1 \cap M_2|/|M_1 \cup M_2| \tag{5.4}$$

Now the similarity $S_M(x, y)$ between species x and y is defined as

$$S_M(x, y) = \sum_{i=1,j=1}^{m_1,m_2} Sim(M_i, M_j)/(m_1 \times m_2) \tag{5.5}$$

Distance score $D_M(x, y)$ is defined as

$$D_M(x, y) = 1 - S_M(x, y) \tag{5.6}$$

These distance scores found among 48 different species have been utilized for creation of the module tree. Our purpose in creating such a tree is to test its novelty in presenting the pathway's evolution [7].

5.3.3 Generation of the reference trees

Quality assessment of alternative phylogenetic trees, *i.e.*, the pathway tree and the module tree, has been done by comparing them with existing standards such as trees based on NCBI's classification and 18S rRNA sequences. Two reference trees, namely, the "NCBI Taxonomy tree" (Figure 5.1(a)) and the "18S rRNA tree" (Figure 5.1(b)), have been created for qualitative analysis of phylogenetic trees derived from Wnt STP data.

Generation of the NCBI taxonomy tree

The NCBI taxonomy tree, as shown in Figure 5.1(a), has been created with the help of NCBI taxonomy database [369]. Newick format of the tree has been saved as a text tree after adding organism names in the "Taxonomy Common tree" page[1]. Then it has been viewed with MEGA version 4.0.2 [352].

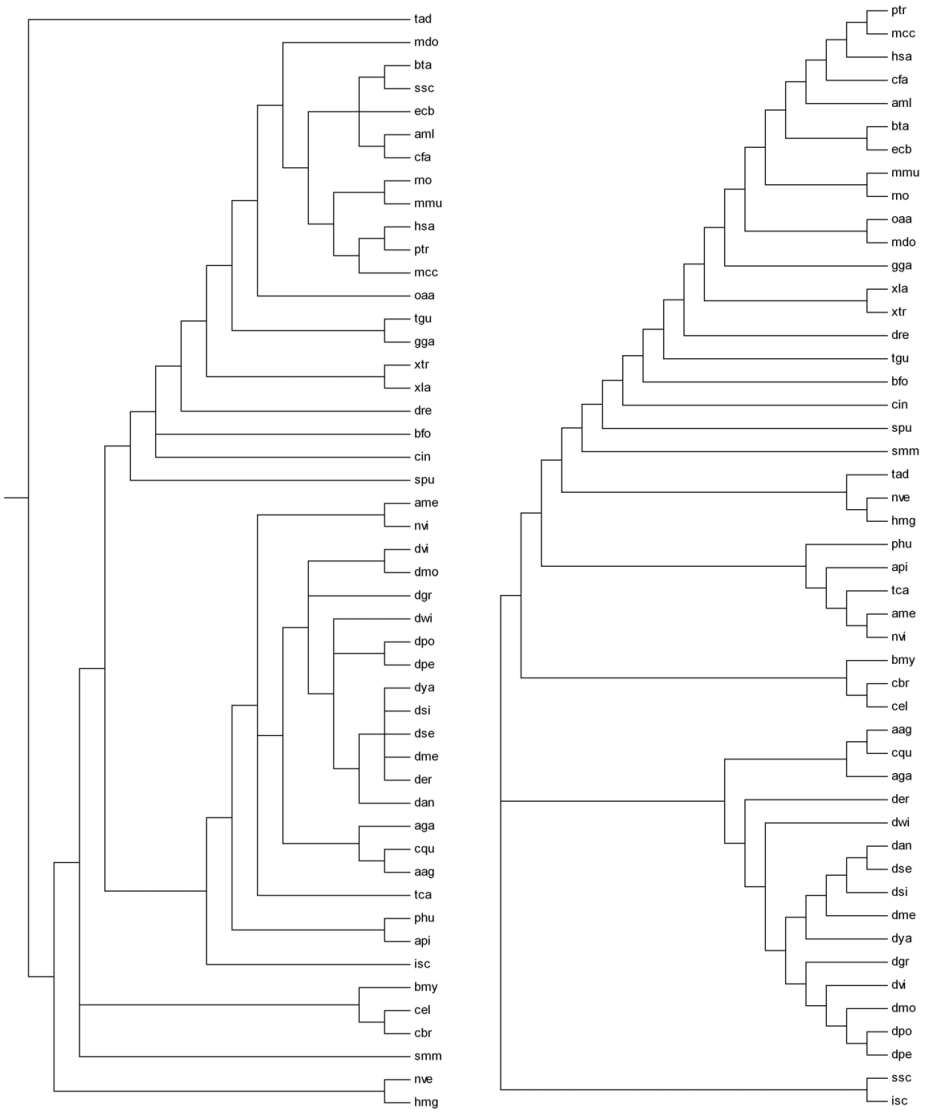
Generation of the 18S rRNA tree

Standard 18S rRNA sequences (Table 5.1) have been used to create the 18S rRNA tree (Figure 5.1(b)). The evolutionary history has been inferred using the Neighbor-Joining method [370]. The optimal tree with the sum of branch length = 2.18407422 has been considered. The evolutionary distances have been computed in the units of the number of base substitutions per site using the Maximum Composite Likelihood method [371]. All positions containing gaps and missing data are eliminated only in pairwise sequence comparisons (Pairwise deletion option). There is a total of 2687 positions in the final dataset. Phylogenetic analyses have been conducted using MEGA version 4.0.2 [352].

5.3.4 Comparison of alternative phylogenetic trees

Nye et al. [372] have developed an algorithm that pairs up each branch in one phylogenetic tree with a matching branch in the second one, and finds the op-

[1]http://www.ncbi.nlm.nih.gov/Taxonomy/CommonTree/wwwcmt.cgi



(a) The NCBI taxonomy tree (b) The 18S rRNA tree

Figure 5.1: The reference phylogenetic trees

timum one-to-one map between branches in the two trees in terms of a topological score. They have developed an applet in Java (http://www.mas.ncl.ac.uk/∼ntmwn/phylo_comparison/pairwise.html) that enables one to explore

the corresponding mapping between the phylogenetic trees interactively, and clearly highlights similar/different parts of the trees, both in terms of topology and branch length. Here, we have considered topology mainly.

Let us now describe the algorithm that compares two phylogenetic trees created from the same set of species. Given two phylogenetic trees T_1 and T_2 that share the same set of leaves L, the algorithm first assigns a score $s(i, j)$ to every pair of edges (i, j) with $i \in T_1$ and $j \in T_2$. Then the algorithm pairs up branches in the two trees to optimize the overall score. This is equivalent to finding a bijection (*i.e.*, a one-to-one and onto correspondence) $f : T_1 \rightarrow T_2$ between the branches of the trees that maximizes the quantity $\sum_{i \in T_1} s(i, f(i))$.

5.4 Results and Discussion

In this section, we describe ranks and positional significance of the species in the pathway tree (Figure 5.2) and the module tree (Figure 5.3). It is to be mentioned here that the species with similar taxonomic ranks coming under the same clade have been marked by continuous rectangles. The species coming under different clades despite having close taxonomical relation have been marked by dotted rectangles. While discussing positional significance of species, we have furnished the similarities in terms of taxonomic ranks (phylum, class and others) to the lowest possible taxonomic rank, which the species under consideration resemble with each other.

5.4.1 The pathway tree

We have compared Wnt STPs of 48 different species, as provided in Table 5.1, for creating a pathway tree (Figure 5.2). The tree has been inferred using the Neighbor-Joining method [370]. These species belong to seven different phyla, most of which (21) belong to the phylum Arthropoda followed by Chordata (19), Nematoda (3), Cnidaria (2), and single species from phyla Echinodermata, Placozoa and Platyhelminthes. As expected, some species have been placed closely in the tree following their taxonomic ranks. A clade

has been formed by dpe, dvi, dan, dmo, dme, dpo and dgr (genus Drosophila). A couple of species belonging to family Culicidae (cqu and aga) and genus Drosophila (dwi and der) of phylum Arthropoda have been placed closely in the tree. The species mmu and bta (phylum Chordata, class Mammalia) have also been placed closely in the tree. The species mcc and mdo (class Mammalia) have formed a clade and so have ptr and ssc. Similarly, the species ecb, tgu and xtr belonging to phylum Chordata have formed a clade. The species cbr and cel (phylum Nematoda, genus Caenorhabditis) have also been included in a clade [7].

On the other hand, we have also found some deviations. The species hmg (phylum Cnidaria) and nvi (phylum Arthropoda) have formed a clade. The species spu (phylum Echinodermata) has formed a clade with oaa (phylum Chordata, class Mammalia) and xla (phylum Chordata, class Amphibia). The species dya (phylum Arthropoda) has formed a clade with a species of phylum Platyhelminthes (smm) rather than with aag (phylum Arthropoda). Together, they have formed another clade with nve (phylum Cnidaria). Possibly these deviations originate due to pressure of phenotypic plasticity on Wnt STP.

5.4.2 The module tree

The module tree (Figure 5.3) has been inferred using the Neighbor-Joining method [370]. Phylogenetic analyses have been conducted by MEGA version 4.0.2 [352]. As before, the species belonging to the same phylum, class, order and family coming under the same clade have been marked by continuous rectangles. The species coming under different clades despite having close taxonomical relation have been marked by dotted rectangles.

The species dpe, dpo, dme and dya have formed a clade (genus Drosophila). Although, aga (family Culicidae) and dvi (family Drosophilidae) have proximity in the tree, they have come under different clades as they belong to two different families. The species dmo and dpe have formed a clade (genus Drosophila), and so have dse and dwi (genus Drosophila), bta and ecb (class Mammalia), cfa, mcc, aml, mmu, ptr and rno (class Mammalia). The species

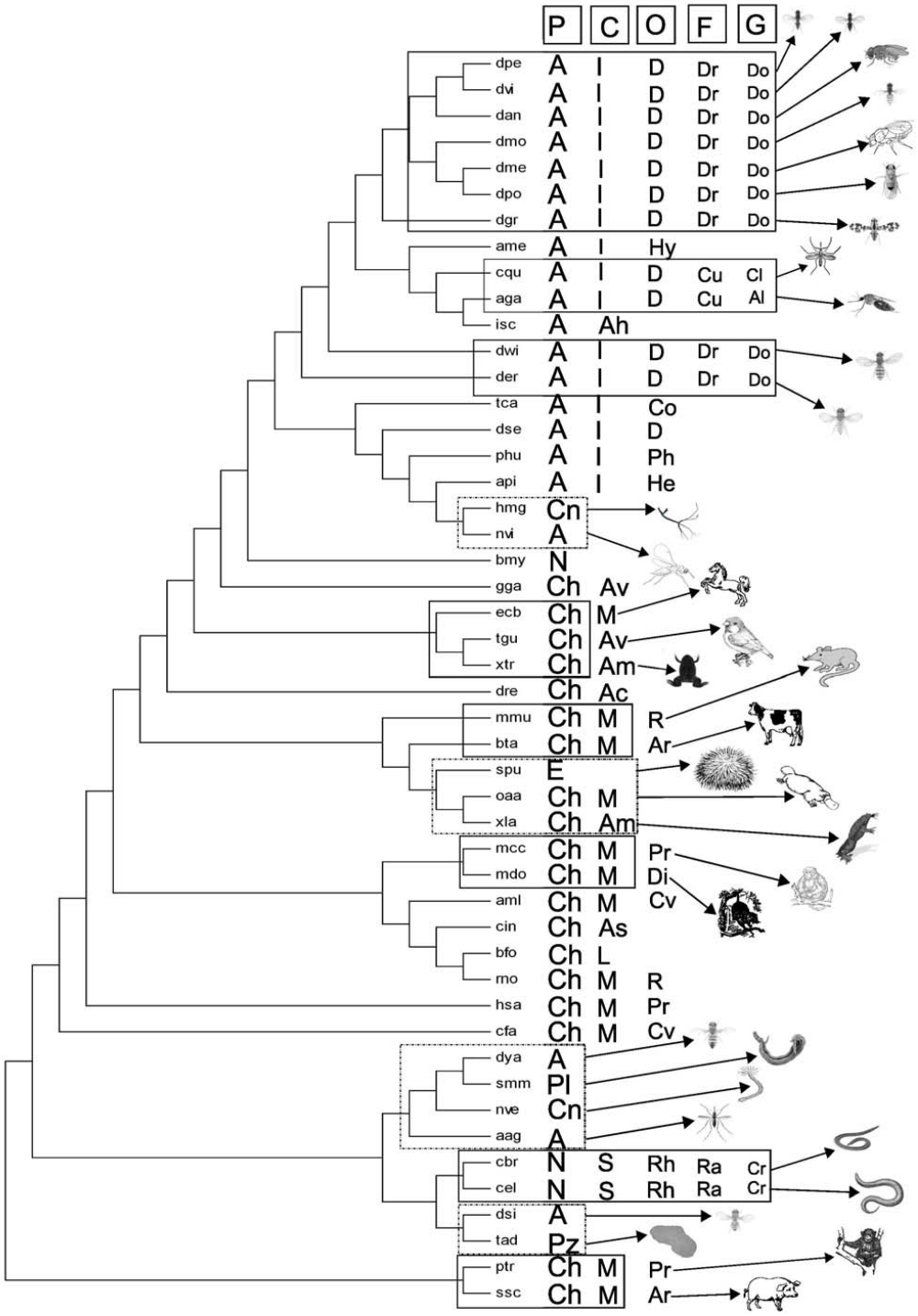


Figure 5.2: The pathway tree constructed from 48 species

The optimal tree with the sum of branch length = 1.88656097 has been considered. Notations used in the figure are listed in Table 5.2.

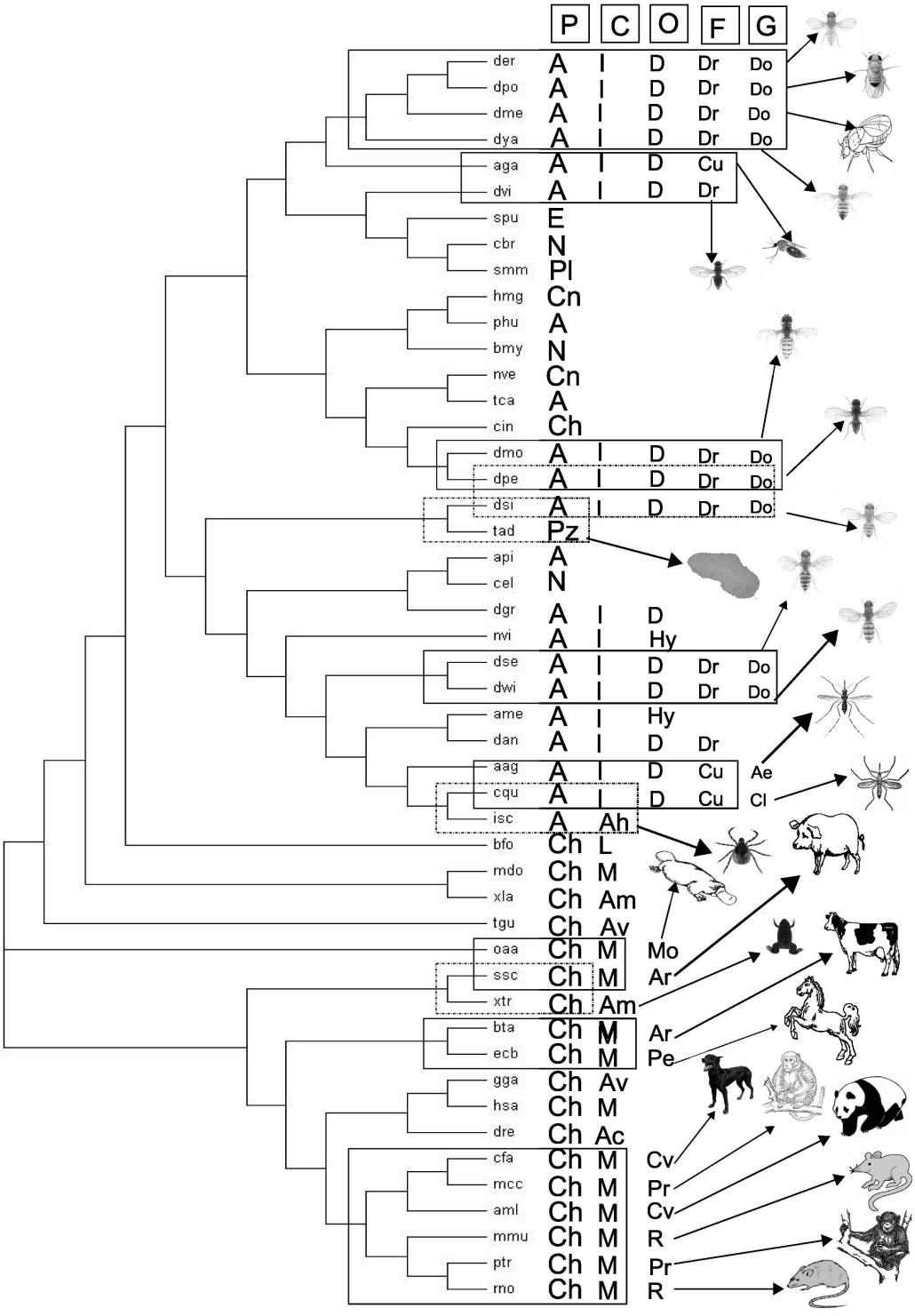


Figure 5.3: The module tree constructed from 48 species

The optimal tree with the sum of branch length = 21.29711960 has been considered. Notations used in the figure are listed in Table 5.2.

aag and cqu have been placed closely in the tree (family Culicidae), and so as oaa and ssc (class Mammalia).

On the other hand, the species dpe and dsi have not formed a clade despite belonging to the same genus (genus Drosophila). However, dsi (phylum Arthropoda) have formed a clade with tad (phylum Placozoa). Another clade has been formed by cqu (class Insecta) and isc (class Arachnida) rather than with aag (class Insecta). Similarly, ssc (class Mammalia) have formed a clade with xtr (class Amphibia) rather than with oaa (class Mammalia) [7].

5.4.3 Finding a better tree

Both the pathway and module trees have shown preservation to contemporary notions regarding evolution of species. Some deviations also have been found in these trees. In order to put a universal measure to their quality, and to determine a better set of factor(s) representing the evolution of Wnt STP, we have followed the concept of alternative phylogenetic tree comparison suggested by [372]. A brief description of this method is furnished in Section 5.3.4.

We have compared the pathway and module trees with the NCBI taxonomy tree and the 18S rRNA tree. The module tree has showed 42.4% topological similarity with the NCBI taxonomy tree and 42.2% similarity with the 18S rRNA tree followed by the pathway tree (38.2% and 37.1% respectively) as given in Table 5.4. Hence, the module tree has outperformed the pathway tree in representing Wnt STP evolution. The module tree has also outperformed the pathway tree, when we have repeated the same protocol with a dataset of 29 species (Table 5.4), for which complete 18S rRNA sequences are available as given in Table 5.1.

Some of the pathways in our dataset comprising 48 species-specific pathways are partially known. In order to strengthen our results, we have considered Wnt STP of a smaller and more complete pathway dataset of 12 species (aml, bta, cfa, dre, ecb, hsa, mcc, mdo, mmu, ptr, rno and xtr). These pathways have varying number (55-60) of nodes. For this dataset too, the module tree has showed maximum similarity with the NCBI taxonomy tree

Table 5.4: Similarity of the pathway tree and the module tree with NCBI taxonomy tree and 18S rRNA tree for 48, 29, and 12 species

	For 48 species		For 29 species		For 12 species	
	NCBI tree	18S rRNA tree	NCBI tree	18S rRNA tree	NCBI tree	18S rRNA tree
Pathway tree	38.2	37.1	38.2	36.2	36.7	37.3
Module tree	42.4	42.2	48.9	39	45.4	55.4

Similarities are given in percentage.

(45.4%) and the 18S rRNA tree (55.4%) as given in Table 5.4. These findings conclude that modules are better factor in creating phylogenetic trees that bear maximum resemblance with reference trees than pathway topology. Then the module tree has been analyzed for clades that represent evolution of Wnts. But, before that a sneak peek into evolution of Wnt STP among different phyla is necessary.

5.4.4 Wnt evolution and the module tree

Wnt genes and the associated pathway (Wnt STP) show diverse characteristics starting from Placozoa (*T. adherens*) to Chordata (*H. sapiens*). A complete component of Wnt STP is present in Trichoplax, irrespective of its simple body plan that presumably takes part in other functions [373–375]. The Wnt diversity continues to cnidarians which have a defined body-plan indicating use of the Wnt STP. They possess 14 Wnt orthologs belonging to 12 sub-families. An additional WntA is present, which does not have any human counter-part [376–378]. This diversity of Wnt genes is lost in flatworms. They possess only five sub-families of Wnts, but both the canonical and non-canonical Wnt STPs are found to be functional. Wnt6-Wnt10, Wnt16 and WntA genes are lost. In addition, the Wnt4 gene is lost with rise of parasitism [379], which is probably due to easy access to the genetic machinery of the host organism.

Nematodes have only 5 Wnt ligands and have more super-specialization in the form of 3 distinct β-catenin genes with distinct functionality [380, 381].

Arthropods are characterized by the loss of Wnt16 [379, 382]. The beetle *T. castaneum* (super-phylum Ecdysozoa) retains only 9 Wnt subfamilies, with no duplications [383], while *D. melanogaster* (fruitfly) has just 7 Wnt genes [384]. Echinoderms retain 11 sub-families of Wnt genes with a WntA ortholog, indicating their connection with protostomes. Wnt2 and Wnt11 genes are absent. While absence of Wnt11 is quite common in other metazoans, absence of Wnt2 is an exception [376]. Chordates are characterized by their complexity in Wnt signaling, presence of multiple Wnt antagonists and loss of WntA gene. Humans have 19 Wnt genes, representing 12 subfamilies with 7 duplications [384, 385].

These observations indicate a possible gene duplication event in MRCA (Most Recent Common Ancestor) ∼940 Mya (Million years ago) that continued to placozoans in a subdued manner, then lost in Platyhelminthes and Nemathelminthes. The loss is minimized in Arthropods possibly due to a gene-boom during or before divergence of Echinoderms ∼500 Mya. Echinoderms retain a mixture of old and new Wnt characteristics that flourish extensively in Chordates with loss of early Protostome characteristics [336, 386]. Presumably amidst multiple gene duplication events, the Wnt genes pass through a wormhole like phase; wormhole being a hypothetical topological feature of space time as given in Figure 5.4. Although these gene duplication events do not correlate with the origin of the principal animal groups, they can be related to evolutionary course of Wnt gene family.

When this trend of Wnt pathway evolution has been compared with our constructed module tree (Figure 5.3), we have found some similarities between them. Most of the chordates have proximity with each other. Arthropods have formed two distinct clusters. Placing of the only Placozoan (tad), Platyheminth (smm) and Echinoderm (spu) cannot be commented upon due to their singular presence in the dataset.However, placing of the single Platyheminth (smm) with a Nematode (cbr) has been still justifiable from the fact that species from both the phyla tolerate Wnt diversity loss.

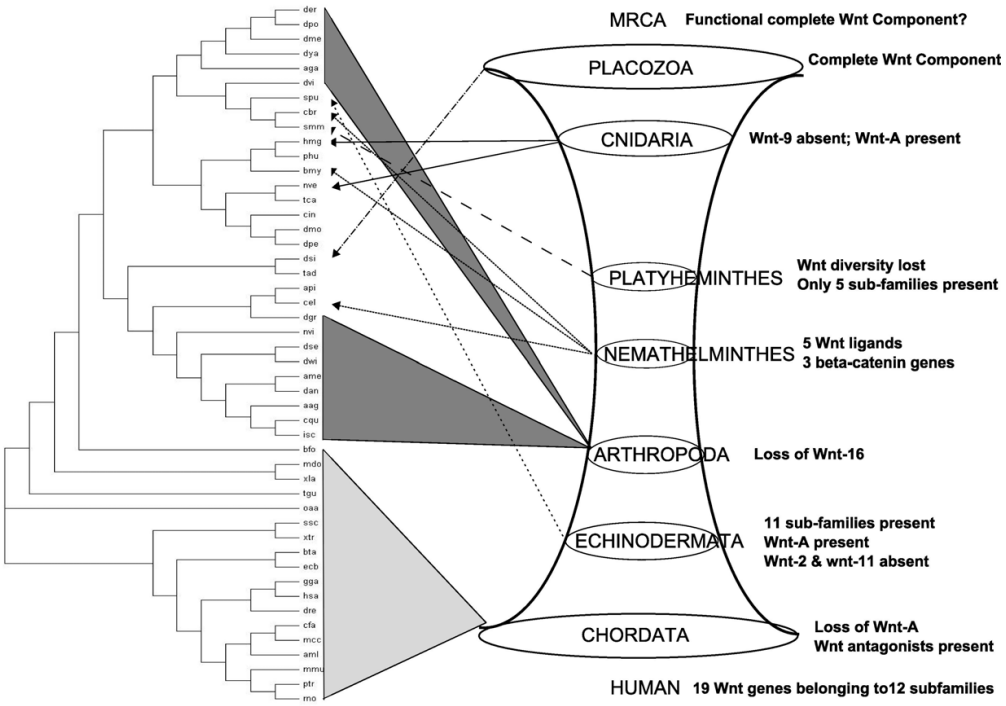


Figure 5.4: Relational aspects between module tree and Wnt STP evolution
Interestingly, Wnt genes pass through a wormhole like phase in time during evolution.

5.5 Conclusive remarks

This chapter emphasized on deriving a way of representing evolution of Wnt STP over various species. Here, we have created two alternate phylogenetic trees, *viz.*, the pathway tree based on pathway topology and the module tree based on modules derived from three datasets of species-specific pathways (comprising 48, 29 and 12 species), and compared them with widely used reference phylogenetic trees, *i.e.*, the NCBI taxonomy tree and the 18S rRNA tree. The module tree has been found to possess maximum resemblance with both the reference trees for all the three datasets. Hence, module comparison serves better for deriving phylogenetic trees from human Wnt STP. Considering modules alone or along with other factors may prove to be beneficial for deriving phylogeny from biochemical pathways.

However, the module tree has showed some species-arrangements which defy the general notion of taxonomy and evolution. This may turn out to be the pressure of phenotypic plasticity that Wnt STP of these phyla faced individually. Moreover, according to one of the standard trees, *i.e.*, the NCBI taxonomy tree, it cannot replace a carefully constructed phylogenetic analysis of molecular or morphological characters. Thus whatever error is there in the global NCBI taxonomy tree, it will reflect in our custom-made NCBI taxonomy tree.

For 48 species, the module tree has formed distinct clades for the genus Drosophila (der, dya, dme, dsi, dan, dwi, dmo, dvi and dgr), class Mammalia (oaa, mdo, mmu, rno, mcc, bta and hsa), family Muridae (mmu and rno), class Insecta ((aga and api) and (aag and ame)), and genus Caenorhabditis (cbr and cel). The module tree has been mapped into major events of Wnt gene evolution, for phyla that have multiple species (Chordates and Arthropods).

Chapter 6

A Wnt Diseasome: Modules in the Diseasome

6.1 Introduction

In this chapter, we have created a manually curated diseasome related to genes of human Wnt STP. The genes are found to be associated with a plethora of diseases including cancers. A diseasome is a combined set of all known disorders or gene associations in a species. It is created by linking the complete set of genetic disorders with the complete list of disease genes [250]. A list of disease genes is not always enough to understand mechanisms of various human diseases [250, 387] for the reason that diseases may not be independent of one another. More connected a disease is to other diseases, the higher is its prevalence and associated mortality rate [251]. For example, certain diseases like Diabetes, Obesity, Gaucher disease and Parkinson disease often co-occur in the same individual. Diseasome-wise studies are needed to understand such situations. In a human disease network, two disorders are linked with each other, if they share at least a common disease gene. On the other hand in a network of genes associated with diseases, there is an edge between two genes, if both of them are associated with atleast a single common disorder. Disease maps are potential knowledge bases that throw light on multiple disease related complicacies. For appropriate disease-specific diagnostic, prognostic and therapeutic approaches, gene-disease association studies provide valuable information [388].

Information on interactions among genes can be used to find disease-related genes. The underlying assumption is that if 'two genes work together, the known association of one with a disease suggests that the other may also be associated with the same disease' [388]. Methods such as ENDEAVOUR [389] and G2D [390] use this assumption. We have used the same assumption but from a different point of view via biomedical literature mining. A disease network is built from the genes involved in the human Wnt signaling pathway to throw light on the Wnt signaling pathway's role in various diseases and their comorbidity.

Biomedical literature is too massive a source to be handled in a general sense. Generalized databases including GEPIS [391], KEGG [167], OMIM

[392], PubMed[1] , STRING [241], TiGER [393], TRANSFAC [394], UniHi [395] constitute parts of this vast resource. There are also purpose specific databases cum tool repositories like Oncomine 3.0 [396], DiseaseMeth [397], DistiLD [398], the Disease Ontology (DO) database [399] and web resources like Network of Cancer Genes [NCG 3.0] [400] among others. We have mined some of these sources for our purpose.

6.2 Methodology

There is a vast amount of existing literature that supports the role of the human Wnt STP in various diseases; the most important being cancers. Some of the articles have been found to implicate the whole pathway for certain diseases, while the others pinpoint certain gene(s) of the pathway and their role in human pathogenesis. We have scanned literature of both kinds to find gene-disease associations for the human Wnt STP. Wnt STP information is taken from KEGG/Pathway database [167]. The literature scanning has been done using the links and references provided in KEGG/Pathway database. The pathway entry (entry id hsa:04310) contains multiple references. The initial referrals for finding associations are provided by these links and references. Especially, important information in one go, has been found from Table 1 provided by Hatsell et al. [401]. Each gene of the human Wnt STP in KEGG/Pathway database is linked to a flat file that contains information about the entry id, gene name, definition, orthology, pathway, class, motif, links to other databases, position in the chromosome, amino acid sequence and nucleotide sequence. The "links to other databases" section is found to contain links of the Wnt STP component related files present in the other databases, namely, Protein database[2] (NCBI), Entrez gene [402], OMIM [403], HGNC [404], HPRD [405], Ensembl [406] and UniProt [407].

- The 'NCBI-GI' link retrieves a flat file from the 'Protein' database of NCBI. The Protein database is a collection of sequences obtained

[1]http://www.ncbi.nlm.nih.gov/sites/entrez?db=pubmed
[2]http://www.ncbi.nlm.nih.gov/protein/

from several sources, including translations from annotated coding regions stored in GenBank [408], RefSeq [409] and TPA[3] [410], as well as records from SwissProt [411], PIR[4], PRF[5] and PDB [412]. The retrieved flat file contains multiple 'REFERENCE' sections dedicated to a query gene. The 'COMMENT' section gives the gist of experimental or computational findings explained in the papers given in the 'REFERENCE' sections. Both these sections have helped us in scanning literature to establish human Wnt gene-disease associations.

- The 'NCBI-GeneID' link takes into Entrez Gene[6] database of NCBI. Entrez Gene is a searchable database of genes, obtained from RefSeq genomes. The file has a 'SUMMARY' section that describes gene function, protein behavior and differential expression of the query gene in different kinds of tissues. The 'Bibliography' section provides us the required citations for the 'SUMMARY' section. Both these sections have helped in our task of finding gene-disease associations.

- 'OMIM' hyperlink takes a flat file into account in response to a query gene present in OMIM (Online Mendelian Inheritance in Man) database. OMIM is a comprehensive, authoritative, and timely compendium of human genes and genetic phenotypes. The files present in OMIM[7] are descriptive in nature with no regular sections. Hence, whole files have to be scanned to detect the gene-disease associations.

- The link 'HGNC' takes a file of the HUGO Gene Nomenclature Committee[8] into account. This file is not helpful in establishing gene-disease associations as it has no summary section.

- 'HPRD' link considers a file of the Human Genome Reference Database[9]. All the information in HPRD has been manually extracted from the lit-

[3]Third Party Annotation database: http://www.ncbi.nlm.nih.gov/genbank/tpa/
[4]http://pir.georgetown.edu/
[5]Protein Reference Database: http://www.prf.or.jp/aboutdb-e.html
[6]http://www.ncbi.nlm.nih.gov/sites/gene
[7]http://www.ncbi.nlm.nih.gov/omim
[8]http://www.genenames.org/index.html
[9]http://www.hprd.org/

erature by expert biologists who read, interpret and analyze the published data. A typical HPRD file has a 'Disease' tab that contains an OMIM id for describing disease relevance of the query gene.

- The 'Ensembl' id takes the user into a file in the Ensembl[10] database. The Ensembl project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online. This page has many useful sections of information on the human Wnt STP genes like gene summary, splice variants, comparative genomics, genetic variations but not gene-disease association information.

- 'UniProt' link takes to a structured file in UniProt[11] database. A standard UniProt file has a 'General annotation (Comments)' section. This section is arranged in multiple subsections. The 'Involvement in disease' subsection has helped us in collecting the gene-diseases association of human Wnt STP components. This subsection contains cited referrals for pathogenicity of a gene, the references being listed at the end of the file.

In summary, the NCBI-GI, NCBI-GeneID, OMIM, UniProt hyperlinks are proved to be useful for our purpose. A list of all the human Wnt STP genes as well as their hyperlinks to the other databases, from which we have extracted the gene-disease associations, is furnished in Table 6.1. We have scanned all these flat and structured files for gene-disease associations of human Wnt STP in the first phase of our literature scanning. In the second phase, we have independently and individually searched via popular search engines for any kind of association of these genes with human pathogenicity available in research articles. From these gene-disease associations, disease-disease associations are extracted with a simple thumb rule: "If Disease 1 is associated with Gene A (supported by article X) and Gene A is associated with Disease 2 (supported by article Y) then Disease 1 is associated with

[10]http://www.ensembl.org/index.html
[11]http://www.uniprot.org/

Disease 2". These associations are then used to build up a Wnt specific disease network and its inherent properties have been studied [8].

Table 6.1: Wnt STP genes and associated links

Serial no.	Human Wnt STP component	KEGG entry id	NCBI-GI	NCBI Gene id	OMIM id	UniProt id
1	LEF1	51176	195222732	51176	153245	B4DG38, Q659G9, Q9UJU2
2	SMAD4/MADH4	4089	4885457	4089	600993	Q13485
3	NLK	51701	149408126	51701	609476	Q9UBE8
4	SOX17	64321	11967991	64321	610928	Q9H6I2
5	CTBP1	1487	61743967	1487	602618	Q13363, Q7Z2Q5
6	CREBBP	1387	119943102	1387	600140	Q4LE28, Q92793
7	RUVBL1	8607	4506753	8607	603449	Q9Y265
8	MYC	4609	71774083	4609	190080	P01106
9	JUN	3725	4758616	3725	165160	P05412
10	FOSL1	8061	4885243	8061	136515	P15407
11	CCND1	595	16950655	595	168461	P24385, Q6FI00
12	PPARD	5467	284807155	5467	600409	Q03181
13	MMP7	4316	4505219	4316	178990	P09237
14	MAP3K7	6885	4507361	6885	602614	O43318
15	CTNNB1	1499	148233338	1499	116806	P35222
16	PSEN1	5663	4506163	5663	104311	P49768
17	CTNNBIP1	56998	9889555	56998	607758	Q5T4V2, Q9NSA3
18	CHD8	57680	282165704	57680	610528	Q9HCK8
19	PRKACA	5566	4506055	5566	601639	P17612
20	CSNK1A1L	122011	269846834	122011	-	Q8N752
21	FBXW11	23291	48928050	23291	605651	Q9UKB1
22	TBL1X	6907	213021186	6907	300196	O60907
23	DVL1	1855	32479521	1855	601365	O14640
24	CXXC4	80319	13376816	80319	611645	Q9H2H0
25	SENP2	59343	54607091	59343	608261	Q9HC62
26	CSNK2A1	1457	4503095	1457	115440	P68400
27	FRAT1	10023	31317236	10023	602503	Q92837
28	APC2	10297	5031587	10297	612034	O95996
29	NKD1	85407	14916433	85407	607851	Q969G9
30	WNT16	51384	17402914	51384	606267	Q9UBV4
31	PORCN	64840	45439329	64840	300651	Q9H237
32	SFRP1	6422	56117838	6422	604156	Q8N474
33	CER1	9350	4885135	9350	603777	O95813
34	WIF1	11197	111125011	11197	605186	Q9Y5W5
35	LRP6	4040	148727288	4040	603507	O75581
36	DKK1	22943	7110719	22943	605189	O94907
37	FZD10	11211	6005762	11211	606147	Q9ULW2
38	RAC1	5879	9845511	5879	602048	A4D2P1, P63000
39	DAAM1	23002	21071077	23002	606626	Q9Y4D1
40	VANGL2	57216	62955805	57216	600533	Q9ULK5
41	PRICKLE1	144165	222136680	144165	608500	B3KVG3, Q96MT3
42	WNT9A	7483	15082261	7483	602863	O14904
43	MAPK8	5599	4506095	5599	601158	P45983
44	RHOA	387	10835049	387	165390	P61586, Q9BVT0
45	ROCK1	6093	4885583	6093	601702	Q13464
46	AXIN1	8312	27501450	8312	603816	O15169
47	CSNK1E	1454	4503093	1454	600863	P49674, Q5U045
48	GSK3B	2932	225903437	2932	605004	P49841, Q6FI27
49	PPP2CA	5515	4506017	5515	176915	B3KUN1, P67775
50	PLCB1	23236	12083581	23236	607120	Q9NQ66
51	CAMK2A	815	25952114	815	114078	A8K161, Q8IWE0
52	CHP	11261	6005731	11261	606988	Q99653
53	PRKCA	5578	4506067	5578	176960	B5BU22, P17252, Q7Z727
54	WNT5A	7474	40806205	7474	164975	P41221
55	NFAT5	10725	164419746	10725	604708	A2RRB4, Q7LA65
56	TP53	7157	120407068	7157	191170	P04637, Q3LRW3, Q53GA5
57	SIAH1	6477	55749557	6477	602212	Q8IUQ4

We have tried to incorporate as many recent articles as possible to this piece of work. But this search is not exhaustive, as everyday some new findings about human Wnt STP are getting published and a considerable number of them are not freely accessible to us. In this article, we concentrate

on the mined information and its analysis rather than the mining techniques. By no means, it is an exhaustive study; still it gives us important ideas about disease relatedness, interdependence and comorbidity.

Analysis of the networks is done by Network Analyzer plug-in release 2.7. [413] of the open-source software package Cytoscape version 2.8.2 [414]. Cytoscape is a popular bioinformatics package for biological network visualization and data integration. Network Analyzer computes a comprehensive set of topological parameters for undirected and directed networks including the number of nodes, edges and connected components; the network diameter, radius, and clustering coefficient as well as the distribution of degrees, neighborhood connectivities, average clustering coefficients, shortest path lengths, number of shared neighbors and stress centrality. In addition, network analyzer constructs the intersection, union, and difference of two networks. It supports the extraction of connected components as separate networks and the removal of self loops.

6.3 Results and Discussion

We have found a total of 112 different kinds of diseases. In general, they are termed as diseases throughout this manuscript for better understanding. Out of these 112 diseases, 66 are different kinds of cancer, while 46 are non-cancerous and *in vivo* events including eyesight problems, heart and lungs complications, mental illness, nervous disorders and organ duplications arising due to developmental and environmental issues, and genetic makeup. Diseases related to Wnt STP affect multiple tissue types (blood, bone marrow, cartilage, endometrial, epithelial, germ, lymph, mesothelial, muscle, nerve and skin) of various organs like artery, bone, brain, breast, ear, elementary canal, esophagus, eye, face, heart, kidney, liver, lungs, mouth, neck, nose, ovary, pancreas, prostate, skeleton, skin, spine, testis, throat, thumb, urinary bladder, uterus and glands (lymph, thyroid, parathyroid, salivary, parotid and sebaceous glands). Some diseases are found strictly in children while the others are found in all age range [8].

These diseases are associated with 48 out of 57 human Wnt STP genes.

Nine genes (CAMK2A, CER1, CHD8, CHP, CSNK1A1L, MAPK8, NKD1, PRKACA and PRKCA) have not been found to be associated with any of the aforementioned diseases. The list of considered genes is given in Table 6.1. Here the term 'association' means: i) the gene is mutated and causes the disease, ii) the gene is experimentally found to have differential expression in diseased tissues, iii) the gene may have role in the disease, iv) the gene is regulated by a third party molecule and possibly that molecule has role in the disease, v) the gene is a therapeutic target to stall the disease progression, and vi) the gene is a therapeutic target to overcome the diseased state. If a gene, by any of the above means, is found to be associated with a disease, we have declared it as a valid gene-disease association. These associations are used to create a gene-disease network (Figure 6.1).

6.3.1 The gene-disease network

The network (Figure 6.1) is made of 200 individual associations among 48 genes and 112 diseases. We have calculated various parameters of this network to study its properties. As no direct gene-gene or disease-disease associations are there in this network, clustering co-efficient[12] has been found to be zero. Nine different connected components[13] have been noticed in this network, the largest being of 144 nodes. The number of connected components indicates the connectivity of a network. A lower number of connected components suggest a stronger connectivity. On the other hand, a very large connected component ensures stronger connectivity among most of the nodes irrespective of the presence of other very small connected components. Here the remaining connected components are of the size of 2 each except one of size 3. The maximum of shortest path lengths between a pair of nodes of

[12]Clustering coefficient of a node n is the ratio N / M, where N is the number of edges between the neighbors of n, and M is the maximum number of edges that could possibly exist between the neighbors of n. The clustering coefficient of a node is always a number between 0 and 1.

[13]In an undirected network, two nodes are connected if there is a path of edges between them. Within a network, all the nodes that are pairwise connected form a connected component.

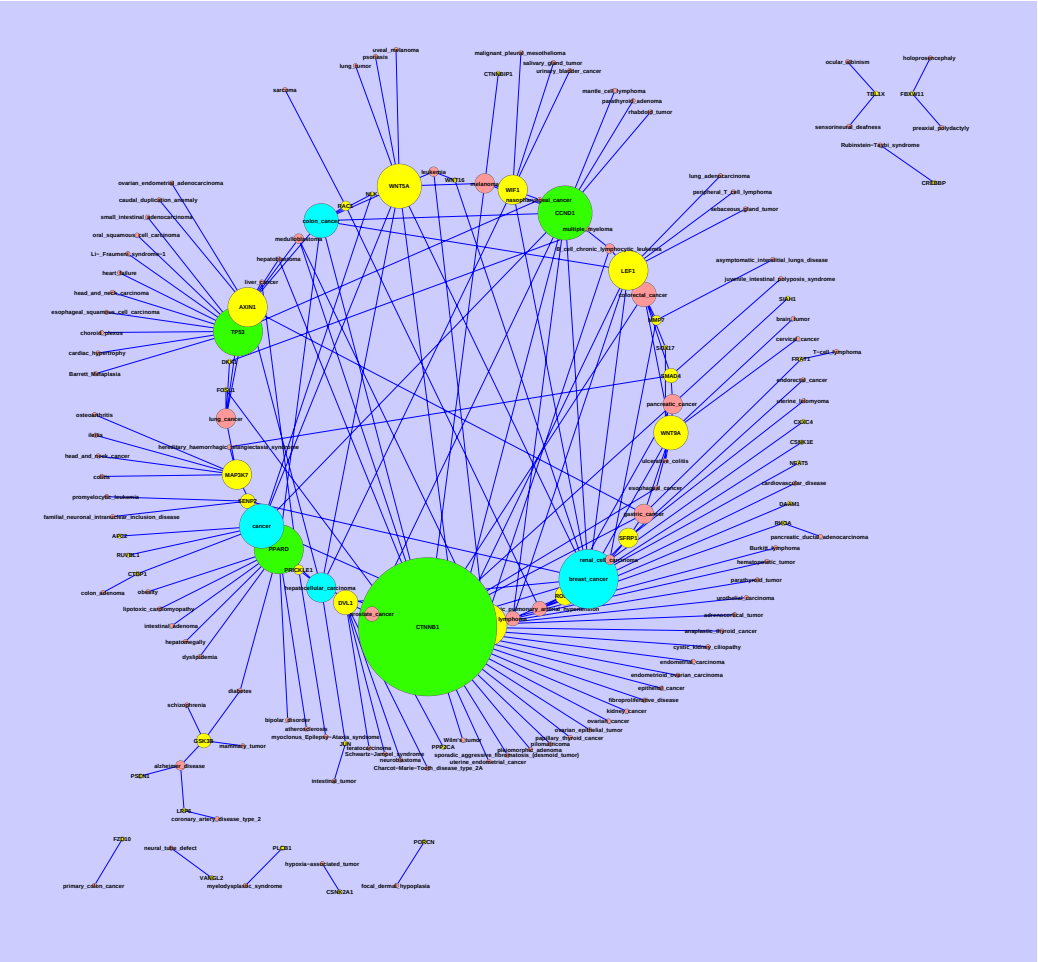


Figure 6.1: The gene-disease network

[Number of genes: 48; number of diseases: 112; number of gene-disease associations: 200. Node sizes are indicative of their degree distribution. Wnt STP genes are depicted in yellow while the four highly connected genes are marked with green. Similarly, diseases are marked with red while the four highly connected diseases are marked with aqua]

the network (network diameter[14]) has been found to be 11 while the shortest possible path length (network radius[15]) being 1.

Network centralization score that defines network topology, is 0.168. Net-

[14]Network diameter is the largest distance (in terms of edges) between two nodes. The network diameter and the shortest path length distribution may indicate small-world properties of the analyzed network.

[15]Network radius is the minimum among the non-zero path-lengths of the nodes in the network.

works whose topologies resemble a star[16] have centralization score close to 1, whereas decentralized networks are characterized by having centralization score close to 0. In that sense, the gene-disease network is decentralized, *i.e.*, clear-cut hub nodes are not present. However, there is a lot of highly connected nodes. Colorectal-cancer is the only disease common among associated diseases of the highly connected nodes representing genes CTNNB1, CCND1 and PPARD. Altogether these genes, along with Tp53, are associated with 55 diseases (48.67% of the total number of diseases). Likewise, among the diseases, breast cancer is highly prevalent, which is found to be associated with 13 genes of the Wnt STP followed by colon cancer (9 genes) and hepatocellular carcinoma (7 genes). These observations confirm that the one disease-one target gene concept is not applicable to the genes of an established pathway, and diseases are way beyond complex events as more and more number of genes are being found to be associated with them.

The average shortest path length, also known as the characteristic path length, is 4.261. It gives the expected distance between two connected nodes. The average number of neighbors (2.484) indicates the average connectivity of a node in the network. A normalized version of this parameter is the network density that lies between 0 and 1. It shows how densely the network is populated with edges (ignoring self-loops and duplicated edges). A network which contains no edge with solely isolated nodes has the density value of 0. In contrast, the density of a clique (a network where each node is connected with every other node) is 1. A network density of 0.016 indicates that the gene-disease network is not densely populated. There is a lot of sparse connections in the network. Logarithmic scatter plot of node-degree[17] distributions of the gene-disease network is given in Figure 6.2. A power low[18] ($y = ax^b$) can be fitted to the points of the plot where y = log(number of nodes), x = log(degree), a = 53.08 and b = -1.433 (correlation value = 0.982

[16]In star topology, each node is connected to central hub(s) with a point-to-point connection. The hub(s) represent single point of control as well as single point of failure.

[17]Node degree of a node n is the number of edges linked to n.

[18]A power law is a mathematical relationship between two quantities. When the frequency of an event varies as a power of some attribute of that event (e.g. its size), the frequency is said to follow a power law.

and R-squared value = 0.888). The correlation value represents correlation between the data points and corresponding points on the line. R-squared value is also known as coefficient of determination. It depicts the relative confidence of a model for fitting into a hypothesized property. Here high correlation and R-squared values depict that the gene-disease network is a scale-free network [8].

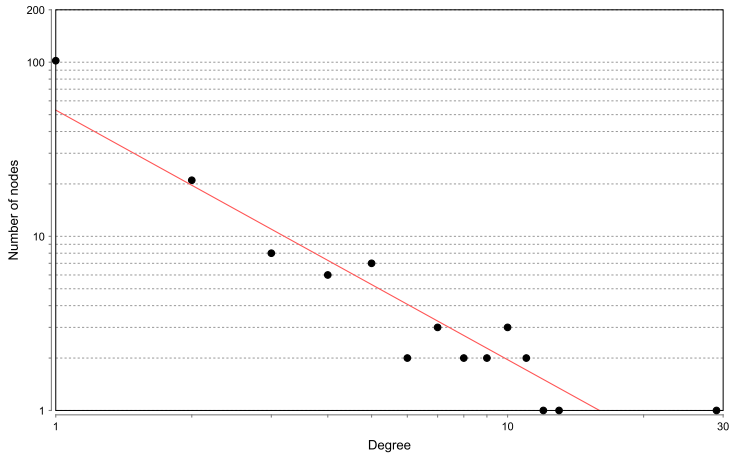


Figure 6.2: Logarithmic scatter plot of node-degree distributions of the gene-disease network
A power low ($y = ax^b$) can be fitted to the points of the plot where a = 53.08 and b = -1.433 (correlation value = 0.982 and R-squared value = 0.888).The network is a scale-free network.

6.3.2 The disease network

It is a network of 107 nodes (diseases) out of total 112 diseases associated among themselves with 823 edges as seen in Figure 6.3. Only 5 diseases (hypoxia-associated tumor, myelodysplastic syndrome, neural tube defect, primary colon cancer and Rubinstein-Taybi syndrome) are found not to be associated with any other disease. They turned out to be the 2-node components of the gene-disease network (Figure 6.1). Other related parameters of the network are given in Table 6.2. Here we discuss some of them to reveal the inherent properties of the disease network. A high clustering co-efficient (0.809) indicates presence of modules in the network. On an average, each

node is connected to 15 neighbors (13.27% of the total nodes). It indicates the high rate of comorbidity among the diseases.

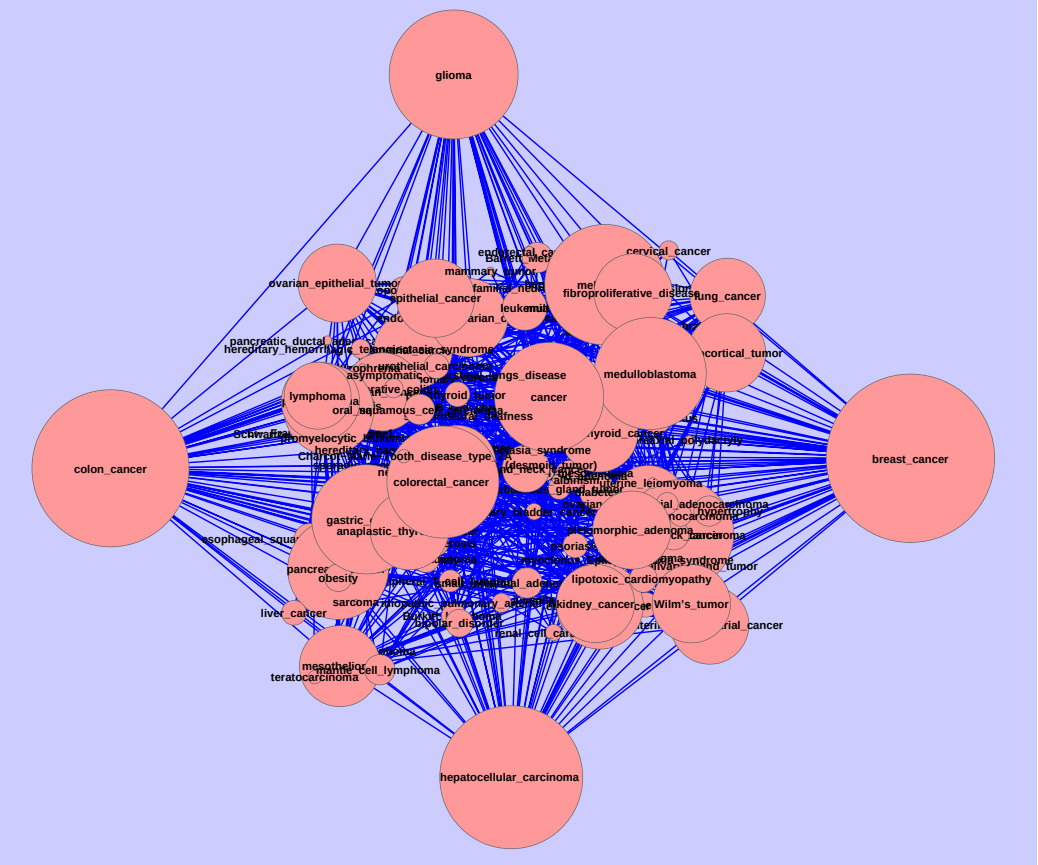


Figure 6.3: The disease network

Node size is proportional to degree of nodes. The four maximally connected nodes are presented as four corner nodes.

Four maximally connected diseases are breast cancer, colon cancer, hepatocellular carcinoma and glioma in decreasing order. They are depicted as the four corner nodes in Figure 6.3. Nodes are sized proportionately according to their degree. First neighbors of breast cancer have included the other three maximally connected nodes along with most of the highly connected nodes of the network confirming our suspicion that pathway related diseases are linked among themselves as seen in Figure 6.4. Logarithmic scatter plot of node-degree distributions of the disease network is given in Figure 6.5.

A power low ($y = ax^b$) can be fitted to the points of the plot where y = log(number of nodes), x = log(degree), a = 10.151 and b = -0.548 (correlation value = 0.372 and R-squared value = 0.375). A low correlation and R-squared value depict that this network is not scale free [8].

Table 6.2: Network parameters of the disease network in Figure 6.3

Parameter	Value
Clustering co-efficient	0.809
Connected Components	3
Network diameter	5
Network radius	1
Network centralization	0.426
Shortest paths	10716 (92%)
Characteristic path length	2.2
Average no. of neighbors	15.241
Network density	0.142

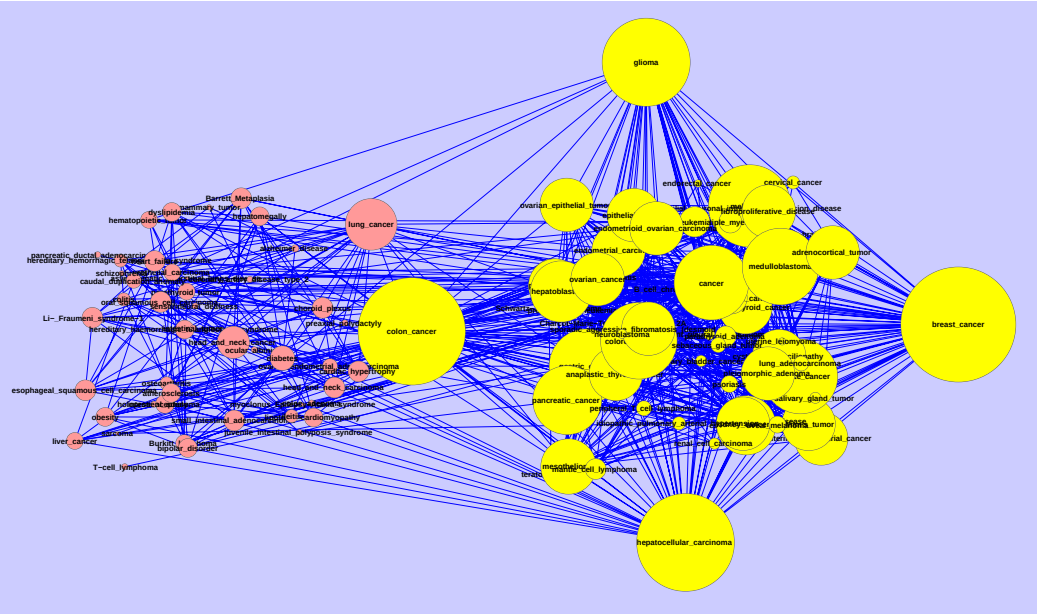


Figure 6.4: First-neighbors of maximally connected node representing “breast cancer” (in yellow)

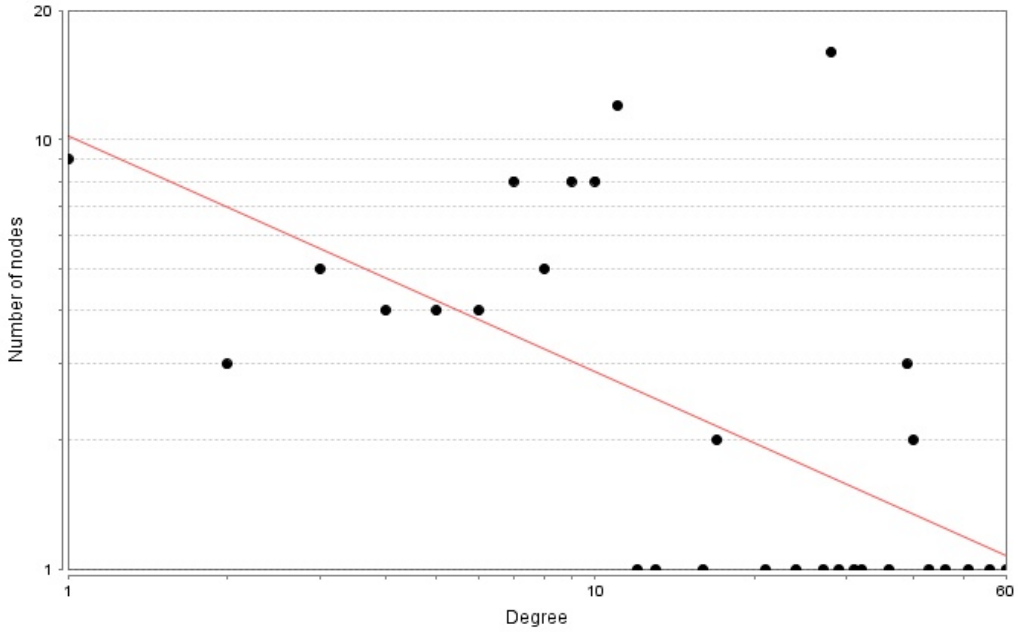


Figure 6.5: Logarithmic scatter plot of node-degree distributions of the disease network
A power low ($y = ax^b$) can be fitted to the points of the plot where a = 10.151 and b = -0.548 (correlation value = 0.372 and R-squared value = 0.375). The network is not scale-free.

6.3.3 Cancerous and non-cancerous disease networks

Here we have categorized the aforementioned 112 Wnt STP associated diseases into two major categories, *i.e.*, 66 cancerous and 46 non-cancerous diseases. We have divided the disease network (in Figure 6.3) into three separate components, *viz.*, a cancerous disease network (Figure 6.6), a non-cancerous disease network (Figure 6.7) and a network linking the cancerous and non-cancerous diseases (Figure 6.8).

The cancerous disease network is a close knit cluster of 61 diseases and 471 unique disease-disease interactions as shown in Figure 6.6. Five types of cancers are found not to be connected with any other kind of cancers, *viz.*, hypoxia associated tumors, mammary tumor, myelodysplastic syndrome, pancreatic ductal adenocarcinoma and primary colon cancer. "Breast cancer" is the maximally connected disease with 40 interactions with other types of cancers (Figure 6.9). Diseases in the network are densely connected. Clustering

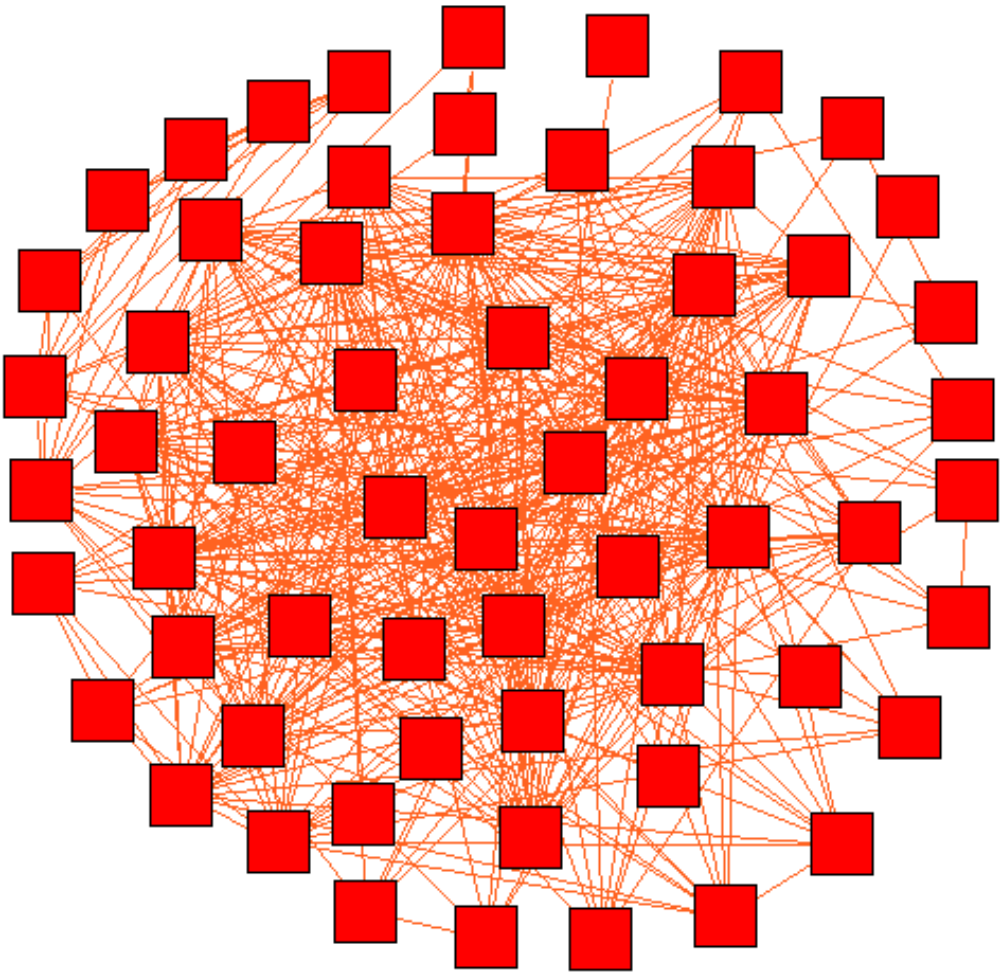


Figure 6.6: The Cancer network
The cancerous disease network is a close knit cluster of 61 diseases and 471 unique disease-disease interactions.

coefficient is 0.801 and on an average each disease shared approximately 15 neighbors with each other. The node "breast cancer", its interactions, first neighbors and their interactions cover 95% of the whole cancer disease network, indicating high comorbidity among Wnt STP related cancers (Figure 6.10).

The non-cancerous disease network (Figure 6.7)) is plotted from 30 non-cancerous diseases and their 47 unique interactions. Out of 46 non-cancerous diseases, 16 have not been found to be present in this network (asymptomatic

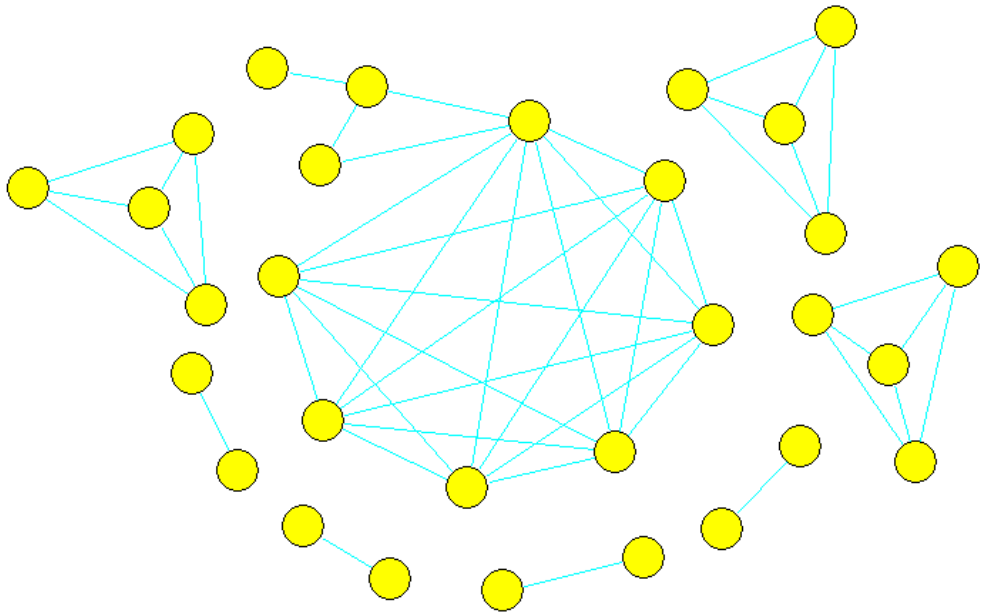


Figure 6.7: The Non-cancerous disease network
The non-cancerous disease network is plotted from 30 non-cancerous diseases and their 47 unique interactions.

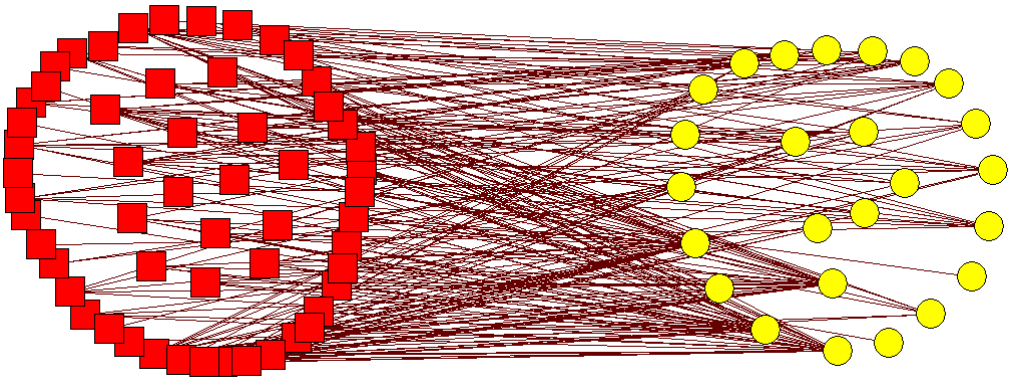


Figure 6.8: Links among cancerous and non-cancerous diseases
Cancerous diseases are marked with red squares while nodes representing the non-cancerous diseases are marked with yellow circles.

interstitial lungs disease, caudal duplication anomaly, familial neuronal intranuclear inclusion disease, focal dermal hypoplasia, juvenile intestinal polyposis syndrome, myoclonus-epilepsy-ataxia syndrome, neuronal tube defect,

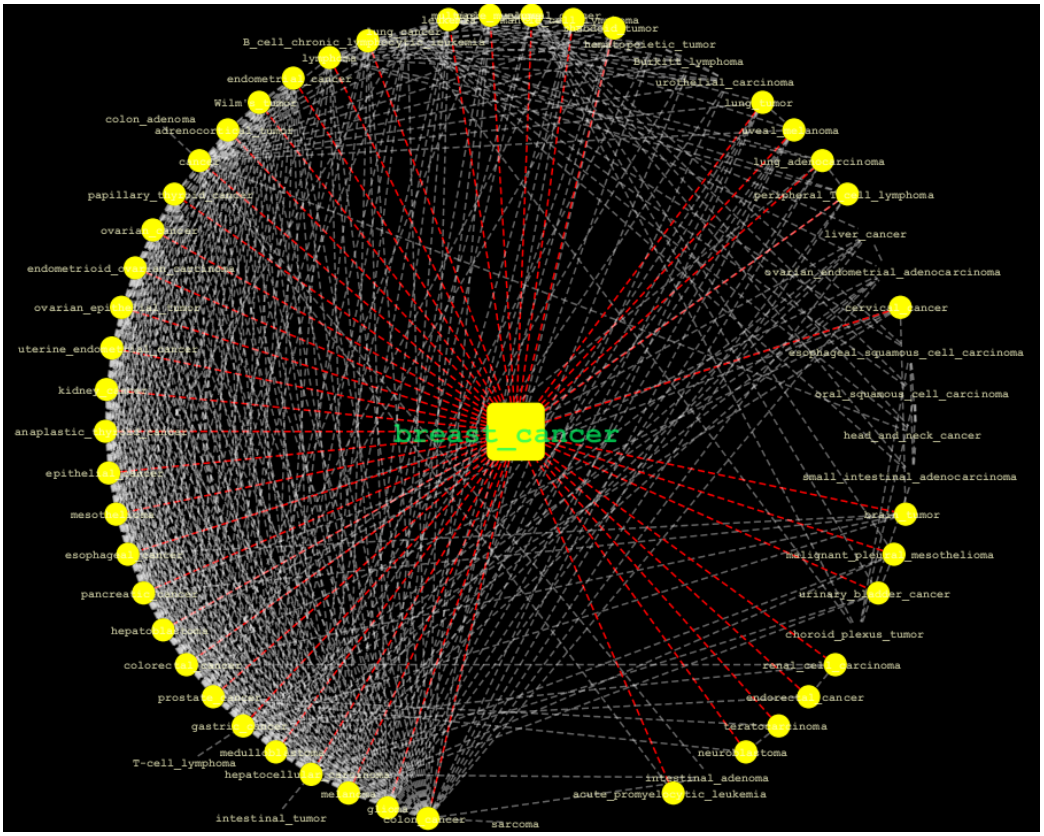


Figure 6.9: The maximally connected node "breast cancer" with its adjacent edges, and first neighbors in the cancerous disease network

parathyroid tumor, parathyroid adenoma, pleiomorphic adenoma, psoriasis, Rubinstein-Taybi syndrome, salivary gland tumor, sebaceous gland tumor, ulcerative colitis and uterine leiomyoma). Unlike the cancerous disease network, this network is sparsely connected. Clustering coefficient is less (0.663) and the diseases share very few neighbors with each other ($\sim$3) indicating less comorbidity and individualistic behavior of these diseases.

6.3.4 Co-morbidity

In a disease network, the nodes are diseases and the edges are the established associations among the diseases. Here, we have established associations among the diseases via common genes. Hence two disease are associated

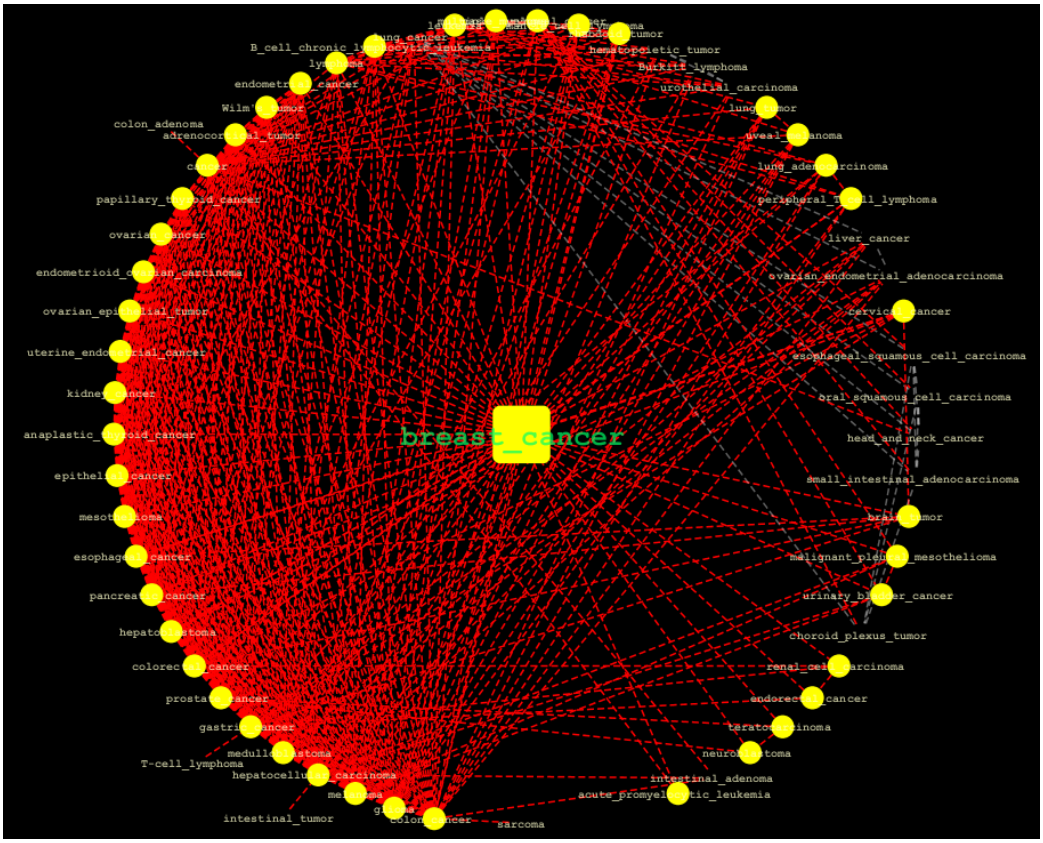


Figure 6.10: The maximally connected node "breast cancer" with its edges, and first neighbors and their edges in the cancerous disease network

if they share atleast one common causative gene. Likewise, there will be multiple diseases and many associations among themselves, if all the genes of a pathway are under consideration. These associations are nothing but indications of existing comorbidity among these diseases. An overall network view of the gene-disease network (Figure 6.1), disease network (Figure 6.3), cancer network (Figure 6.6), non-cancerous disease network (Figure 6.7), and the link network between the cancerous and non-cancerous diseases (Figure 6.11) is provided in Table 6.3. The denser is the network, the higher is its comorbidity due to high rate of shared neighbors. As evident from the table, the cancerous disease network is the densest network (network density: 0.257) with high degree of interaction among the diseases (average number of neigh-

bors: ∼15). On the other hand, the non-cancerous disease network is quite sparse (network density: 0.108) with less number of interactions among the diseases (∼3). These facts conclude that Wnt STP related cancerous diseases show high degree of comorbidity among themselves, while the non-cancerous diseases do not.

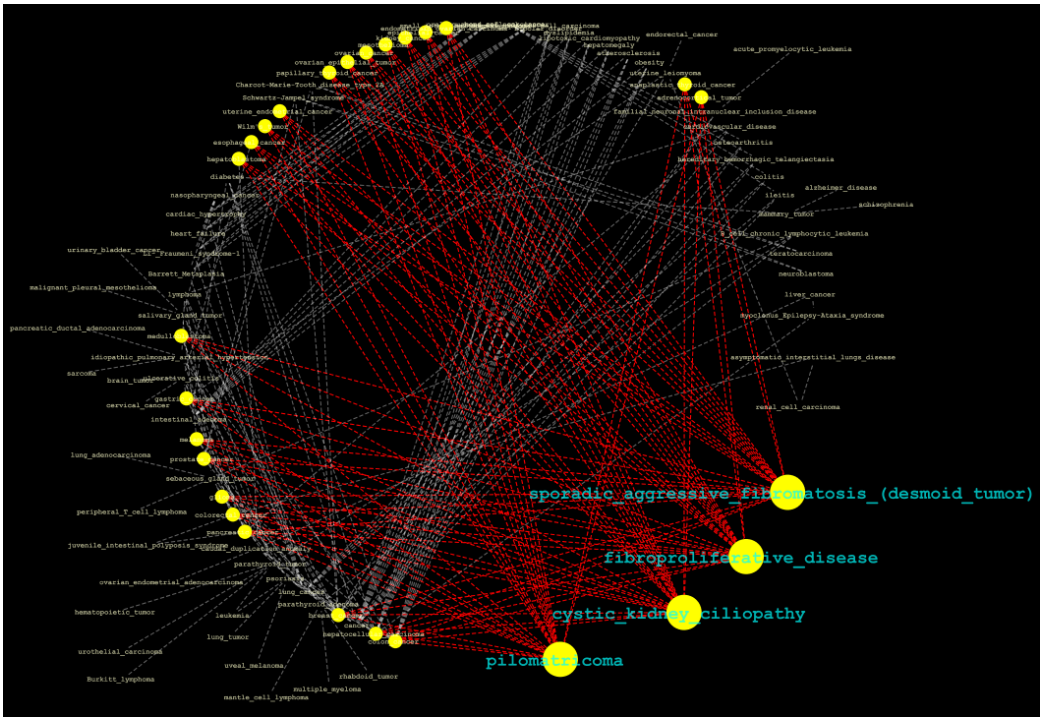


Figure 6.11: The maximally connected nodes of the link network
The network is plotted out of 95 diseases and 243 unique associations among cancerous and non-cancerous diseases. The maximally connected nodes "pilomatricoma", "cystic kidney ciliopathy", "fibroproliferative disease" and "desmoid tumor" are shown in this figure with their adjacent edges, and first neighbors in the connection network among cancerous and non-cancerous diseases.

Table 6.3: Properties of various networks

Type of network	Clustering coefficient	Average number of neighbors	Network density	Network heterogeneity
Gene-disease network	0	2.484	0.016	1.323
Disease network	0.809	15.241	0.142	0.882
Cancer network	0.801	15.443	0.257	0.741
Non-cancerous disease network	0.663	3.133	0.108	0.626
Link network	0	5.116	0.054	0.968

6.3.5 Modules in the cancerous disease network

The cancerous disease network (Figure 6.6) is a dense network with high degree of interaction among the nodes and it has not resolved into separate components like the non-cancerous disease network (Figure 6.7). So, for a deeper level of inspection, we have divided the 61 different types of cancerous diseases according to the tissue or organ they affect. Hence, the cancerous diseases are divided into 15 categories (blood, brain, colon, esophagus, intestine, kidney, liver, lungs, lymph, multiple organs, ovary, thyroid gland, urinary bladder, uterus and single diseases). When a particular type of cancer is found to affect more than one type of tissue or organ (*e.g.*, colorectal cancer affects colon and rectum), it is listed in the "multiple organ" category. When a single type of tissue/organ specific diseases and associations among them are considered, 10 different modules have emerged (Figure 6.12). They are named numerically starting from module 1 in an anti-clock wise manner with the original cancerous disease network (Figure 6.6) in the center. Module 1 shows associations among blood related cancers. Multiple myeloma and B-cell chronic lymphocytic leukemia are associated via gene CCND1. Leukemia and B-cell chronic lymphocytic leukemia, hematopoetic tumor and B-cell chronic lymphocytic leukemia, and leukemia and hematopoetic tumor are linked via gene MYC. Associated liver cancers are included in module 2. Anaplastic thyroid cancer and papillary thyroid cancer (cancers of the thyroid gland) have been associated via gene CTNNB1 in module 3. Brain related cancers (glioma and medulloblastoma) are in module 4.

Module 5 includes kidney related cancers. Wilm's tumor and kidney cancer, Wilm's tumor and adrenocortical tumor, and kidney cancer and adrenocortical tumor are associated via gene CTNNB1. Module 6 showcases diseases of tissue or organ categories (bone, breast, epithelium, eyes, mammary gland, mesothelium, mouth, nerve, pancreas, pancreatic duct, prostate, rectum, skin, stomach, T-cell, throat and testis) in which a single disease is listed that we could not consider independently. Instead, we have found out associations among such types of cancers and listed them under a separate module. It has turned out to be the dense module of the cancerous disease

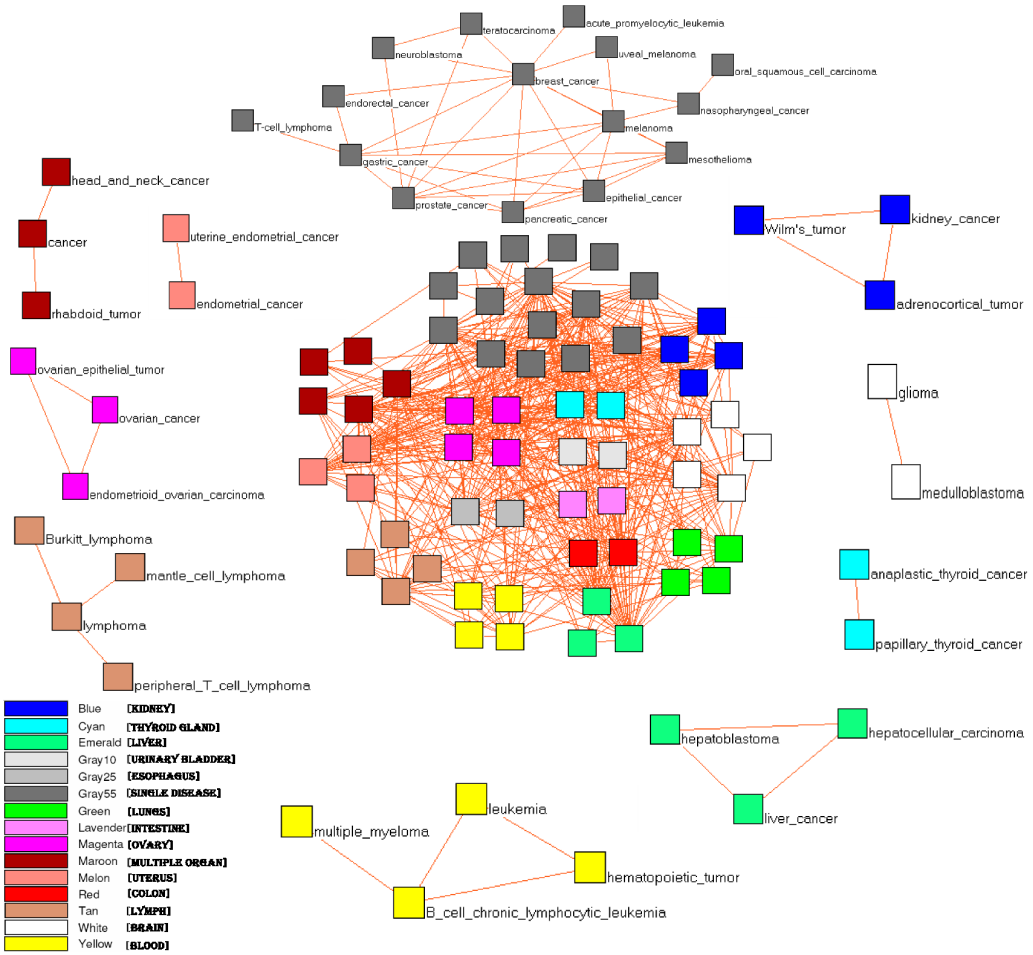


Figure 6.12: Modules of the cancerous disease network

Modules are named numerically starting with Module 1 (containing "multiple myeloma", "leukemia", "hematopoetic tumor" and "B-cell chronic lymphocytic leukemia") from mid-bottom anti-clock-wise manner.

network. Associations among cancers that affect multiple types of tissue or organ are shown in module 7. Associated cancers of uterus are in module 8. Module 9 lists cancers of ovary. Ovarian epithelial tumor and ovarian cancer, ovarian epithelial tumor and endometrioid ovarian carcinoma, and ovarian cancer and endometrioid ovarian carcinoma are associated via gene CTNNB1. Module 10 contains lymph related cancers (Burkitt lymphoma, mantle cell lymphoma, lymphoma, peripheral T-cell lymphoma) and their associations. We have not found any association among colon, esophagus,

intestine, lungs and urinary bladder associated diseases.

6.3.6 Modules in the non-cancerous disease network

The non-cancerous disease network (Figure 6.7) has been divided into 8 separate modules as seen in Figure 6.13. Module 1 contains eye and ear sensory disorders ("ocular albinism" and "sensorineuronal deafness" linked via gene TBL1X). Congenital disorders belong to module 2 ("Holoprosencephaly" and "preaxial polydactyly" linked via gene FBXW11). Generally high blood pressure in the arteries of the lungs ("idiopathic pulmonary hypertension") makes the heart work harder to force the blood through these vessels. Over time this leads to "cardiovascular diseases". These two disorders belong to module 3. Module 4 contains muscle and nerve related disorders affecting human body movement and flexibility ("Charcot-Marie-Tooth disease type 2A" and "Schwartz-Jampel syndrome" linked via gene DVL1). Inflammatory bowel diseases came under the same component ("ileitis" and "colitis" linked via gene MAP3K7 in module 5).

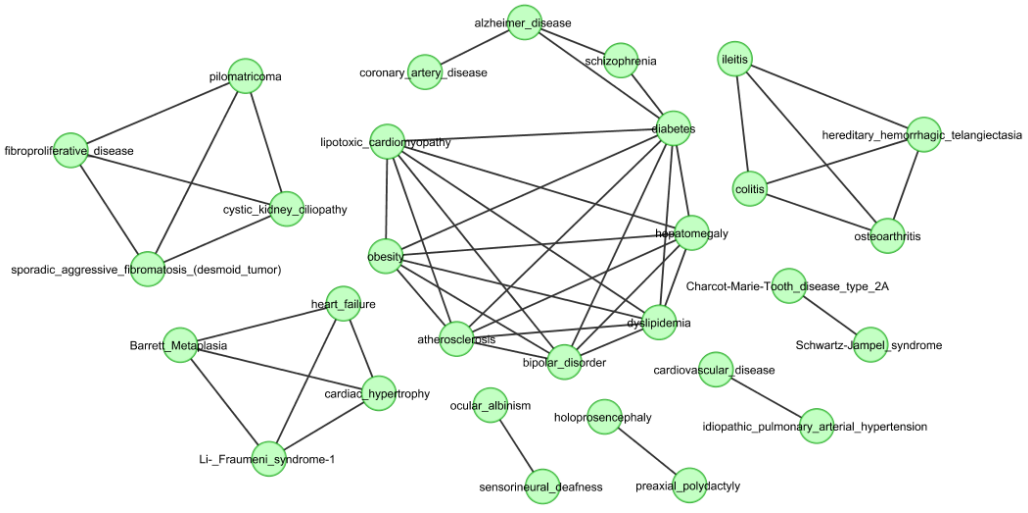


Figure 6.13: Modules of the non-cancerous disease network
Modules are named numerically starting with Module 1 (containing "ocular albinism" and "sensorineuronal deafness") from mid-bottom anti-clock-wise manner.

Module 6 is the largest component of non-cancerous disease network

that has clustered obesity related complicacies ("diabetes", "coronary heart disease", "dyslipidemia", "hepatomegaly", "lipotoxic cardiomyopathy" and "obesity") along with brain disorders like "Alzheimer's disease", "Schizophrenia" and "bipolar disorder". It is well known that overweight issues due to bad food habits and over-eating in early and adult life increase risk of mental illness and mood disorders. Obesity and diabetes significantly and independently increase risk for Alzheimer's disease [415]. It is one of the most common physical health problems among patients with severe and persistent mental illnesses, such as schizophrenia [416]. Studies have reported that up to 60% of individuals with schizophrenia and 68% of those with bipolar disorder are overweight/obese [417]. On the other hand, patients with bipolar disorder, in particular, are at greater risk for overweight and obesity than individuals in the general population [418].

Module 7 represents an aggregate of tumors and cyst related disorders ("pilomatricoma", "desmoid tumor", "fibroproliferative disease" and "cystic kidney ciliopathy"). They are the four highly connected non-cancerous diseases which have maximum (23) associations with cancerous diseases as evident from the link network between cancerous and non-cancerous diseases (Figure 6.11). The network is plotted out of 95 diseases and 243 unique associations among cancerous and non-cancerous diseases. Module 8 contains heart related disorders ("cardiac hypertrophy", "heart failure") along with "Barrett Metaplasia" and "Li-Fraumeni syndrome-1".

6.4 The human Wnt diseasome database

The gene-disease and disease-disease associations analyzed in this chapter can be downloaded from the online human Wnt diseasome database server[19] in .csv format. A snap shot of the webserver is provide in Figure 6.14. The homepage of the online database provides an introduction of the diseasome, along with some useful links for further reference. One can download the whole or partial data from the browse page of the webserver by selecting all the genes or a few genes of interest in the browse page. Data can be

[19]http://www.isical.ac.in/∼rajat/diseasome/index.php

downloaded at multiple levels (Gene-disease associations along with their citations, unique diseases associated with the selected genes and disease-disease associations.

6.5 Conclusive remarks

In this chapter we have built and analyzed a manually curated diseasome (disease map) from 57 genes of the Wnt STP. Disease pathways or diseasomes are potential knowledge bases that throw light on multiple disease related complicacies. Disease pathway studies help one to uncover potential knowledge about their overall function that can contribute to cure human diseases. In this chapter, we have built a gene-disease network from the genes of the human Wnt STP. In this network, a single disease is found to be associated with many genes, *e.g.*, breast and colon cancer among others. Moreover, a single gene (*e.g.*, CTNNB1 and CCND1 among others) is found to be associated with many diseases showing inter-linkage among genes and diseases of a signal transduction pathway. A disease network has been inferred from the gene-disease network. It throws light on disease comorbidity among Wnt STP associated diseases. When divided into cancerous, non-cancerous and link (between cancerous and non-cancerous disease) networks, they have showcased their individual network properties. The cancerous network has been found to be denser; hence the cancerous diseases are more co-morbid. In addition, we have found out some independent co-morbid disease modules in both the cancerous and non-cancerous disease networks. If presence of one disease influences occurrence or risk of the other disease(s) then they definitely suggest personalized systems medicine rather than targeting one gene or one disease at a time for therapeutic purposes. The whole data analyzed in this chapter is available in a publicly accessible webserver[20].

[20]http://www.isical.ac.in/∼rajat/diseasome/index.php

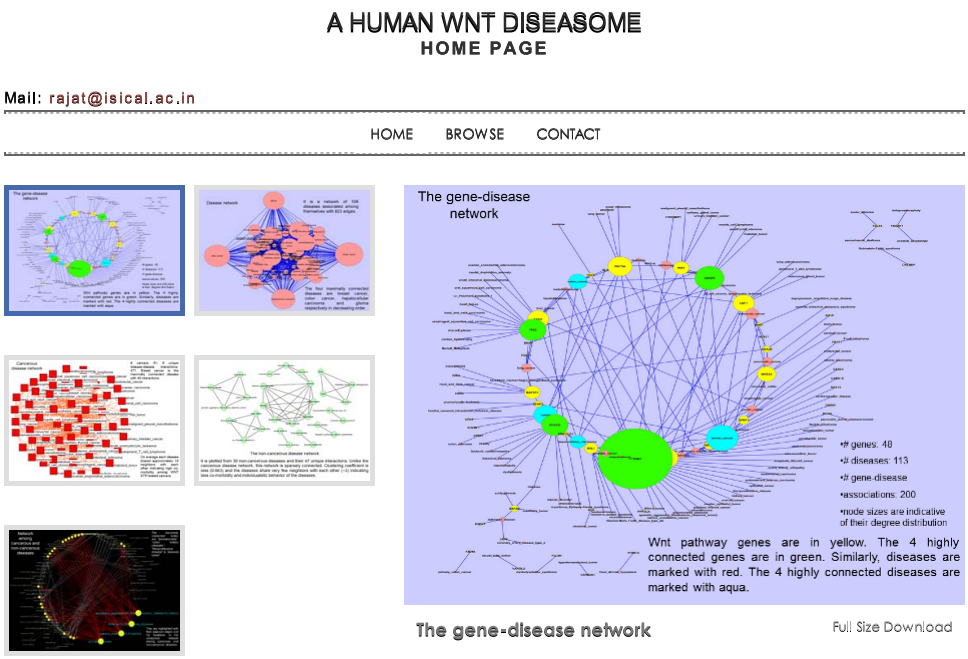


Figure 6.14: The Wnt Diseasome Webserver
It is available at http://www.isical.ac.in/~rajat/diseasome/index.php

Chapter 7

Conclusions and Scopes for Further Research

7.1 Conclusive remarks

Signal Transduction Pathways (STPs) represent sophisticated intracellular biological phenomena. They operate individually as well as in groups. Members of one pathway communicate among themselves as well as with other pathways (inter-pathway communication). They altogether exist and act like a spider's web like system. The turning point is that some disturbances in this 'signaling web' create disease like states in human body. In literal sense, in a string mesh, when we pull one string, all the other connected strings feel the pull with varied effects. Similarly, in a signaling web, multiple genes get affected by behavioral inconsistency of a gene. These disturbances need to be understood properly for devising any kind of remedy. Hence STPs are hot research topic *in vivo*, *in vitro* and *in silico*. Knowledge about them will enhance our capability to track down the root cause for signaling related diseases and their cellular effects.

In this chapter, we conclude the thesis, along with scopes for further research. We have done extensive review on MAPK, Ca^{2+} and Wnt STPs in Chapter 2. MAPK STPs are generalized processes taking part in most of the household events of cells. Ca^{2+} STP is important as gene expression events are effected by concentration as well as mode of entry of Ca^{2+} ions inside a cell. Wnt STPs are involved in multiple biological processes. Perturbations of these pathways are associated with various human pathologies including cancers. The succeeding chapters have used these STPs, especially the Wnt STP, as data. All the three pathways have been considered as data pathways in Chapter 3 for creation of modules. Human Wnt STP has been used as the data pathway for performance comparison of some partitioning algorithms in Chapter 4 and creation of a Wnt diseasome in Chapter 6. Species-specific Wnt STPs have been used as data for deriving a phylogenetic tree from modules in Chapter 5. Wnt STP genes and their associated diseases have been manually curated for construction of a human Wnt diseasome in Chapter 6.

We have developed an algorithm for modularizing STPs in Chapter 3. The algorithm has been applied to human MAPK, Ca^{2+} and Wnt STPs. We have found some significant modules from human MAPK STP. We have

successfully conferred biological significance to the modules obtained from human Ca^{2+} STP. Except one very small module, the remaining seven modules of Wnt STP were found to be biologically significant. The algorithm has successfully identified functional modules from MAPK, Ca^{2+} and Wnt STPs by centralizing the highly connected nodes. Some significant modules have been obtained from the human MAPK STP for complexity level (c-value) of 3. The comparative study of MAPK STPs among 7 species have shown gradual development of the pathway from *P. troglodytes* to *H. sapiens* via *S. scrofa*, *P. troglodytes*, *B. taurus* and *R. norvegicus*. We have successfully conferred biological significance to the modules obtained from human Ca^{2+} STP for complexity level for $c = 3$. The comparative study indicates gradual increase in development of Ca^{2+} STPs starting from *P. troglodytes* to *H. sapiens* via *C. familiaris*, *S. scrofa*, *B. taurus*, *M. musculus* and *R. norvegicus*. Eight modules have been obtained from the human Wnt STP for $c = 3$. Modules of 31 species-specific Wnt STPs have been compared to find conservation among them for the same c-value. Module PLC has been found to be the most conserved module, being present in a maximum number of 17 species.

In Chapter 4, we have done a comparative analysis among five different partitioning algorithms. Three of them follow graph partitioning techniques while the rest two follow community finding and modularization techniques respectively. The partitions generated by these algorithms have been validated by comparing the level of extent to which they could associate to GO terms. A new GO attribute based score (Functional Enrichment score) has been designed for validating these modules. The score establishes a validity index among GO attributes. It can be extended for performance measurement of any kind of partitions/clusters/modules created from biological networks with existing ontology knowledge. Superior performance of the Modularization algorithm in comparison with some traditional graph partitioning and community finding algorithms has been reflected in this chapter.

Chapter 5 has emphasized on deriving a way of representing evolution of Wnt STPs over various species based on the modules obtained by the modularization algorithm. In this chapter, three phylogenetic trees, *viz.*, pathway,

module, and gene trees have been created from different sets of factors derived from three datasets of species-specific STPs (comprising 48, 29 and 12 species), and compared them with widely used reference phylogenetic trees, *i.e.*, NCBI taxonomy tree and 18S rRNA tree. The module tree has been found to bear maximum resemblance with both the reference trees for all the three datasets. Thus modules can be considered as a valid factor for deriving phylogenetic trees from biological pathways. For 48 species, the module tree has displayed distinct clades for the genus Drosophila (*D. erecta*, *D. yakuba*, *D. melanogaster*, *D. simulans*, *D. ananassae*, *D. willistoni*, *D. mojavensis*, *D. virilis* and *D. grimshawi*), class Mammalia (*O. anatinus*, *M. domestica*, *M. musculus*, *R. norvegicus*, *M. mulatta*, *B. taurus* and *H. sapiens*), family Muridae (*M. musculus* and *R. norvegicus*), class Insecta ((*A. gambiae* and *A. pisum*) and (*A. aegypti* and *A. mellifera*)) and genus Caenorhabditis (*C. briggsae* and *C. elegans*). The tree has showed conservation at class level and gradual divergence as we proceed towards the lower ranks, in accordance with the basic principle of evolution.

In Chapter 6, we have built a manually curated diseasome (disease map) with 57 genes of the Wnt STP. Single diseases have been found to be associated with many genes (*e.g.*, breast cancer and colon cancer) in this diseasome. Moreover, single genes (*e.g.*, CTNNB1 and CCND1) have been found to be associated with many diseases showing inter-linkage among genes and diseases of a signal transduction pathway. A disease network has been inferred from the gene-disease network. The disease network has indicated disease comorbidity among Wnt STP associated diseases. When divided into cancerous, non-cancerous and link (between cancerous and non-cancerous disease) networks, they have showcased their individual network properties. The cancerous network has been found to be denser than the non-cancerous disease network. Hence, cancers are more co-morbid than the non-cancerous diseases. In addition, we have found out some independent co-morbid disease modules in both the cancerous and non-cancerous disease networks. Some of these modules have clustered diseases known to co-exist in a single patient (*e.g.*, module 6 of the non-cancerous disease network that links obesity and lifestyle related disorders with brain disorders). Thus presence of one disease

influences occurrence or risk of the other disease(s). Such cases definitely indicate towards personalized systems medicine rather than targeting one gene or one disease at a time for therapeutic purposes. With this inference I conclude the present thesis.

7.2 Scopes for further research

Here we briefly mention some of the scopes for future work. Modularization algorithm described in Chapter 3, works on undirected networks. Designing it to work on directed as well as weighted directed networks will be of value. At present state of the modularization, the user must create multiple sets of modules from a data network for varied c-values and then analyze these sets to choose the optimum set of modules. Such a process can lead to ambiguity as sometimes the user is not well acquainted with the data network to analyze the obtained modules. The process becomes more cumbersome with large networks, where manual analysis of the modules is almost impossible. For such cases, developing a decision process for automatic selection of the optimum c-value forms an integral part of future development of the algorithm.

In Chapter 4, we have done a comparison among a five partitioning algorithms to showcase the major loop-holes in traditional partitioning approaches, when it came to their application to biological pathways, specifically STPs. However, there have been many recent developments in module detection techniques as described in Chapter 1. Including some of them in the comparison will form a strong case for more exhaustive comparison. Moreover, the algorithms have been tested only on human Wnt STP. Extending the data range to some more STPs as well as other kind of biochemical pathways will be of interest to the authors.

In Chapter 5, we have compared a pathway tree constructed with topology dissimilarities with a module tree derived from module dissimilarities. Inclusion of another tree, namely, species tree constructed from gene sequences would have bettered the comparison. Deriving species trees from multiple

gene trees is another major area of work [419]. Concatenation[1] [420] and majority voting[2] approaches have been used earlier for creating a species tree from multiple gene trees [421]. But poor results of these approaches have generated a demand for more smarter methods. In this context, 'Global LAteSt Split (GLASS)' [422] is a simple algorithm for reconstructing phylogenies from multiple gene trees in the presence of incomplete lineage sorting, *i.e.*, when the topology of the gene trees may differ from that of the species tree. Given sufficient genes, it returns the correct topology. However, it systematically overestimates divergence times, leading to biased estimates of species tree branch lengths. This drawback has been overcome in iGLASS [423]. Using these methods one can derive a consensus species tree. The species tree then can be compared with module tree and pathway tree to show case the better tree for representing pathway phylogeny. But, some of the species-specific pathways are not completely discovered yet and some gene sequences are not available. Completion of this task is dependant on availability of data in future.

Work done in Chapter 6 can be advanced with the help of automated literature-mining approaches. Goal of such approaches is to allow researchers to identify necessary information more efficiently, uncover relationships obscured by the sheer volume of available information, and in general to shift the burden of information overload from the researcher to the computer by applying algorithmic, statistical and data management methods to the vast amount of biomedical knowledge that exists in the literature as well as the free text fields of biomedical databases.

[1]Concatenation of the sequences originating from several genes that in turn is used to infer a species tree from the combined data.

[2]Inferring multiple gene trees and taking the most common reconstruction (*i.e.*, take a majority vote) as a species tree.

Appendix

This section lists additional tables in support of valid attribute and functional enrichment score based results given in Chapter 4.

Table A1: Valid attribute score of the individual partitions of algorithms. BP - Biological Process, CC - Cellular Component and GF - Go Full.

		Number of valid attributes		
Algorithm	Partition number	BP	CC	GF
Farhat's	1	0	0	1
	2	3	1	4
	3	1	0	1
	4	13	0	18
	5	0	1	1
	6	1	0	1
	7	2	0	3
	8	1	0	1
	9	0	0	0
	10	2	0	6
	11	8	0	11
Greedy	1	15	3	25
	2	1	0	1
	3	1	0	1
	4	1	0	1
	5	1	0	1
	6	0	0	0
	7	2	0	2
	8	1	0	1
	9	3	0	7
Modularization	1	26	8	43
	2	1	0	4
	3	2	0	2
	4	4	0	5
	5	3	0	4
	6	1	1	2
	7	0	0	0
	8	0	0	0
Newman's	1	7	0	10
	2	2	0	2
	3	0	0	0
	4	0	0	0
	5	89	10	129
	6	70	5	101
	7	62	10	89
	8	11	0	12
Kernighan-Lin's	1	13	0	15
	2	52	12	91

Table A2: Valid attribute score of the algorithms

	Sum total number of valid attributes		
Algorithm	BP	CC	GF
Farhat's	31	2	47
Greedy	25	3	39
Modularization	37	9	60
Newman's	241	25	343
Kernighan-Lin's	65	12	106

Valid attribute score of an algorithm for partitioning a network is the sum total of number of valid attributes associated with each of the partitions. Newman's community finding algorithm is associated with maximum number of attributes. BP - Biological Process, CC - Cellular Component and GF - Go Full

Table A3: FE_score of modules obtained by Modularization algorithm for c = 1,2,3, ..., 13

$c = 1$	$c = 2$	$c = 3$	$c = 4$	$c = 5$	$c = 6$	$c = 7$ & 8	$c = 9$, 10 & 11	$c = 12$ & 13
132.3567369	119.9244759	119.9244759	119.9244759	119.9244759	119.9244759	119.9244759	119.9244759	34.8881761
57.96244639	104.8761268	45.35655974	58.43119024	104.7486068	104.7486068	88.44955719	48.92757948	349.6677017
384.2291667	187.762473	173.7102982	82.77784477	50.50737544	74.11148128	86.06524456	349.6677017	0
177.3557692	141.8846154	31.94552087	163.956564	163.956564	349.6677017	349.6677017	0	-
133.0168269	494.8918547	163.956564	0	0	0	0	-	-
0	0	1811.5625	133.0168269	349.6677017	-	-	-	-
558.6200881	349.6677017	349.6677017	349.6677017	0	-	-	-	-
305.7406017	118.2371795	0	0	-	-	-	-	-
177.3557692	350.0483155	-	-	-	-	-	-	-
177.3557692	119.771284	-	-	-	-	-	-	-
158.6582183	0	-	-	-	-	-	-	-
624.1799554	-	-	-	-	-	-	-	-
0	-	-	-	-	-	-	-	-
294.2688433	-	-	-	-	-	-	-	-
996.2594016	-	-	-	-	-	-	-	-
177.3557692	-	-	-	-	-	-	-	-
0	-	-	-	-	-	-	-	-
0	-	-	-	-	-	-	-	-
0	-	-	-	-	-	-	-	-
0	-	-	-	-	-	-	-	-
217.7357681	*180.6421842*	**337.0154526**	*113.4718254*	*112.6863891*	*129.6904531*	*128.8213959*	*129.6299393*	*128.1852926*

FE_score - Functional Enrichment Score. For $c = 3$, the modules are yielding maximum FE_score of 337.0154526. Hence the set of modules obtained for $c = 3$ have been chosen for comparative analysis of algorithms in Chapter 4.

Table A4: FE_score of different sets of partitions obtained by Greedy algorithm

5	6	7	8	9	10	11
191.4548534	230.7742959	168.4048013	150.6208458	935.2898352	341.3357118	239.4433579
17.92163291	70.94230769	88.67788462	173.7291378	118.224359	118.224359	141.8692308
41.04072262	50.07378745	305.9965904	126.6826923	88.67307692	88.67307692	94.14468959
70.57432941	8.949130863	17.98254142	35.23515664	89.53398058	89.53398058	106.4134615
28.44535928	45.98235794	126.4293497	16.26769466	256.1666667	256.1666667	117.7354802
-	0	32.42871949	132.0469101	0	0	107.4407767
-	-	21.48943789	75.52458474	333.0166667	333.0166667	154.3738832
-	-	-	53.31782466	118.2371795	118.2371795	0
-	-	-	-	6.654730102	55.54191025	0
-	-	-	-	-	180.8235294	249.443973
-	-	-	-	-	-	6.406716454
69.88737952	*67.78697997*	*108.7727607*	*95.42810584*	**216.1996105**	*158.1553081*	*110.6610518*

FE_score - Functional Enrichment Score. For 9 partitions, Greedy algorithm is yielding maximum average FE_score of 216.1996105. Hence these 9 partitions have been chosen for comparative analysis of algorithms in Chapter 4.

Table A5: FE_score of different sets of partitions obtained by Farhat's algorithm

5	6	7	8	9	10	11
118.2602669	178.1565851	167.9952778	343.3079372	61.37506607	190.4704813	558.9090909
10.2556992	37.920601	190.8291597	45.78071542	86.97489613	86.97489613	771.6973448
39.00514532	46.4217363	32.33948143	46.66365363	0	0	141.8846154
22.83760227	15.16368983	0	44.13840736	0	0	109.6464339
17.42439328	53.20673077	30.01673833	0	32.4743943	35.61255062	614.8
-	53.20384615	4.671128666	0	697.669927	697.669927	307.4
-	-	88.67788462	0	42.75106659	42.75106659	75.58438193
-	-	-	27.33607946	0	0	106.4076923
-	-	-	-	101.3461538	108.3840812	0
-	-	-	-	-	88.67788462	18.13576221
-	-	-	-	-	-	156.2484099
41.55662139	*64.01219819*	*73.50423865*	*63.40334913*	*113.6212782*	*125.0540887*	**260.0648847**

FE_score - Functional Enrichment Score. For 11 partitions, Farhat's algorithm is yielding maximum average FE_score of 260.0648847. Hence these 11 partitions have been chosen for comparative analysis of algorithms in Chapter 4.

Bibliography

[1] L Nayak and R K De, "An algorithm for modularization of MAPK and calcium signaling pathways: Comparative analysis among different species," *J. Biomed. Inform.*, vol. 40, pp. 726–749, 2007.

[2] L Nayak and R K De, "Modularized study of human calcium signalling pathway," *J. Biosci.*, vol. 32, pp. 1009–1017, 2007.

[3] L Nayak and R K De, "Developmenatl Trend Derived from Modules of Wnt Signaling Pathway," in *Proceedings of the 4th International Conference on Pattern Recognition and Machine Intelligence (PReMI'11)*, 2011, pp. 400–405.

[4] R K De and L Nayak, "MAPK signaling pathways and their recursive modularization," in *Proceedings of the 15th International Conference on Computing (CIC'06)*, 2006, pp. 203–208.

[5] L Nayak and R K De, "Finding better partitions and conserved modules in Wnt signaling pathways," in *The 2012 International Conference on Bioinformatics and Computational Biology (BIOCOMP'12), Las Vegas, USA*, 2012(Accepted).

[6] L Nayak, N Tomar, and R K De, *Recent Trends in Computational Biology and Computational Statistics Applied in Biotechnology and Bioinformatics*, chapter 14, pp. 337–369, New India Publishing Agency (NIPA), New Delhi, India, 2012 (In Press).

[7] L Nayak and R K De, "Deriving a phylogenetic tree from Wnt Signaling Pathways," *Submitted to be considered for a Young Scientist Award by Indian Science Congress*, 2012.

[8] L Nayak and R K De, "Disease Comorbidity and the Human Wnt signaling Pathway: A network-wise Study," *OMICS: A Journal of Integrative Biology*, 2012 (Communicated).

[9] M Wu and C Chan, "Human Metabolic Network: Reconstruction, Simulation, and Applications in Systems Biology," *Metabolites*, vol. 2, pp. 242–253, 2012.

[10] J A Papin, N D Price, S J Wiback, D A Fell, and B O Palsson, "Metabolic pathways in the post-genome era," *Trends in Biochemical Sciences*, vol. 28, no. 5, pp. 250–258, 2003.

[11] D Voet and J G Voet, *Biochemistry*, John Wiley, New York, 1997.

[12] C H Schilling, D Letscher, and B Palsson, "Theory for the Systemic Definition of Metabolic Pathways and their use in Interpreting Metabolic Function from a Pathway-Oriented Perspective," *J. Theor. Biol.*, vol. 203, pp. 229–248, 2000.

[13] R Schweiger, M Linial, and N Linial, "Generative probabilistic models for protein-protein interaction networks--the biclique perspective," *Bioinformatics*, vol. 27, pp. i142–i148, 2011.

[14] N Blow, "Untangling the protein web," *Nature*, vol. 460, pp. 415–418, 2009.

[15] L Hakes, J W Pinney, D L Robertson, and S C Lovell, "Protein-protein interaction networks and biology-what's the connection?," *Nature Biotechnology*, vol. 26, no. 1, pp. 69–72, 2008.

[16] M Michaut, S Kerrien, L Montecchi-Palazzi, F Chauvat, C Cassier-Chauvat, J Aude, P Legrain, and H Hermjakob, "InteroPORC: automated inference of highly conserved protein interaction networks," *Bioinformatics*, vol. 24, no. 14, pp. 1625–1631, 2008.

[17] P Bork, L J Jensen, C V Mering, A K Ramani, I Lee, and E M Marcotte, "Protein interaction networks from yeast to human," *Current Opinion in Structural Biology*, vol. 14, pp. 292–299, 2004.

[18] G Altay and F Emmert-Streib, "Inferring the conservative causal core of gene regulatory networks," *BMC Systems Biology*, vol. 4, no. 132, 2010.

[19] M Bansal, V Belcastro, A Ambesi-Impiombato, and D D Bernardo, "How to infer gene networks from expression profiles," *Molecular Systems Biology*, vol. 3, no. 78, 2007.

[20] H D Jong, "Modeling and Simulation of Genetic Regulatory Systems: A Literature Review," *Journal of Computational Biology*, vol. 9, no. 1, pp. 67–103, 2002.

[21] M T Laub and M Goulian, "Specificity in Two-Component Signal Transduction Pathways," *Annu. Rev. Genet.*, vol. 41, pp. 121–45, 2007.

[22] J M Berg, J L Tymoczk, and L Stryer, *Biochemistry, 5th edition*, W H Freeman, New York, 2002.

[23] R B Bourret and R E Silversmith, "Two-component signal transduction," *Curr. Opin. Microbiol.*, vol. 13, no. 2, pp. 113–115, 2010.

[24] A M Stock, V L Robinson, and P N Goudreau, "Two-Component Signal Transduction," *Annu. Rev. Biochem.*, vol. 69, pp. 183–215, 2000.

[25] W R Burack and A S Shaw, "Signal transduction: hanging on a scaffold," *Current Opinion in Cell Biology*, vol. 12, pp. 211–216, 2000.

[26] R Seger and E G Krebs, "The MAPK signaling cascade," *FASEB J*, vol. 9, pp. 726–35, 1995.

[27] D E Clapham, "Calcium Signaling," *Cell*, vol. 131, pp. 1047–1058, 2007.

[28] D E Clapham, "Calcium Signaling," *Cell*, vol. 80, pp. 259–268, 1995.

[29] K M Cadigan and Y I Liu, "Wnt signaling: complexity at the surface," *J. Cell Sci.*, vol. 119, pp. 395–402, 2006.

[30] K M Cadigan and R Nusse, "Wnt signaling: a common theme in animal development," *Genes and Development*, vol. 11, pp. 3286–3305, 1997.

[31] E Nadal and F Posas, "Elongating under stress," *Genetics Research International*, vol. 2011, no. 326286, 2011.

[32] R Alonso-Monge, E Roman, D M Arana, J Pla, and C Nombela, "Fungi sensing environmental stress," *Clin. Microbiol. Infect.*, vol. 15, no. Suppl. 1, pp. 17–19, 2009.

[33] A J Morris and C C Malbon, "Physiological Regulation of G Protein-Linked Signaling," *Physiological Reviews*, vol. 79, no. 4, pp. 1373–1413, 1999.

[34] E Falkenstein, H Tillmann, M Christ, M Feuring, and M Wehling, "Multiple Actions of Steroid HormonesA Focus on Rapid, Nongenomic Effects," *Pharmacol. Rev.*, vol. 52, pp. 513–555, 2000.

[35] B Alberts, D Bray, and J Lewis, *Molecular Biology of the Cell, 3rd edition*, Garland Science, New York, 1994.

[36] J E Dumont, S Dremier, I Pirson, and C Maenhaut, "Cross signaling, cell specificity, and physiology," *Am. J. Physiol. Cell Physiol.*, vol. 283, pp. C2–C28, 2002.

[37] B Alberts, A Johnson, and J Lewis, *Molecular Biology of the Cell, 4th edition*, Garland Science, New York, 2002.

[38] R Webster, S Maxwell, H Spearman, K Tai, O Beckstein, M Sansom, and D Beeson, "A novel congenital myasthenic syndrome due to decreased acetylcholine receptor ion-channel conductance," *Brain*, , no. Pt 4, pp. 1070–1080, 2012.

[39] C Toyoshima and N Unwin, "Ion channel of acetylcholine receptor reconstructed from images of postsynaptic membranes," *Nature*, vol. 336, no. 6196, pp. 247–250, 1988.

[40] M Congreve and F Marshall, "The impact of GPCR structures on pharmacology and structure-based drug design," *British Journal of Pharmacology*, vol. 159, no. 5, pp. 986–996, 2010.

[41] W K Kroeze, D J Sheffler, and B L Roth, "G-protein-coupled receptors at a glance," *Journal of Cell Science*, vol. 116, pp. 4867–4869, 2003.

[42] U Gether and B K Kobilka, "G Protein-coupled Receptors," *The Journal of Biological Chemistry*, vol. 273, no. 29, pp. 17979–17982, 1998.

[43] H E Hamm, "How activated receptors couple to G proteins," *Proc. Nat. Acad. Sci., USA*, vol. 98, no. 9, pp. 4819–4821, 2001.

[44] G Pearson, F Robinson, G T Beers, B E Xu, M Karandikar, K Berman, and M H Cobb, "Mitogen-activated protein (MAP) kinase pathways: regulation and physiological functions.," *Endocr Rev.*, vol. 22, no. 2, pp. 153–183, 2001.

[45] H J Schaeffer and M J Webber, "Mitogen-Activated Protein Kinases: Specific Messages from Ubiquitous Messengers," *Molecular And Cellular Biology*, vol. 19, no. 4, pp. 2435–2444, 1999.

[46] R J Davis, "The Mitogen-activated Protein Kinase Signal Transduction Pathway," *The Journal of Biological Chemistry*, vol. 268, no. 20, pp. 14553–14556, 1993.

[47] A Citri and Y Yarden, "EGF-ERBB signalling: towards the systems level," *Nat. Rev. Mol. Cell Biol.*, vol. 7, no. 7, pp. 505–516, 2006.

[48] K Oda, Y Matsuoka, A Funahashi, and H Kitano, "A comprehensive pathway map of epidermal growth factor receptor signaling," *Mol. Sys. Biol.*, vol. 1, no. 2005.0010, 2005.

[49] Y. Yarden and M X Sliwkowski, "Untangling the ErbB Signalling Network," *Nat. Rev. Mol. Cell Biol.*, vol. 2, pp. 127–137, 2001.

[50] M J Hilton, X Tu, X Wu, S Bai, H Zhao, T Kobayashi, H M Kronenberg, S L Teitelbaum, F P Ross, R Kopan, and F Long, "Notch signaling maintains bone marrow mesenchymal progenitors by suppressing osteoblast differentiation," *Nat. Med.*, vol. 14, no. 3, pp. 306–314, 2008.

[51] U Fiuza and A M Arias, "Cell and molecular biology of notch," *Journal of Endocrinology*, vol. 194, pp. 459–474, 2007.

[52] K Tanigaki and T Honjo, "Regulation of lymphocyte development by Notch signaling," *Nat. Immunology*, vol. 8, no. 5, pp. 451–456, 2007.

[53] S J Bray, "Notch signalling: a simple pathway becomes complex," *Nat. Rev. Mol. Cell Biol.*, vol. 7, no. 9, pp. 678–89, 2006.

[54] L Miele, "Notch Signaling," *Clin Cancer Res.*, vol. 12, no. 4, pp. 1074–1079, 2006.

[55] C W Wilson and P Chuang, "Mechanism and evolution of cytosolic Hedgehog signal transduction," *Development*, vol. 137, pp. 2079–2094, 2010.

[56] T R Brglin, "The Hedgehog protein family," *Genome Biology*, vol. 9, pp. 241, 2008.

[57] D J Robbins and M Hebrok, "Hedgehogs: la dolce vita Workshop on Hedgehog-Gli Signaling in Cancer and Stem Cells," *EMBO reports*, vol. 8, no. 5, pp. 451–455, 2007.

[58] J Briscoe and P Thrond, "Hedgehog Signaling: From the Drosophila Cuticle to Anti-Cancer Drugs," *Developmental Cell*, vol. 8, pp. 143–151, 2005.

[59] M Shi, J Zhu, R Wang, X Chen, L Mi, T Walz, and T A Springer, "Latent TGF-β structure and activation," *Nature*, vol. 474, pp. 343–351, 2011.

[60] S Itoh and P Dijke, "Negative regulation of TGF-β receptor/Smad signal transduction," *Current Opinion in Cell Biology*, vol. 19, pp. 176–184, 2007.

[61] A Agrotis, N Kalinina, and A Bobik, "Transforming Growth Factor-*beta*, Cell Signaling and Cardiovascular Disorders," *Current Vascular Pharmacology*, vol. 3, pp. 55–61, 2005.

[62] M Kowanetz and N Ferrara, "Vascular Endothelial Growth Factor Signaling Pathways: Therapeutic Perspective," *Clin. Cancer Res.*, vol. 12, no. 17, pp. 5018–5022, 2006.

[63] A Olsson, A Dimberg, J Kreuger, and L Claesson-Welsh, "VEGF receptor signalling in control of vascular function," *Nat. Rev. Mol. Cell Biol.*, vol. 7, pp. 359–371, 2006.

[64] F J Giles, "The Vascular Endothelial Growth Factor (VEGF) Signaling Pathway: A Therapeutic Target in Patients with Hematologic Malignancies," *Oncologist*, vol. 6, pp. 32–39, 2001.

[65] O V Smirnova, T Y Ostroukhova, and R L Bogorad, "JAK-STAT pathway in carcinogenesis: Is it relevant to cholangiocarcinoma progression?," *World J. Gastroenterol.*, vol. 13, no. 48, pp. 6478–6491, 2007.

[66] K Grote, M Luchtefeld, and B Schieffer, "JANUS under stress–role of JAK/STAT signaling pathway in vascular diseases," *Vascul. Pharmacol.*, vol. 43, pp. 357–363, 2005.

[67] J S Rawlings, K M Rosler, and D A Harrison, "The JAK/STAT signaling pathway," *Journal of Cell Science*, vol. 118, no. 8, pp. 1281–1283, 2004.

[68] C W Schindler, "JAK-STAT signaling in human disease," *J. Clin. Invest.*, vol. 109, pp. 1133–1137, 2002.

[69] I Greenwald, *Introduction to signal transduction(September 9, 2005), WormBook*, WormBook, New York, 2005.

[70] L Nayak and R K De, "Signal Transduction Pathway Resources for Everyday Research," *Everyman's Science*, 2012 (Revision Communicated).

[71] M G Wilkinson and J B A Millar, "Control of the eukaryotic cell cycle by MAP kinase signaling pathways," *The FASEB Journal*, vol. 14, no. 14, pp. 2147–2157, 2000.

[72] E K Kim and E Choi, "Pathological roles of MAPK signaling pathways in human diseases," *BBA - Molecular Basis of Disease,*, vol. 1802, no. 4, pp. 396–405, 2010.

[73] Y Aoki, T Niihori, Y Narumi, S Kure, and Y Matsubara, "The RAS/MAPK syndromes: novel roles of the ras pathway in human genetic disorders," *Hum Mutat.*, vol. 29, no. 8, pp. 992–1006, 2008.

[74] W E Tidyman and K A Rauen, "The RASopathies: Developmental syndromes of Ras/MAPK pathway dysregulation," *Curr. Opin. Genet. Dev.*, vol. 19, no. 3, pp. 230–236, 2009.

[75] M C Lawrence, A Jivan, C Shao, L Duan, D Goad, E Zaganjor, J Osborne, K McGlynn, S Stippec, S Earnest, W Chen, and M H Cobb, "The roles of MAPKs in disease," *Cell Res.*, vol. 18, no. 4, pp. 436–442, 2008.

[76] C S Patil and K L Kirkwood, "p38 MAPK signaling in oral-related diseases," *J. Dent. Res.*, vol. 86, no. 9, pp. 812–825, 2007.

[77] P P Roux and J Blenis, *Apoptosis, Cell Signaling, and Human Diseases: Molecular Mechanisms*, pp. 135–149, Humana Press, 2007.

[78] B D Gelb and M Tartaglia, "Noonan syndrome and related disorders: dysregulated RAS-mitogen activated protein kinase signal transduction," *Hum. Mol. Genet.*, vol. 15, no. 2, pp. R220–R226, 2006.

[79] K L Dunn, P S Espino, B Drobic, S He, and J R Davie, "The ras-MAPK signal transduction pathway, cancer and chromatin remodeling," *Biochem. Cell Biol.*, vol. 83, pp. 1–14, 2005.

[80] H Schramek, "MAP kinases: from intracellular signals to physiology and disease," *News Physiol Sci.*, vol. 17, pp. 62–67, 2002.

[81] E J Cartwright, D Oceandy, C Austin, and L Neyses, "Ca^{2+} signalling in cardiovascular disease: the role of the plasma membrane calcium pumps," *Sci. China Life Sci.*, vol. 54, no. 8, pp. 691–698, 2011.

[82] E Carafoli, "The plasma membrane calcium pump in the hearing process: physiology and pathology," *Sci. China Life Sci.*, vol. 54, no. 8, pp. 686–690, 2011.

[83] R Ficarella, L F Di, M Bortolozzi, S Ortolano, F Donaudy, M Petrillo, S Melchionda, A Lelli, T Domi, L Fedrizzi, D Lim, G E Shull, P Gasparini, M Brini, F Mammano, and E Carafoli, "A functional study of plasma-membrane calcium-pump isoform 2 mutants causing digenic deafness," *Proc. Natl. Acad. Sci., USA*, vol. 104, no. 5, pp. 1516–1521, 2007.

[84] J M Schultz, Y Yang, A J Caride, A G Filoteo, A R Penheiter, A Lagziel, R J Morell, S A Mohiddin, L Fananapazir, A C Madeo, J T Penniston, and A J Griffith, "Modification of Human Hearing Loss by Plasma-Membrane Calcium Pump PMCA2," *N. Engl. J. Med.*, vol. 352, no. 15, pp. 1557–1564, 2005.

[85] S Somlo and B Ehrlich, "Human disease: calcium signaling in polycystic kidney disease," *Curr. Biol.*, vol. 11, pp. R356–R360, 2001.

[86] C Supnet and I Bezprozvanny, "Presenilins as endoplasmic reticulum calcium leak channels and Alzheimer's disease pathogenesis," *Sci. China Life Sci.*, vol. 54, no. 8, pp. 744–751, 2011.

[87] M P Mattson and S L Chan, "Neuronal and glial calcium signaling in Alzheimer's disease," *Cell Calcium*, vol. 34, no. (4-5), pp. 385–397, 2003.

[88] I Bezprozvanny, "Calcium signaling and neurodegenerative diseases," *Trends Mol. Med.*, vol. 15, pp. 89–100, 2009.

[89] S Feske, "Calcium signalling in lymphocyte activation and disease," *Nat. Rev. Immunol.*, vol. 7, no. 9, pp. 690–702, 2007.

[90] E Carafoli, "Calcium signaling: a tale for all seasons," *Proc. Natl. Acad. Sci., USA*, vol. 99, no. 3, pp. 1115–1122, 2002.

[91] J R Prosperi, H H Luu, and K H Goss, *Dysregulation of the Wnt Pathway in Solid Tumors*, pp. 81–128, Springer, 2011.

[92] R Fodde, J Kuipers, C Rosenberg, R Smits, M Kielman, C Gaspar, J H V Es, C Breukel, J Wiegant, R H Giles, and H Clevers, "Mutations in the APC tumour suppressor gene cause chromosomal instability," *Nat. Cell Biol.*, vol. 3, pp. 433–438, 2001.

[93] P Polakis, "Wnt signaling and cancer," *Bioinformatics*, vol. 14, pp. 1837–1851, 2000.

[94] B T MacDonald, K Tamai, and X He, "Wnt/β-Catenin signaling: Components, Mechanisms, and Diseases," *Dev. Cell*, vol. 17, no. 1, pp. 9–26, 2009.

[95] A Klaus and W Birchmeier, "Wnt signalling and its impact on development and cancer," *Nat. Rev. Cancer.*, vol. 8, no. 5, pp. 387–398, 2008.

[96] J Luo, J Chen, Z L Deng, X Luo, W X Song, K A Sharff, N Tang, R C Haydon, H H Luu, and T C He, "Wnt signaling and human diseases: what are the therapeutic implications?," *Lab. Invest.*, vol. 87, no. 2, pp. 97–103, 2007.

[97] M Tennis, M V Scoyk, and R A Winn, "Role of the Wnt Signaling Pathway and Lung Cancer," *J. Thorac. Oncol.*, vol. 2, no. 10, pp. 889–892, 2007.

[98] A Gregorieff and H Clevers, "Wnt signaling in the intestinal epithelium: from endoderm to cancer," *Genes and Development*, vol. 19, pp. 877–890, 2005.

[99] J H Es, N Barker, and H Clevers, "You Wnt some, you lose some: oncogenes in the Wnt signaling pathway," *Current Opinion in Genetics and Development*, vol. 13, pp. 28–33, 2003.

[100] N R Gough, "Understanding Wnt's Role in Osteoarthritis," *Sci. Signal.*, vol. 4, no. 172, pp. ec134, 2011.

[101] S Niemann, C Zhao, F Pascu, U Stahl, U Aulepp, L Niswander, J L Weber, and U Muller, "Homozygous wNT3 Mutation Causes Tetra-Amelia in a Large Consanguineous Family," *Am. J. Hum. Genet.*, vol. 74, no. 3, pp. 558–563, 2004.

[102] F S Rabelo, L M D Mota, R A Lima, F A Lima, G B Barra, J F D Carvalho, and A A Amato, "The wnt signaling pathway and rheumatoid arthritis," *Autoimmun. Rev.*, vol. 9, no. 4, pp. 207–210, 2010.

[103] C M Laine, B D Chung, M Susic, T Prescott, O Semler, T Fiskerstrand, P DEufemia, M Castori, M Pekkinen, E Sochett, W G Cole, C Netzer, and O Makitie, "Novel mutations affecting LRP5 splicing in patients with osteoporosis-pseudoglioma syndrome (OPPG)," *Eur. J. Hum. Genet.*, vol. 19, no. 8, pp. 875–881, 2011.

[104] Y Gong, R B Slee, N Fukai, G Rawadi, S Roman-Roman, A M Reginato, H Wang, T Cundy, F H Glorieux, D Lev, M Zacharin, K Oexle, J Marcelino, W Suwairi, S Heeger, G Sabatakos, S Apte, W N Adkins, J Allgrove, M Arslan-Kirchner, J A Batch, P Beighton, G C M Black, R G Boles, L M Boon, C Borrone, H G Brunner, G F Carle, B Dallapiccola, A D Paepe, B Floege, M L Halfhide, B Hall, R C Hennekam, T Hirose, A Jans, H Juppner, C A Kim, K Keppler-Noreuil, A Kohlschuetter, D LaCombe, M Lambert, E Lemyre, T Letteboer, L Peltonen, R S Ramesar, M Romanengo, H Somer, E Steichen-Gersdorf, B Steinmann, B Sullivan, A Superti-Furga, W Swoboda, M V D Boogaard, W V Hul, M Vikkula, M Votruba, B Zabel, T Garcia, R Baron, B R Olsen, and M L Warman, "LDL Receptor-Related Protein 5 (LRP5) Affects Bone Accrual and Eye Development," *Cell*, vol. 107, pp. 513–524, 2001.

[105] C Toomes, H M Bottomley, R M Jackson, K V Towns, S Scott, D A Mackey, J E Craig, L Jiang, Z Yang, R Trembath, G Woodruff, C Y Gregory-Evans, K Gregory-Evans, M J Parker, G C Black, L M Downey, K Zhang, and C F Inglehearn, "Mutations in LRP5 or FZD4 Underlie the common Familial Exudative Vitreoretinopathy Locus on

Chromosome 11q," *Am. J. Hum. Genet.*, vol. 74, no. 4, pp. 721–730, 2004.

[106] E Matalova, J Fleischmannova, P T Sharpe, and A S Tucker, "Tooth Agenesis: from Molecular Genetics to Molecular Dentistry," *J. Dent. Res.*, vol. 87, no. 7, pp. 617–623, 2008.

[107] L Lammi, S Arte, M Somer, H Jarvinen, P Lahermo, I Thesleff, S Pirinen, and P Nieminen, "Mutations in AXIN2 Cause Familial Tooth Agenesis and Predispose to Colorectal Cancer," *Am. J. Hum. Genet.*, vol. 74, no. 5, pp. 1043–1050, 2004.

[108] M V Belzen, O Bartsch, D Lacombe, D J Peters, and R C Hennekam, "Rubinstein-Taybi syndrome (CREBBP, EP300)," *Eur. J. Hum. Genet.*, vol. 19, no. 1, pp. 118–120, 2011.

[109] B Frank, M Hoffmeister, N Klopp, T Illig, J Chang-Claude, and H Brenner, "Single nucleotide polymorphisms in Wnt signaling and cell death pathway genes and susceptibility to colorectal cancer," *Carcinogenesis*, vol. 31, no. 8, pp. 1381–1386, 2010.

[110] D Ciznadija, R Tothill, M L Waterman, L Zhao, D Huynh, R M Yu, M Ernst, S Ishii, T Mantamadiotis, T J Gonda, R G Ramsay, and J Malaterre, "Intestinal adenoma formation and MYC activation are regulated by cooperation between MYB and Wnt signaling," *Cell Death. Differ.*, vol. 16, no. 11, pp. 1530–1538, 2009.

[111] C Y Logan and R Nusse, "The Wnt signaling pathway in development and disease," *Annual Review of Cell and Developmental Biology*, vol. 20, pp. 781–810, 2004.

[112] T N Martic, N Pecina-Slaus, V Kusec, T Kokotovic, H Musinovic, D Tomas, and M Zeljko, "Changes of AXIN-1 and Beta-Catenin in Neuroepithelial Brain Tumors," *Pathol. Oncol. Res.*, vol. 16, no. 1, pp. 75–79, 2010.

[113] M C Curia, M Zuckermann, L D Lellis, T Catalano, R Lattanzio, G Aceto, S Veschi, A Cama, J B Otte, M Piantelli, R Mariani-Costantini, F Cetta, and P Battista, "Sporadic childhood hepatoblastomas show activation of β-catenin, mismatch repair defects and p53 mutations," *Mod. Pathol.*, vol. 21, no. 1, pp. 7–14, 2008.

[114] T Jin, "The WNT signalling pathway and diabetes mellitus," *Diabetologia*, vol. 51, no. 10, pp. 1771–1780, 2008.

[115] S H Lee, C Demeterco, I Geron, A Abrahamsson, F Levine, and P Itkin-Ansari, "Islet Specific Wnt Activation in Human Type II Dsiabetes," *Exp. Diabetes Res.*, vol. 2008, no. 728763, 2008.

[116] K Pulkkinen, S Murugan, and S Vainio, "Wnt signaling in kidney development and disease," *Organogenesis*, vol. 4, no. 2, pp. 55–59, 2008.

[117] L Tickenbrock, S Hehn, B Sargin, G Evers, P R Ng, C Choudhary, W E Berdel, C Muller-Tidow, and H Serve, "Activation of Wnt signaling in cKit-ITD mediated transformation and imatinib sensitivity in acute myeloid leukemia," *Int. J. Hematol.*, vol. 88, no. 2, pp. 174–180, 2008.

[118] S A Ugur and A Tolun, "Homozygous WNT10b mutation and complex inheritance in Split-Hand/Foot Malformation," *Hum. Mol. Genet.*, vol. 17, no. 17, pp. 2644–2653, 2008.

[119] L Adaimy, E Chouery, H Megarbane, S Mroueh, V Delague, E Nicolas, H Belguith, P D Mazancourt, and A Megarbane, "Mutation in WNT10A is Associated with an Autosomal Recessive Ectodermal Dysplasia: The Odonto-onycho-dermal Dysplasia," *Am. J. Hum. Genet.*, vol. 81, no. 4, pp. 821–828, 2007.

[120] E Bowley, D B OGorman, and B S Gan, "β-Catenin Signaling in Fibroproliferative Disease," *J. Surg. Res.*, vol. 138, no. 1, pp. 141–150, 2007.

[121] A Mani, J Radhakrishnan, H Wang, A Mani, M A Mani, C Nelson-Williams, K S Carew, S Mane, H Najmabadi, D Wu, and R P Lifton, "LRP6 mutation in a Family with Early Coronary Disease and Metabolic Risk Factors," *Science*, vol. 315, no. 5816, pp. 1278–1282, 2007.

[122] M D Thompson and S P Monga, "WNT/β-Catenin Signaling in Liver Health and Disease," *Hepatology*, vol. 45, no. 5, pp. 1298–1305, 2007.

[123] C Christodoulides, A Scarda, M Granzotto, G Milan, E D Nora, J Keogh, G D Pergola, H Stirling, N Pannacciulli, J K Sethi, G Federspil, A Vidal-Puig, I S Farooqi, S O'Rahilly, and R Vettor, "WNT10B mutations in human obesity," *Diabetologia*, vol. 49, no. 4, pp. 678–684, 2006.

[124] J Reischl, S Schwenke, J M Beekman, U Mrowietz, S Sturzebecher, and J F Heubach, "Increased Expression of Wnt5a in Psoriatic Plaques," *J. Invest. Dermatol.*, vol. 127, no. 1, pp. 163–169, 2007.

[125] C G Woods, S Stricker, P Seemann, R Stern, J Cox, E Sherridan, E Roberts, K Springell, S Scott, G Karbani, S M Sharif, C Toomes, J Bond, D Kumar, L Al-Gazali, and S Mundlos, "Mutations in WNT7A Cause a Range of Limb Malformations, Including Fuhrmann syndrome and Al-Awadi/Raas-Rothschild/Schinzel Phocomelia Syndrome," *Am. J. Hum. Genet.*, vol. 79, no. 2, pp. 402–408, 2006.

[126] M A Koay and M A Brown, "Genetic disorders of the LRP5-Wnt signalling pathway affecting the skeleton," *Trends in Molecular Medicine*, vol. 11, pp. 129–137, 2005.

[127] R Levasseur, D Lacombe, and M C Vernejoul, "LRP5 mutations in osteoporosis-pseudoglioma syndrome and high-bone-mass disorders," *Joint Bone Spine*, vol. 72, pp. 207–214, 2005.

[128] S Mangioni, P Vigano, D Lattuada, A Abbiati, M Vignali, and A M D Blasio, "Overexpression of the Wnt5b Gene in Leiomyoma Cells: Implications for a Role of the Wnt Signaling Pathway in the Uterine Benign

Tumor," *J. Clin. Endocrinol. Metab.*, vol. 90, no. 9, pp. 5349–5355, 2005.

[129] S Nagayama, C Fukukawa, T Katagiri, T Okamoto, T Aoyama, N Oyaizu, M Imamura, J Toguchida, and Y Nakamura, "Therapeutic potential of antibodies against FZD 10, a cell-surface protein, for synovial sarcomas," *Oncogene*, vol. 24, no. 41, pp. 6201–6212, 2005.

[130] M L Johnson, G Gong, W Kimberling, S M Recker, D B Kimmel, and R R Recker, "Linkage of a Gene Causing High Bone Mass to Human Chromosome 11 (11q12-13)," *Am. J. Hum. Genet.*, vol. 60, pp. 1326–1332, 1997.

[131] L V Wesenbeeck, E Cleiren, J Gram, R K Beals, O Benichou, D Scopelliti, L Key, T Renton, C Bartels, Y Gong, M L Warman, M Vernejoul, J Bollerslev, and W V Hul, "Six Novel Missense Mutations in the LDL Receptor-Related Protein 5 (LRP5) Gene in Different Conditions with an Increased Bone Density," *Am. J. Hum. Genet.*, vol. 72, pp. 763–771, 2003.

[132] V Church, T Nohno, C Linker, C Marcelle, and P Francis-West, "Wnt regulation of chondrocyte differentiation," *J. Cell Sci.*, vol. 115, pp. 4809–4818, 2002.

[133] N Kozlovsky, R H Belmaker, and G Agam, "GSK-3 and the neurodevelopmental hypothesis of schizophrenia," *Eur. Neuropsychopharmacol.*, vol. 12, no. 1, pp. 13–25, 2002.

[134] T Miyaoka, H Seno, and H Ishino, "Increased expression of Wnt-1 in schizophrenic brains," *Schizophr. Res.*, vol. 38, no. 1, pp. 1–6, 1999.

[135] N J Szerlip, A Pedraza, D Chakravarty, M Azim, J McGuire, Y Fang, T Ozawa, E C Holland, J T Huse, S Jhanwar, M A Leversha, T Mikkelsen, and C W Brennan, "Intratumoral heterogeneity of receptor tyrosine kinases EGFR and PDGFRA amplification in glioblastoma defines subpopulations with distinct growth factor response," *Proc. Natl. Acad. Sci., USA*, vol. 109, no. 8, pp. 3041–3046, 2012.

[136] C Lopez-Gines, R Gil-Benso, R Ferrer-Luna, R Benito, E Serna, J Gonzalez-Darder, V Quilis, D Monleon, B Celda, and M Cerda-Nicolas, "New pattern of EGFR amplification in glioblastoma and the relationship of gene copy number with gene expression profile," *Modern Pathology*, vol. 23, pp. 856–865, 2010.

[137] S Britsch, *Advances in Anatomy, Embryology and Cell Biology*, pp. 1–65, Springer Verlag, 2007.

[138] N E Hynes and G MacDonald, "ErbB receptors and signaling pathways in cancer," *Current Opinion in Cell Biology*, vol. 21, pp. 177–184, 2009.

[139] N E Hynes and H A Lane, "ErbB receptors and signaling pathways in cancer," *Nat. Rev. Cancer*, vol. 5, pp. 341–354, 2005.

[140] M E Fortini, "Notch Signaling: The Core Pathway and Its Posttranslational Regulation," *Dev. Cell*, vol. 16, pp. 633–647, 2009.

[141] F M Watt, S Estrach, and C A Ambler, "Epidermal Notch signalling: differentiation, cancer and adhesion," *Current Opinion in Cell Biology*, vol. 20, pp. 171–179, 2008.

[142] F Radtke, F Schweisguth, and W Pear, "The Notch gospel Workshop on Notch Signalling in Development and Cancer," *EMBO reports*, vol. 6, no. 12, pp. 1120–1125, 2005.

[143] S Stylianou, R B Clarke, and K Brennan, "Aberrant activation of notch signaling in human breast cancer," *Cancer Res.*, vol. 66, pp. 1517–1525, 2006.

[144] V Bolos, J Grego-Bessa, and J L Pompa, "Notch Signaling in Development and Cancer," *Endocrine Reviews*, vol. 28, no. 3, pp. 339–363, 2007.

[145] T Gridley, "Notch signaling and inherited disease syndromes," *Human Molecular Genetics*, vol. 12, pp. R9–R13, 2003.

[146] A C Lin, B L Seeto, J M Bartoszko, M A Khoury, H Whetstone, L Ho, C Hsu, A S Ali, and B A Alman, "Modulating hedgehog signaling can attenuate the severity of osteoarthritis," *Nat. Med.*, vol. 15, no. 12, pp. 1421–1426, 2009.

[147] R R Singh, J H Cho-Vega, Y Davuluri, S Ma, F Kasbidi, C Milito, P A Lennon, E Drakos, L J Medeiros, R Luthra, and F Vega, "Sonic Hedgehog Signaling Pathway Is Activated in ALK-Positive Anaplastic Large Cell Lymphoma," *Cancer Res.*, vol. 69, no. 6, pp. 2550–2558, 2009.

[148] C Zhao, A Chen, C H Jamieson, M Fereshteh, A Abrahamsson, J Blum, H Y Kwon, J Kim, J P Chute, D Rizzieri, M Munchhof, T VanArsdale, P A Beachy, and T Reya, "Hedgehog signalling is essential for maintenance of cancer stem cells in myeloid leukaemia," *Nature*, vol. 458, no. 9, pp. 776–779, 2009.

[149] E H Epstein, "Basal cell carcinomas: attack of the hedgehog," *Nat. Rev. Cancer*, vol. 8, pp. 743–754, 2008.

[150] P A Beachy, S S Karhadkar, and D M Berman, "Tissue repair and stem cell renewal in carcinogenesis," *Nature*, vol. 432, pp. 324–331, 2004.

[151] A F Baas, J Medic, R Slot, C G Kovel, A Zhernakova, R H Geelkerken, S E Kranendonk, S M Sterkenburg, D E Grobbee, A P Boll, C Wijmenga, J D Blankensteijn, and Y M Ruigrok, "Association of the tgf-b receptor genes with abdominal aortic aneurysm," *European Journal of Human Genetics*, vol. 18, pp. 240–244, 20010.

[152] N Garcia-Fernandez and M L M Molina, "TGF-β Made Easy," *The Open Urology & Nephrology Journal*, vol. 2, pp. 1–5, 2009.

[153] S Giampieri, C Manning, S Hooper, L Jones, C S Hill, and E Sahai, "Localized and reversible TGFβ signalling switches breast cancer cells from cohesive to single cell motility," *Nat. Cell Biol.*, vol. 11, no. 11, pp. 1287–1296, 2009.

[154] S Zhou, P Buckhaults, L Zawel, F Bunz, G Riggins, J L Dai, S E Kern, K W Kinzler, and B Vogelstein, "Targeted deletion of Smad4 shows it is required for transforming growth factor β and activin signaling in colorectal cancer cells," *Proc. Natl. Acad. Sci., USA*, vol. 95, pp. 2412–2416, 1998.

[155] S Bornstein, R White, S Malkoski, M Oka, G Han, T Cleaver, D Reh, P Andersen, N Gross, S Olson, C Deng, S Lu, and X Wang, "Smad4 loss in mice causes spontaneous head and neck cancer with increased genomic instability and inflammation," *J. Clin. Invest.*, vol. 119, no. 11, pp. 3408–3419, 2009.

[156] M Korc, "Smad4: gatekeeper gene in head and neck squamous cell carcinoma," *J. Clin. Invest.*, vol. 119, no. 11, pp. 3208–3212, 2009.

[157] K J Gordon and G C Blobe, "Role of transforming growth factor-beta superfamily signaling pathways in human disease," *Biochem. Biophys. Acta*, vol. 1782, pp. 197–228, 2008.

[158] C A Bertuccio, "Relevance of VEGF and Nephrin Expression in Glomerular Diseases," *Journal of Signal Transduction*, vol. 2011, no. 718609, 2011.

[159] M Shibuya, "Differential Roles of Vascular Endothelial Growth Factor Receptor-1 and Receptor-2 in Angiogenesis," *Journal of Biochemistry and Molecular Biology*, vol. 39, no. 5, pp. 469–478, 2006.

[160] D C Rajalakshmi, A R K Gopalakrishnan, and C C Kartha, *Signal Transduction in the Cardiovascular System in Health and Disease*, pp. 301–326, Springerlink Publishers, 2008.

[161] M B Marrero, A K Banes-Berceli, D M Stern, and D C Eaton, "Role of the jAK/STAT signaling pathway in diabetic nephropathy," *Am. J. Physiol. Renal Physiol.*, vol. 290, pp. F762–F768, 2006.

[162] A B Pernis and P B Rothman, "JAK-STAT signaling in asthma," *The Journal of Clinical Investigation*, vol. 109, no. 10, pp. 1279–1283, 2002.

[163] E Mascareno, M El-Shafei, N Maulik, M Sato, Y Guo, D K Das, and M A Q Siddiqui, "JAK/STAT Signaling Is Associated With Cardiac Dysfunction During Ischemia and Reperfusion," *Circulation*, vol. 104, pp. 325–329, 2001.

[164] A C Ward, I Touw, and A Yoshimura, "The Jak-Stat pathway in normal and perturbed hematopoiesis," *Blood*, vol. 95, no. 1, pp. 19–29, 2000.

[165] D Nishimura, "A view from the Web: BioCarta," *Biotech Software & Internet Report*, vol. 2, no. 3, pp. 117–120, 2001.

[166] S Yamamoto, N Sakai, H Nakamura, H Fukagawa, K Fukuda, and T Takagi, "INOH: ontology-based highly structured database of signal transduction pathways," *Database (Oxford)*, vol. 2011, no. bar052, 2011.

[167] M Kanehisa and S Goto, "KEGG: Kyoto Encyclopedia of Genes and Genomes," *Nucleic Acids Research*, vol. 28, pp. 27–30, 2000.

[168] K Kandasamy, S S Mohan, R Raju, S Keerthikumar, G S Kumar, A K Venugopal, D Telikicherla, J D Navarro, S Mathivanan, C Pecquet, S K Gollapudi, S G Tattikota, S Mohan, H Padhukasahasram, Y Subbannayya, R Goel, H K Jacob, J Zhong, R Sekhar, V Nanjappa, L Balakrishnan, R Subbaiah, Y L Ramachandra, B A Rahiman, T S Prasad, J X Lin, J C Houtman, S Desiderio, J C Renauld, S N Constantinescu, O Ohara, T Hirano, M Kubo, S Singh, P Khatri, S Draghici, G D Bader, C Sander, W J Leonard, and A Pandey, "NetPath: a public resource of curated signal transduction pathways," *Genome Biol.*, vol. 11, no. 1, pp. R3, 2010.

[169] R Raju, V Nanjappa, L Balakrishnan, A Radhakrishnan, J K Thomas, J Sharma, M Tian, S M Palapetta, T Subbannayya, N R Sekhar, B Muthusamy, R Goel, Y Subbannayya, D Telikicherla, M Bhattacharjee, S M Pinto, N Syed, M S Srikanth, G J Sathe, S Ahmad, S N Chavan, G S Kumar, A Marimuthu, T S Prasad, H C Harsha,

B A Rahiman, O Ohara, G D Bader, S S Mohan, W P Schiemann, and A Pandey, "NetSlim: high-confidence curated signaling maps," *Database (Oxford)*, vol. 2011, no. bar032, 2011.

[170] C F Schaefer, K Anthony, S Krupa, J Buchoff, M Day, T Hannay, and K H Buetow, "PID: the Pathway Interaction Database," *Nucleic Acids Res.*, vol. 37, pp. D674–D679, 2009.

[171] A Dilek, M E Belviranli, and U Dogrusoz, "VISIBIOweb: visualization and layout services for BioPAX pathway models," *Nucleic Acids Research*, vol. 38, pp. W150–W154, 2010.

[172] L Matthews, G Gopinath, M Gillespie, M Caudy, D Croft, B D Bono, P Garapati, J Hemish, H Hermjakob, B Jassal, A Kanapin, S Lewis, S Mahajan, B May, E Schmidt, I Vastrik, G Wu, E Birney, L Stein, and P DEustachio, "Reactome knowledgebase of human biological pathways and processes," *Nucleic Acids Res.*, vol. 37, pp. D619–D622, 2009.

[173] G Joshi-Tope, M Gillespie, I Vastrik, P DEustachio, E Schmidt, B D Bono, B Jassal, G R Gopinath, G R Wu, L Matthews, S Lewis, E Birney, and L Stein, "Reactome: a knowledgebase of biological pathways," *Nucleic Acids Research*, vol. 33, pp. D428–D432, 2005.

[174] N R Gough, "Science's signal transduction knowledge environment: the connections maps database," *Ann. N. Y. Acad. Sci.*, vol. 971, pp. 585–587, 2002.

[175] K I Fukuda and T Takagi, "Knowledge Representation of Signal Transduction Pathways," *Bioinformatics*, vol. 17, pp. 8290–837, 2001.

[176] L D F Costa, "The Hierarchical Backbone of Complex Networks," *Physical Review Letters*, vol. 93, no. 098702, 2004.

[177] G Karypis, E H Han, and V Kumar, "Multilevel refinement for hierarchical clustering," Tech. Rep. TR-99-020, Department of Computer Science, University of Minnesota, Minneapolis, 1999.

[178] E Ravasz and A Barabasi, "Hierarchical organization in complex networks," *Physical Review E*, vol. 67, 2003.

[179] U Elsner, "Graph Partitioning: A survey," Tech. Rep. 97-27, Technische Universitat Chemnitz, Chemnitz, Germany, 1997.

[180] P O Fjallstrom, "Algorithms for Graph Partitioning: A Survey," Tech. Rep. 98-010, Linkoping University, Sweden, Electronic Articles in Computer and Information Science, 1998.

[181] W W Larue, E Komp, S Schaffer, V S Frost, K S Shanmugan, and D Reznik, "A block oriented paradigm for modeling communication networks," in *Conference Record, 'A New Era', IEEE*, 1990, pp. 689–695.

[182] M Ederer, T Sauter, E Bullinger, E D Gilles, and F Allgwer, "An approach for dividing models of biological reaction networks into functional units," *SIMULATION*, vol. 79, no. 12, pp. 703–716, 2003.

[183] R Guimera and L A N Amaral, "Functional cartography of complex metabolic networks," *Nature*, vol. 433, pp. 895–900, 2005.

[184] S Zhao and S Li, "A co-module approach for elucidating drug-disease associations and revealing their molecular basis," *Bioinformatics*, vol. 28, no. 7, pp. 955–961, 2012.

[185] S Dutt, "New Faster Kernighan-Lin-Type Graph-Partitioning Algorithms," in *Proc. of Intl. Conf. Computer-Aided Design*, 1993, pp. 370–377.

[186] M E J Newman, "Finding community structure in networks using the eigenvectors of matrices," *Physical Review E*, vol. 74, 2006.

[187] M E J Newman, "Modularity and community structure in networks," in *Proc. Natl. Acad. Sci., USA*, 2006, vol. 103, pp. 8577–8582.

[188] M E J Newman, "Detecting community structure in networks," *The European Physics Journal B*, vol. 38, pp. 321–330, 2004.

[189] J Zhu, H Xiao, X Shen, J Wang, J Zou, L Zhang, D Yang, W Ma, C Yao, X Gong, M Zhang, Y Zhang, and Z Guo, "Viewing cancer genes from co-evolving gene modules," *Bioinformatics*, vol. 26, no. 7, pp. 919–924, 2010.

[190] M J Cowley, M Pinese, K S Kassahn, N Waddell, J V Pearson, S M Grimmond, A V Biankin, S Hautaniemi, and J Wu, "PINA v2.0: mining interactome modules," *Nucleic Acids Res.*, vol. 40, pp. D862–D865, 2012.

[191] R Yang, B J J Daigle, L R Petzold, and F J Doyle-III, "Core module biomarker identification with network exploration for breast cancer metastasis," *BMC Bioinformatics*, vol. 13, no. 12, 2012.

[192] L Hua, H Lin, D Li, L Li, and Z Liu, "Mining Functional Gene Modules Linked with Rheumatoid Arthritis Using a SNP-SNP Network," *Genomics Proteomics Bioinformatics*, vol. 10, no. 1, pp. 23–34, 2012.

[193] G V Sridharan, S Hassoun, and K Lee, "Identification of biochemical network modules based on shortest retroactive distances," *PLoS Comput Biol.*, vol. 7, no. 11, pp. e1002262, 2011.

[194] J H Cho, K Wang, and D J Galas, "An integrative approach to inferring biologically meaningful gene modules," *BMC Syst. Biol.*, vol. 5, no. 117, 2011.

[195] J T Hsu, C H Peng, W P Hsieh, C Y Lan, and C Y Tang, "A novel method to identify cooperative functional modules: study of module coordination in the *Saccharomyces cerevisiae* cell cycle," *BMC Bioinformatics*, vol. 12, no. 281, 2011.

[196] V Jayaswal, M Lutherborrow, D D Ma, and Y H Yang, "Identification of microRNA-mRNA modules using microarray data," *BMC Genomics*, vol. 12, no. 138, 2011.

[197] P Zarrineh, A C Fierro, A Sanchez-Rodriguez, B D Moor, K Engelen, and K Marchal, "COMODO: an adaptive coclustering strategy to

identify conserved coexpression modules between organisms," *Nucleic Acids Res.*, vol. 39, no. 7, pp. e41, 2011.

[198] S Sun, X Dong, Y Fu, and W Tian, "An iterative network partition algorithm for accurate identification of dense network modules," *Nucleic Acids Research*, vol. 40, no. 3, pp. e18, 2012.

[199] Q Yu, G Li, and J Huang, "MOfinder: A Novel Algorithm for Detecting Overlapping Modules from Protein-Protein Interaction Network," *Journal of Biomedicine and Biotechnology*, vol. 2012, no. 103702, 2012.

[200] W Hendrix, A M Rocha, K Padmanabhan, A Choudhary, K Scott, J R Mihelcic, and N F Samatova, "DENSE: efficient and prior knowledge-driven discovery of phenotype-associated protein functional modules," *BMC Syst. Biol.*, vol. 5, no. 172, 2011.

[201] P Jia, S Zheng, J Long, W Zheng, and Z Zhao, "dmGWAS: dense module searching for genome-wide association studies in protein-protein interaction networks," *Bioinformatics*, vol. 27, no. 1, pp. 95–102, 2011.

[202] S G Konietzny, L Dietz, and A C McHardy, "Inferring functional modules of protein families with probabilistic topic models," *BMC Bioinformatics*, vol. 12, no. 141, 2011.

[203] Q J Jiao, Y K Zhang, L N Li, and H B Shen, "BinTree seeking: A Novel Approach to Mine Both Bi-Sparse and Cohesive Modules in Protein Interaction Networks," *PLoS One*, vol. 6, no. 11, pp. e27646, 2011.

[204] C J Augeri and H H Ali, "New Graph-Based Algorithms for Partitioning VLSI Circuits," in *Proc. of Circuits and Systems (ISCAS '04)*, 2004, vol. 4, pp. 521–524.

[205] C H Chen, "Graph partitioning for concurrent test scheduling in VLSI circuit," in *Proc. of 28th ACM/IEEE Design Automation Conference*, 1991, vol. 18.4, pp. 287–290.

[206] L P Cordella, P Foggia, C Sansone, and M Vento, "Fast Graph Matching for detecting CAD image components," in *Proc. of 15th International Conference on Pattern Recognition*. IEEE, 2000, vol. 2, pp. 1034–1037.

[207] A Pothen, *Parallel Numerical Algorithms*, chapter Graph partitioning algorithms with applications to scientific computing, pp. 323–368, Kluwer Academic Press, 1997.

[208] M W Berry, B Henderickson, and P Raghavan, *The Mathematics of Numerical Analysis*, vol. 32 of *Lectures in Applied Mathematics*, chapter Sparse matrix reordering schemes for browsing hypertext, pp. 99–123, American Mathematical Society, 1996.

[209] G Karypis and V Kumar, "Parallel Multilevel Graph Partitioning," in *Proc. of the 10th International Parallel Processing Symposium (IPPS '96)*, 1996, pp. 314–319.

[210] G Kedem and H Watanabe, "Graph optimization techniques for IC layout and compaction," in *Proc. of the 20th conference on Design automation*. IEEE, 1983, pp. 113–120.

[211] B Hendrickson and R Leland, "The Chaco User's Guide Version 2," Tech. Rep. SAND94-2692, Sandia National Laboratories, Albuquerque, NM, 1994.

[212] M Girvan and M E J Newman, "Community structure in social and biological networks," in *Proc. Natl. Acad. Sci., USA*, 2002, vol. 99, pp. 7821–7826.

[213] S Fortunato, V Latora, and M Marchiori, "A method to Find community structures based on information centrality," *Physical Review E*, vol. 70, no. 056104, 2004.

[214] N Imafuji and M Kitsuregawa, "Effects of Maximum Flow Algorithm on Identifying Web Community," in *Proc. of the 4th international*

workshop on Web information and data management (WIDM '02), 2002, pp. 43–48.

[215] S Yang, "Exploring complex networks by walking on them," *Physical Review E*, vol. 71, no. 016107, 2005.

[216] M Latapy and P Pons, "Computing communities in large networks using random walks," in *Proceedings of the 20th International Symposium on Computer and Information Sciences (ISCIS'05)*, 2005, vol. 3733, pp. 284–293.

[217] M E J Newman, "Fast algorithm for detecting community structure in networks," *Physical Review E*, vol. 69:066133, 2004.

[218] P Holme, M Huss, and H Jeong, "Subnetwork hierarchies of biochemical pathways," *Bioinformatics*, vol. 19, no. 4, pp. 532–538, 2003.

[219] H Jeong, B Tombor, R Albert, Z N Oltvai, and A L Barabasi, "The large-scale organization of metabolic networks," *Nature*, vol. 407, no. 6804, pp. 545–658, 2000.

[220] H Hu, X Yan, Y Huang, J Han, and X J Zhou, "Mining coherent dense subgraphs across massive biological networks for functional discover," *Bioinformatics*, vol. 21, no. (Suppl. 1), pp. i213–i221, 2005.

[221] S M Patra and S Vishveshwara, "Backbone cluster identification in proteins by a graph theoretical method," *Biophysical Chemistry*, vol. 84, no. 1, pp. 13–25, 2000.

[222] S Schuster, T Pfeiffer, F Moldenhauer, I Koch, and T Dandeke, "Exploring the pathway structure of metabolism: decomposition into subnetworks and application to *Mycoplasma pneumonia*," *Bioinformatics*, vol. 18, no. 2, pp. 351–361, 2002.

[223] A Wagner and D A Fell, "Small world inside large metabolic networks," in *Proc. Royal Society B*, 2001, vol. 268, pp. 1803–1810.

[224] M E J Newman and M Girvan, "Finding and evaluating community structure in networks," *Phys. Rev. E*, vol. 69, no. 026113, 2004.

[225] D M Wilkinson and B A Huberman, "A method for finding communities of related genes," in *Proc. Natl. Acad. Sci., USA*, 2004, vol. 101, pp. 5241–5248.

[226] M E J Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45, no. 2, pp. 167–256, 2003.

[227] S Maslov and K Sneppen, "Specificity and stability in topology of protein networks," *Science*, vol. 296, no. 5569, pp. 910–913, 2002.

[228] J Stelling, S Klamt, K Bettenbrock, S Schuster, and E D Gilles, "Metabolic network structure determines key aspects of functionality and regulation," *Nature*, vol. 420, no. 6912, pp. 190–193, 2002.

[229] N Guelzim, S Bottani, P Bourgine, and F Kepes, "Topological and causal structure of the yeast transcriptional regulatory network," *Nature genetics*, vol. 31, no. 1, pp. 60–63, 2002.

[230] D J Raine and V Norris, "Network structure of metabolic pathways," *Journal of Biological Physics and Chemistry*, vol. 1, no. 2, pp. 89–94, 2001.

[231] R Milo, S Shen-Orr, S Itzkovitz, N Kashtan, D Chklovskii, and U Alon, "Network motifs: Simple building blocks of complex networks," *Science*, vol. 298(5594), pp. 824–827, 2002.

[232] S S Shen-Orr, R Milo, S Mangan, and U Alon, "Network motifs in the transcriptional regulation network of escherichia coli," *Nature genetics*, vol. 31, no. 1, pp. 64–68, 2002.

[233] E Segal, M Shapira, A Regev, D Peer, D Botstein, D Koller, and N Friedman, "Module networks: ideantifying regulatory modules and their condition-specific regulators from gene expression data," *Nat. Genet.*, vol. 34, pp. 166–176, 2003.

[234] K Macropol, T Can, and A K Singh, "RRW: repeated random walks on genome-scale protein networks for local cluster discovery," *BMC Bioinformatics*, vol. 10, pp. 283, 2009.

[235] W S Verwoerd, "A new computational method to split large biochemical networks into coherent subnets," *BMC Systems Biology*, vol. 5, no. 25, 2011.

[236] M Mete, F Tang, X Xu, and N Yuruk, "A structural approach for finding functional modules from large biological networks," *BMC Bioinformatics*, vol. 9(Suppl 9), no. S19, 2008.

[237] P H Lee and D Lee, "Modularized learning of genetic interaction networks from biological annotations and mRNA expression data," *Bioinformatics*, vol. 21, pp. 2739–2747, 2005.

[238] R L Chang, F Luo, S Johnson, and R H Scheuermann, "Deterministic Graph-Theoretic Algorithm for Detecting Modules in Biological Interaction Networks," *International Journal of Bioinformatics Research and Application*, vol. 6, no. 6, pp. 101–119, 2010.

[239] L Kaufman and P J Rousseeuw, *Finding groups in data. An introduction to cluster analysis*, Wiley, New York, 1990.

[240] Y Xiao and M R Segal, "Identification of yeast transcriptional regulation networks using multivariate random forests," *PLoS Comput Biol.*, vol. 5, no. 6, pp. e1000414, 2009.

[241] D Szklarczyk, A Franceschini, M Kuhn, M Simonovic, A Roth, P Minguez, T Doerks, M Stark, J Muller, P Bork, L J Jensen, and C V Mering, "The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored," *Nucleic Acids Res.*, vol. 39, pp. D561–D568, 2011.

[242] J A Papin, J L Reed, and B O Palsson, "Hierarchical thinking in network biology: the unbiased modularization of biochemical networks," *Trends in Biochemical Sciences*, vol. 29, no. 12, pp. 641–647, 2004.

[243] N J Eungdamrong and R Iyengar, "Modeling cell signaling networks," *Biol. Cell*, vol. 96, no. 5, pp. 355–362, 2004.

[244] R N Neves and R Iyengar, "Modeling of signaling networks," *BioEssays*, vol. 24, no. 12, pp. 1110–1117, 2002.

[245] J Saez-rodriguez, A Kremling, H Conzelmann, K Bettenbrock, and E D Gilles, "Modular analysis of signal transduction networks," *Control Systems Magazine*, vol. 24, no. 4, pp. 35–52, 2004.

[246] E Grafahrend-Belau, F Schreiber, M Heiner, A Sackmann, B H Junker, S Grunwald, A Speer, K Winder, and I Koch, "Modularization of biochemical networks based on classification of Petri net t-invariants," *BMC Bioinformatics*, vol. 9, no. 90, 2008.

[247] G Chartrand and O R Oellermann, *Applied and algorithmic graph theory*, McGraw Hill, New York, 1993.

[248] C Farhat, "A simple and effcient automatic FEM domain decomposer," *Computers and Structures*, vol. 28, pp. 579–602, 1988.

[249] B W Kernighan and S Lin, "An Efficient Heuristic Procedure for Partitioning Graphs," *The Bell System Technical Journal*, vol. 49, pp. 291–307, 1970.

[250] K I Goh, M E Cusick, D Valle, B Childs, M Vidal, and A L Barabasi, "The human disease network," *Proc. Natl. Acad. Sci., USA*, vol. 104, pp. 8685–8690, 2007.

[251] D S Lee, J Park, K A Kay, N A Christakis, Z N Oltvai, and A L Barabasi, "The implications of human metabolic network topology for disease comorbidity," *Proc. Natl. Acad. Sci., USA*, vol. 105, pp. 9880–9885, 2008.

[252] T G Boulton, G D Yancopoulos, J S Gregory, C Slaughter, C Moomaw, J Hsu, and M H Cobb, "An insulin-stimulated protein kinase similar to yeast kinases involved in cell cycle control," *Science*, vol. 249, no. 4964, pp. 64–67, 1990.

[253] T G Boulton, S H Nye, D J Robbins, N Y Ip, E Radziejewska, S D Morgenbesser, R A DePinho, N Panayotatos, M H Cobb, and G D Yancopoulos, "ERKs: a family of protein-serine/threonine kinases that are activated and tyrosine phosphorylated in response to insulin and NGF.," *Cell*, vol. 65, no. 4, pp. 663–675, 1991.

[254] N G Ahn and E G Krebs, "Evidence for an Epidermal Growth Factor-stimulated Protein Kinase Cascade in Swiss 3T3 cells. Activation of serine peptide kinase activity by myelin basic protein kinases in vitro," *J. Biol. Chem.*, vol. 265, no. 20, pp. 11495–11501, 1990.

[255] L B Ray and T W Sturgill, "Insulin-stimulated microtubule-associated protein kinase is phosphorylated on tyrosine and threonine in vivo," *Proc. Natl. Acad. Sci., USA*, vol. 85, no. 11, pp. 3753–3757, 1988.

[256] A J Rossomando, D M Payne, M J Weber, and T W Sturgill, "Evidence that pp42, a major tyrosine kinase target protein, is a mitogen-activated serine/threonine protein kinase," *Proc. Natl. Acad. Sci., USA*, vol. 86, no. 18, pp. 6940–6943, 1989.

[257] W Kolch, M Calder, and D Gilbert, "When kinases meet mathematics: the systems biology of MAPK signalling," *FEBS Letters*, vol. 579, pp. 1891–1895, 2005.

[258] G Taj, P Agarwal, M Grant, and A Kumar, "MAPK machinery in plants: recognition and response to different stresses through multiple signal transduction pathways," *Plant Signal Behav.*, vol. 5, no. 11, pp. 1370–1378, 2010.

[259] G L Johnson and R Lapadat, "Mitogen-Activated Protein Kinase Pathways Mediated by ERK, JNK, and p38 Protein Kinases," *Science*, vol. 298, no. 5600, pp. 1911–1912, 2002.

[260] C Huang, K Jacobson, and M D Schaller, "MAP kinases and cell migration," *J. Cell Sci.*, vol. 117, pp. 4619–4628, 2004.

[261] R E Chen and J Thorner, "Function and regulation in MAPK signaling pathways: lessons learned from the yeast *Saccharomyces cerevisiae*," *Biochim Biophys Acta.*, vol. 1773, no. 8, pp. 1311–1340, 2007.

[262] S Ringer, "A further contribution regarding the influence of the different constituents of the blood on the contraction of the heart," *J. Physiol.*, vol. 4, no. 1, pp. 29–42, 1883.

[263] H Streb, R F Irvine, M J Berridge, and I Schulz, "Release of Ca^{2+} from a nonmitochondrial intracellular store in pancreatic acinar cells by inositol-1,4,5-trisphosphate," *Nature*, vol. 306, no. 5938, pp. 67–69, 1983.

[264] K S Cuthbertson and P H Cobbold, "Phorbol ester and sperm activate mouse oocytes by inducing sustained oscillations in cell Ca^{2+}," *Nature*, vol. 316, no. 6028, pp. 541–542, 1985.

[265] L Combettes, G Dupont, and J B Parys, "New mechanisms and functions in Ca^{2+} signalling," *Biol. Cell*, vol. 96, no. 1, pp. 1–2, 2004.

[266] K G Baimbridge, M R Celio, and J H Rogers, "Calcium binding proteins in the nervous system," *Trends Neurosciences*, vol. 15, pp. 303–308, 1992.

[267] C W Heizmann and W Hunziker, "Intracellular calcium binding proteins: more sites than insights," *Trends in Biochemical Sciences*, vol. 16, pp. 98–103, 1991.

[268] D A Zacharias, S J Dalrymple, and E E Strehler, "Transcript distribution of plasma membrane ca^{2+} pump isoforms and splice variants in the human brain," *Molecular Brain Research*, vol. 28, pp. 263–272, 1995.

[269] A Bruce, J Alexander, L Julian, R Martin, R Keith, and W Peter, "Cell communication," in *Molecular Biology of The Cell*, pp. 831–906. Garland Science, New York, 4th edition, 2002.

[270] T Pozzan, R Rizzuto, P Volpe, and J Meldolesi, "Molecular and cellular physiology of intracellular calcium stores," *Physiol. Rev.*, vol. 74, pp. 595–636, 1994.

[271] M J Berridge, "Elementary and global aspects of calcium signaling," *J. Physiol.*, vol. 499, pp. 291–306, 1997.

[272] A J Golumbfskie, C J Mundy, A K Kubota, A Nichols, and A A Quong, "A three dimensional model of intercellular calcium signaling in epithelial cells," in *Proceedings of the 25th Annual International Conference of the IEEE EMBS: 17-21 December 2003, Cancun, Mexico*, 2003, pp. 2694–2697.

[273] A Levchenko, "Stochastic modeling of spatially localized events in protein kinase A and Ca^{2+} signaling," in *Proceedings of the Second Joint EMBS/BMES Conference: 23-26 October 2002, Houston, TX, USA*, 2002.

[274] W Echevarria, M F Leite, M T Guerra, W R Zipfel, and M H Nathanson, "Regulation of calcium signals in the nucleus by a nucleoplasmic reticulum," *Nat. Cell Biol.*, vol. 5, pp. 440–446, 2003.

[275] R W Tsien and R Y Tsien, "Calcium channels, stores and oscillations," *Ann. Rev. Cell Biol.*, vol. 6, pp. 715–760, 1990.

[276] A R Means, "Calcium, calmodulin and cell cycle regulation," *FEBS Lett.*, vol. 347, pp. 1–4, 1994.

[277] R Schreiber, "Ca^{2+} Signaling, Intracellular pH and Cell Volume in Cell Proliferation," *J. Membr. Biol.*, vol. 205, no. 3, pp. 129–137, 2005.

[278] H Bading, D D Ginty, and M E Greenberg, "Regulation of gene expression in hippocampal neurons by distinct calcium signaling pathways," *Science*, vol. 260, pp. 181–186, 1993.

[279] P A Negulescu, N Shastri, and M D Cahalan, "Intracellular calcium dependence of gene expression in single T lymphocytes," in *Proc. Natl. Acad. Sci., USA*, 1994, pp. 2873–2877.

[280] K Burns, B Duggan, E A Atkinson, K S Famulski, M Nemer, R C Bleackley, and M Michalak, "Modulation of gene expression by calreticulin binding to the glucocorticoid receptor," *Nature*, vol. 367, pp. 476–480, 1994.

[281] P Nicotera, B Zhivotovsky, and S Orrenius, "Nuclear calcium transport and the role of calciumin apoptosis," *Cell Calcium*, vol. 16, pp. 279–288, 1994.

[282] M Puceat and M Jaconi, "Ca^{2+} signalling in cardiogenesis," *Cell Calcium*, vol. 38, no. 3-4, pp. 383–389, 2005.

[283] M Brini and E Carafoli, "Calcium Pumps in Health and Disease," *Physiol. Rev.*, vol. 89, no. 4, pp. 1341–1378, 2009.

[284] F Rijsewijk, M Schuermann, E Wagenaar, P Parren, D Weigel, and R Nusse, "The Drosophila homology of the mouse mammary oncogene *int-1* is identical to the segment polarity gene *wingless*," *Cell*, vol. 50, pp. 649–657, 1987.

[285] R P Sharma, "Wingless - a new mutant in D. melanogaster," *Drosoph. Inf. Serv.*, vol. 50, no. 134, 1973.

[286] R P Sharma and V L Chopra, "Effect of the *Wingless* (wg^1) Mutation on Wing and Haltere Development in *Drosophila melanogaster*," *Developmental Biology*, vol. 48, pp. 461–465, 1976.

[287] R Nusse and H E Varmus, "Many tumors induced by the mouse mammary tumor virus contain a provirus integrated in the same region of the host genome," *Cell*, vol. 31, no. 1, pp. 99–109, 1982.

[288] R Nusse, A Ooyen, D Cox, Y K T Fung, and H Varmus, "Mode of proviral activation of a putative mammary oncogene (*int*-1) on mouse chromosome 15," *Nature*, vol. 307, pp. 131–136, 1984.

[289] C Nusslein-Volhard and E Wieschaus, "Mutations affecting segment number and polarity in *Drosophila*," *Nature*, vol. 287, pp. 795–801, 1980.

[290] J Wu and S M Cohen, "Repression of Teashirt marks the initiation of wing development," *Development*, vol. 129, pp. 2411–2418, 2002.

[291] J Huelsken and W Birchmeier, "New aspects of Wnt signaling pathways in higher vertebrates," *Current Opinion in Genetics and Development*, vol. 11, pp. 547–553, 2001.

[292] M Bienz, "Spindles cotton on to junctions, APC and EB1," *Nat. Cell Biol.*, vol. 3, pp. E67–E68, 2001.

[293] M Kuhl, L C Sheldahl, M Park, J R Miller, and R T Moon, "The Wnt/Ca^{2+} pathway: a new vertebrate Wnt signaling pathway takes shape," *Trends in Genetics*, vol. 16, pp. 279–283, 2000.

[294] C J Thorpe, A Schlesinger, and B Bowerman, "Wnt signalling in *Caenorhabditis elegans*: regulating repressors and polarizing the cytoskeleton," *Trends Cell Biol.*, vol. 10, pp. 10–17, 2000.

[295] J Huelsken and J Behrens, "The Wnt signalling pathway," *J. Cell Sci.*, vol. 115, pp. 3977–3978, 2002.

[296] M Montcouquiol, E B Crenshaw-III, and M W Kelley, "Noncanonical Wnt Signaling and Neural Polarity," *Annu. Rev. Neurosci.*, vol. 29, pp. 363–386, 2006.

[297] J Hulsken and J Behrens, "The Wnt signalling pathway," *J. Cell Sci.*, vol. 113, pp. 3545, 2000.

[298] J Behrens, J P V Kries, M Kuhl, L Bruhn, D Wedlich, R Grosschedl, and W Birchmeier, "Functional interaction of β-catenin with the transcription factor LEF-1," *Nature*, vol. 382, pp. 638–642, 1996.

[299] M Molenaar, M Wetering, M Oosterwegel, J Peterson-Maduro, S Godsave, V Korinek, J Roose, O Destree, and H Clevers, "XTcf-3 Transcription Factor Mediates β-Catenin-Induced Axis Formation in Xenopus Embryos," *Cell*, vol. 86, pp. 391–399, 1996.

[300] M Kuhl, L C Sheldahl, C C Malbon, and R T Moon, "Ca^{2+}/Calmodulin-dependent Protein Kinase IIIs Stimulated by Wnt and Frizzled Homologs and Promotes Ventral Cell Fates in *Xenopus*," *The Journal of Biological Chemistry*, vol. 275, pp. 12701–12711, 2000.

[301] R Bayly and J D Axelrod, "Pointing in the right direction: new developments in the field of planar cell polarity," *Nat. Rev. Genet.*, vol. 12, no. 6, pp. 385–391, 2011.

[302] J D Axelrod, "Progress and challenges in understanding planar cell polarity signaling," *Semin. Cell Dev. Biol.*, vol. 20, no. 8, pp. 964–971, 2009.

[303] M Simons and M Mlodzik, "Planar Cell Polarity Signaling: From Fly Development to Human Disease," *Annu. Rev. Genet.*, vol. 42, pp. 517–540, 2008.

[304] R T Moon and K Shah, "Developmental biology: Signalling polarity," *Nature*, vol. 417, pp. 239–240, 2002.

[305] K Willert, J D Brown, E Danenberg, A W Duncan, I L Weissman, T Reya, J R Yates-III, and R Nusse, "Wnt proteins are lipid-modified and can act as stem cell growth factors," *Nature*, vol. 423, pp. 448–452, 2003.

[306] Y Cui, P J Niziolek, B T MacDonald, C R Zylstra, N Alenina, D R Robinson, Z Zhong, S Matthes, C M Jacobsen, R A Conlon, R Brommage, Q Liu, F Mseeh, D R Powell, Q M Yang, B Zambrowicz, H Gerrits, J A Gossen, X He, M Bader, B O Williams, M L Warman, and A G Robling, "Lrp5 functions in bone to regulate bone mass," *Nat. Med.*, vol. 17, no. 6, pp. 684–691, 2011.

[307] A Wodarz and R Nusse, "Mechanisms of Wnt signaling in development," *Annual Review of Cell and Developmental Biology*, vol. 14, pp. 59–88, 1998.

[308] D Pinto and H Clevers, "Wnt, stem cells and cancer in the intestine," *Biol. Cell*, vol. 97, pp. 185–196, 2005.

[309] W E Lowry, C Blanpain, J A Nowak, G Guasch, L Lewis, and E Fuchs, "Defining the impact of β-catenin/Tcf transactivation on epithelial stem cells," *Genes and Development*, vol. 19, pp. 1596–1611, 2005.

[310] T Reya, A W Duncan, L Ailles, J Domen, and D C Scherer, "A role for Wnt signalling in self-renewal of haematopoietic stem cells," *Nature*, vol. 423, pp. 409–414, 2003.

[311] A Patapoutian and L F Reichardt, "Roles of Wnt proteins in neural development and maintenance," *Current Opinion in Neurobiology*, vol. 10, pp. 392–399, 2000.

[312] N C Inestrosa and E Arenas, "Emerging roles of wnts in the adult nervous system," *Nat. Rev. Neurosci.*, vol. 11, no. 2, pp. 77–86, 2010.

[313] P Geetha-Loganathan, S Nimmagadda, and M Scaal, "Wnt signaling in limb organogenesis," *Organogenesis*, vol. 4, no. 2, pp. 109–115, 2008.

[314] J Huelsken, R Vogel, B Erdmann, G Cotsarelis, and W Birchmeier, "β-Catenin Controls Hair Follicle Morphogenesis and Stem Cell Differentiation in the Skin," *Cell*, vol. 105, pp. 533–545, 2001.

[315] G Taylor, M S Lehrer, P J Jensen, T Sun, and R M Lavker, "Involvement of Follicular Stem Cells in Forming Not Only the Follicle but Also the Epidermis," *Cell*, vol. 102, pp. 451–461, 2000.

[316] P S Eriksson, E Perfilieva, T Bjork-Eriksson, A Alborn, C Nordborg, D A Peterson, and F H Gage, "Neurogenesis in the adult human hippocampus," *Nature Medicine*, vol. 4, pp. 1313–1317, 1998.

[317] S E Ross, N Hemati, K A Longo, C N Bennett, P C Lucas, R L Erickson, and O A MacDougald, "Inhibition of Adipogenesis by Wnt Signaling," *Science*, vol. 289, pp. 950–953, 2000.

[318] V L Church and P Francis-West, "Wnt signalling during limb development," *Int. J. Dev. Biol.*, vol. 46, pp. 927–936, 2002.

[319] C Brandon, L M Eisenberg, and C A Eisenberg, "WNT signaling modulates the diversification of hematopoietic cells," *Blood*, vol. 96, pp. 4132–4141, 2000.

[320] H Liu, M M Fergusson, J J Wu, I I Rovira, J Liu, O Gavrilova, T Lu, J Bao, D Han, M N Sack, and T Finkel, "Wnt Signaling Regulates Hepatic Metabolism," *Sci. Signal.*, vol. 4, no. 158, pp. ra6, 2011.

[321] K M Cadigan and M Peifer, "Wnt Signaling from Development to Disease: Insights from Model Systems," *Cold Spring Harb. Perspect. Biol.*, vol. 1, no. 2, pp. a002881, 2009.

[322] L M Boyden, J Mao, J Belsky, L Mitzner, A Farhi, M A Mitnick, D Wu, K Insogna, and R P Lifton, "High Bone Density Due to a Mutation in LDL-ReceptorRelated Protein 5," *N. Engl. J. Med.*, vol. 346, pp. 1513–1521, 2002.

[323] C Haumaitre, M Fabre, S Cormier, C Baumann, A L Delezoide, and S Cereghini, "Severe pancreas hypoplasia and multicystic renal dysplasia in two human fetuses carrying novel HNF1beta/MODY5 mutations," *Hum. Mol. Genet.*, vol. 15, no. 15, pp. 2363–2375, 2006.

[324] C N Qian, J Knol, P Igarashi, F Lin, U Zylstra, B T Teh, and B O Williams, "Cystic Renal Neoplasia Following Conditional Inactivation of Apc in Mouse Renal Tubular Epithelium," *J. Biol. Chem.*, vol. 280, no. 5, pp. 3938–3945, 2005.

[325] K Surendran, S P McCaul, and T C Simon, "A role for Wnt-4 in renal fibrosis," *Am. J. Physiol. Renal Physiol.*, vol. 282, no. 3, pp. F431–F441, 2002.

[326] M L Gumz, H Zou, P A Kreinest, A C Childs, L S Belmonte, S N LeGrand, K J Wu, B A Luxon, M Sinha, A S Parker, L Z Sun, D A Ahlquist, C G Wood, and J A Copland, "Secreted frizzled-related

protein 1 loss contributes to tumor phenotype of clear cell renal cell carcinoma," *Clin. Cancer Res.*, vol. 13, no. 16, pp. 4740–4749, 2007.

[327] H Yuzugullu, K Benhaj, N Ozturk, S Senturk, E Celik, A Toylu, N Tasdemir, M Yilmaz, E Erdal, K C Akcali, N Atabey, and M Ozturk, "Canonical Wnt signaling is antagonized by noncanonical Wnt5a in hepatocellular carcinoma cells," *Mol. Cancer*, vol. 8, no. 90, 2009.

[328] M Ashburner, 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, "Gene Ontology: tool for the unification of biology," *Nature Genetics*, vol. 25, pp. 25–29, 2000.

[329] S Maere, K Heymans, and M Kuiper, "BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks," *Bioinformatics*, vol. 21, pp. 3448–3449, 2005.

[330] Y Komiya and R Habas, "Wnt signal transduction pathways," *Organogenesis*, vol. 4, no. 2, pp. 68–75, 2008.

[331] P Shannon, A Markiel, O Ozier, N S Baliga, J T Wang, D Ramage, N Amin, B Schwikowski, and T Ideker, "Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks," *Genome Res.*, vol. 13, pp. 2498–2504, 2003.

[332] A Bhattacharya and R K De, "Divisive Correlation Clustering Algorithm (DCCA) for grouping of genes: detecting varying patterns in expression profiles," *Bioinformatics*, vol. 24, pp. 1359–1366, 2008.

[333] M A Felix, "An inversion in the wiring of an intercellular signal: evolution of Wnt signaling in the nematode vulva," *Bioessays*, vol. 27, no. 8, pp. 765–769, 2005.

[334] B Hendriks and E Reichmann, "Wnt signaling: a complex issue," *Biol. Res.*, vol. 35, no. 2, pp. 277–286, 2002.

[335] R Janssen, G M Le, M Pechmann, F Poulin, R Bolognesi, E E Schwager, C Hopfen, J K Colbourne, G E Budd, S J Brown, N M Prpic, C Kosiol, M Vervoort, W G Damen, G Balavoine, and A P McGregor, "Conservation, loss, and redeployment of Wnt ligands in protostomes: implications for understanding the evolution of segment formation," *BMC Evol. Biol.*, vol. 10, no. 374, 2010.

[336] A Pires-daSilva and R J Sommer, "The evolution of signalling pathways in animal development," *Nat. Rev. Genet.*, vol. 4, no. 1, pp. 39–49, 2003.

[337] O Kuchaiev, T Milenkovic, V Memisevic, W Hayes, and N Przulj, "Topological network alignment uncovers biological function and phylogeny," *J. R. Soc. Interface*, vol. 7, no. 50, pp. 1341–1354, 2010.

[338] R Y Pinter, O Rokhlenko, E Yeger-Lotem, and M Ziv-Ukelson, "Alignment of metabolic pathways," *Bioinformatics*, vol. 21, no. 16, pp. 3401–3408, 2005.

[339] C V Forst and K Schulten, "Phylogenetic analysis of metabolic pathways," *J. Mol. Evol.*, vol. 52, no. 6, pp. 471–489, 2001.

[340] B P Kelley, B Yuan, F Lewitter, R Sharan, B R Stockwell, and T Ideker, "PathBLAST: a tool for alignment of protein interaction networks," *Nucleic Acids Res.*, vol. 32, pp. W83–W88, 2004.

[341] M A Ovacik and I P Androulakis, "Enzyme sequence similarity improves the reaction alignment method for cross-species pathway comparison," *Toxicol. Appl. Pharmacol.*, vol. In Press, 2010.

[342] S Hariharaputran, T Topel, T Oberwahrenbrock, and R Hofestadt, "Alignment of Linear Biochemical Pathways Using Protein Structural Classification," *Nature Proceedings*, 2008.

[343] M Heymans and A Singh, "Deriving phylogenetic trees from the similarity analysis of metabolic pathways," *Bioinformatics*, vol. 19, pp. i138–i146, 2003.

[344] C W Chang, P C Lyu, and M Arita, "Reconstructing phylogeny from metabolic substrate-product relationships," *BMC Bioinformatics*, vol. 12, no. Suppl. 1, pp. S27, 2011.

[345] W Ali and C M Deane, "Functionally guided alignment of protein interaction networks for module detection," *Bioinformatics*, vol. 25, no. 23, pp. 3166–3173, 2009.

[346] J C Clemente, K Satou, and G Valiente, "Reconstruction of phylogenetic relationships from metabolic pathways based on the enzyme hierarchy and the gene ontology," *Genome Inform.*, vol. 16, no. 2, pp. 45–55, 2005.

[347] Y Tohsato and Y Nishimura, "Metabolic Pathway Alignment Based on Similarity between Chemical Structures," *Information and Media Technologies*, vol. 3, no. 1, pp. 191–200, 2008.

[348] F Ay, M Kellis, and T Kahveci, "SubMAP: aligning metabolic pathways with subnetwork mappings," *J. Comput. Biol.*, vol. 18, no. 3, pp. 219–235, 2011.

[349] L Liao, S Kim, and J Tomb, "Genome comparisons based on profiles of metabolic pathways," in *Proceedings of the 6th International Conference on Knowledge-Based Intelligent Information and Engineering Systems (KES'02)*, 2002, pp. 469–476.

[350] Y Li, D D Ridder, M J D Groot, and M J Reinders, "Metabolic pathway alignment between species using a comprehensive and flexible similarity measure," *BMC Syst. Biol.*, vol. 2, no. 111, 2008.

[351] J Felsenstein, "PHYLIP - Phylogeny Inference Package (Version 3.2)," *Cladistics*, vol. 5, pp. 164–166, 1989.

[352] K Tamura, J Dudley, M Nei, and S Kumar, "MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0," *Molecular Biology and Evolution*, vol. 24, pp. 1596–1599, 2007.

[353] E Pruesse, C Quast, K Knittel, B M Fuchs, W Ludwig, J Peplies, and F O Glockner, "SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB," *Nucleic Acids Res*, vol. 35, pp. 7188–7196, 2007.

[354] D E Stage and T H Eickbush, "Sequence variation within the rRNA gene loci of 12 Drosophila species," *Genome Res*, vol. 17, pp. 1888–1897, 2007.

[355] D A Benson, I Karsch-Mizrachi, D J Lipman, J Ostell, and E W Sayers, "GenBank," *Nucleic Acids Res*, vol. 39, pp. D32–37, 2011.

[356] L C Latta-IV, J W Bakelar, R A Knapp, and M E Pfrender, "Rapid evolution in response to introduced predators II: the contribution of adaptive plasticity," *BMC Evol. Biol.*, vol. 7, no. 21, 2007.

[357] P Hunter, "The human impact on biological diversity," *EMBO Rep.*, vol. 8, no. 4, pp. 316–318, 2007.

[358] W E Bradshaw and C M Holzapfel, "Evolutionary response to rapid climate change," *Science*, vol. 312, no. 5779, pp. 1477–1478, 2006.

[359] M B Davis, R G Shaw, and J R Etterson, "Evolutionary Responses to Changing Climate," *Ecology*, vol. 86, no. 7, pp. 1704–1714, 2005.

[360] D K Skelly, L N Joseph, H P Possingham, L K Freidenburg, T J Farrugia, M T Kinnison, and A P Hendry, "Evolutionary Responses to Climate Change," *Conserv. Biol.*, vol. 21, no. 5, pp. 1353–1355, 2007.

[361] B F Kochin, J J Bull, and R Antia, "Parasite Evolution and Life History Theory," *PLoS Biol.*, vol. 8, no. 10, pp. e1000524, 2010.

[362] R M Lowe, S A Ward, and R H Crozier, "The evolution of parasites from their hosts: intra- and interspecific parasitism and Emery's rule," *Proc. Biol. Sci.*, vol. 269, no. 1497, pp. 1301–1305, 2002.

[363] D N Frank, "XplorSeq: A software environment for integrated management and phylogenetic analysis of metagenomic sequence data," *BMC Bioinformatics*, vol. 9, no. 420, 2008.

[364] K S Pfennig, "Evolution of pathogen virulence: the role of variation in host phenotype," *Proc. Biol. Sci.*, vol. 268, no. 1468, pp. 755–760, 2001.

[365] P Gienapp, C Teplitsky, J S Alho, J A Mills, and J Merila, "Climate change and evolution: disentangling environmental and genetic responses," *Mol. Ecol.*, vol. 17, no. 1, pp. 167–178, 2008.

[366] C Teplitsky, J A Mills, J S Alho, J W Yarrall, and J Merila, "Bergmann's rule and climate change revisited: disentangling environmental and genetic responses in a wild bird population," *Proc. Natl. Acad. Sci., USA*, vol. 105, no. 36, pp. 13492–13496, 2008.

[367] O Kuchaiev, A Stevanovic, W Hayes, and N Przulj, "GraphCrunch 2: Software tool for network modeling, alignment and clustering," *BMC Bioinformatics*, vol. 12, no. 24, 2011.

[368] N Przulj, "Biological network comparison using graphlet degree distribution," *Bioinformatics*, vol. 23, no. 2, pp. e177–e183, 2007.

[369] S Federhen, "The NCBI Taxonomy database," *Nucleic Acids Res.*, vol. 40, pp. D136–D143, 2012.

[370] N Saitou and M Nei, "The neighbor-joining method: a new method for reconstructing phylogenetic trees," *Mol Biol Evol*, vol. 4, pp. 406–425, 1987.

[371] K Tamura, M Nei, and S Kumar, "Prospects for inferring very large phylogenies by using the neighbor-joining method," *Proc. Natl. Acad. Sci., USA*, vol. 101, no. 11030, 2004.

[372] T M W Nye, P Lio, and W R Gilks, "A novel algorithm and web-based tool for comparing two alternative phylogenetic trees," *Bioinformatics*, vol. 22, no. 117, 2006.

[373] M Adamska, S M Degnan, K M Green, M Adamski, A Craigie, C Larroux, and B M Degnan, "Wnt and TGF-β expression in the sponge *Amphimedon queenslandica* and the origin of metazoan embryonic patterning," *PLoS One*, vol. 2, no. 10, pp. e1031, 2007.

[374] P N Lee, S Kumburegama, H Q Marlow, M Q Martindale, and A H Wikramanayake, "Asymmetric developmental potential along the animal-vegetal axis in the anthozoan cnidarian, *Nematostella vectensis*, is mediated by Dishevelled," *Dev. Biol.*, vol. 310, no. 1, pp. 169–186, 2007.

[375] M Srivastava, E Begovic, J Chapman, N H Putnam, U Hellsten, T Kawashima, A Kuo, T Mitros, A Salamov, M L Carpenter, A Y Signorovitch, M A Moreno, K Kamm, J Grimwood, J Schmutz, H Shapiro, I V Grigoriev, L W Buss, B Schierwater, S L Dellaporta, and D S Rokhsar, "The *Trichoplax* Genome and the Nature of Placozoans," *Nature*, vol. 454, no. 7207, pp. 955–960, 2008.

[376] J C Croce, S Y Wu, C Byrum, R Xu, L Duloquin, A H Wikramanayake, C Gache, and D R McClay, "A genome-wide survey of the evolutionarily conserved Wnt pathways in the sea urchin *Strongylocentrotus purpuratus*," *Dev. Biol.*, vol. 300, no. 1, pp. 121–131, 2006.

[377] A Kusserow, K Pang, C Sturm, M Hrouda, J Lentfer, H A Schmidt, U Technau, A V Haeseler, B Hobmayer, M Q Martindale, and T W Holstein, "Unexpected complexity of the *Wnt* gene family in a sea anemone," *Nature*, vol. 433, no. 7022, pp. 156–160, 2005.

[378] P N Lee, K Pang, D Q Matus, and M Q Martindale, "A WNT of things to come: evolution of Wnt signaling and polarity in cnidarians," *Semin. Cell Dev. Biol.*, vol. 17, no. 2, pp. 157–167, 2006.

[379] N Riddiford and P D Olson, "*Wnt* gene loss in flatworms," *Dev. Genes Evol.*, vol. 221, no. 4, pp. 187–197, 2011.

[380] D M Eisenmann, *Wnt signaling, WormBook*, pp. 1–17, 2005.

[381] M A Herman, *Wnt Signaling in C. elegans*, chapter 12, pp. 184–209, Kluwer Academic/Plenum Publishers, 2003.

[382] S Murat, C Hopfen, and A P McGregor, "The function and evolution of *Wnt* genes in arthropods," *Arthropod Struct. Dev.*, vol. 39, no. 6, pp. 446–452, 2010.

[383] R Bolognesi, L Farzana, T D Fischer, and S J Brown, "Multiple *Wnt* genes are required for segmentation in the short-germ embryo of *Tribolium castaneum*," *Curr. Biol.*, vol. 18, no. 20, pp. 1624–1629, 2008.

[384] S J Cho, Y Valles, V C J Giani, E C Seaver, and D A Weisblat, "Evolutionary Dynamics of the *wnt* Gene Family: A Lophotrochozoan Perspective," *Mol. Biol. Evol.*, vol. 27, no. 7, pp. 1645–1658, 2010.

[385] R J Garriock, A S Warkman, S M Meadows, S DAgostino, and P A Krieg, "Census of Vertebrate Wnt genes: Isolation and Developmental Expression of Xenopus Wnt2, Wnt3, Wnt9a, Wnt9b, Wnt10a, and Wnt16," *Dev. Dyn.*, vol. 236, no. 5, pp. 1249–1258, 2007.

[386] T Miyata and H Suga, "Divergence pattern of animal gene families and relationship with the Cambrian explosion," *Bioessays*, vol. 23, no. 11, pp. 1018–1027, 2001.

[387] A L Barabasi, "Network Medicine From Obesity to the "Diseasome"," *N Engl J Med*, vol. 357, no. 4, pp. 404–407, 2007.

[388] N Tiffin, M A Andrade-Navarro, and C Perez-Iratxeta, "Linking genes to diseases: its all in the data," *Genome Medicine*, vol. 1, no. 77, 2009.

[389] L C Tranchevent, R Barriot, S Yu, S V Vooren, P V Loo, B Coessens, B D Moor, S Aerts, and Y Moreau, "ENDEAVOUR update: a web resource for gene prioritization in multiple species," *Nucleic Acids Res. (Web Server issue)*, vol. 36, pp. W377–W384, 2008.

[390] C Perez-Iratxeta, P Bork, and M A Andrade-Navarro, "Update of the G2D tool for prioritization of gene candidates to inherited diseases," *Nucleic Acids Res. (Web Server issue)*, vol. 35, pp. W212–W216, 2007.

[391] Y Zhang, D A Eberhard, G D Frantz, P Dowd, T D Wu, Y Zhou, C Watanabe, S M Luoh, P Polakis, K J Hillan, W I Wood, and Z Zhang, "GEPIS-quantitative gene expression profiling in normal and cancer tissues," *Bioinformatics*, vol. 20, no. 15, pp. 2390–2398, 2004.

[392] A Hamosh, A F Scott, J S Amberger, C A Bocchini, and V A McKusick, "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders," *Nucleic Acids Res.*, vol. 33, pp. D514–D517, 2005.

[393] X Liu, X Yu, D J Zack, H Zhu, and J Qian, "TiGER: A database for tissue-specific gene expression and regulation," *BMC Bioinformatics*, vol. 9, no. 271, 2008.

[394] E Wingender, "The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation," *Brief. Bioinform.*, vol. 9, no. 4, pp. 326–332, 2008.

[395] G Chaurasia, S Malhotra, J Russ, S Schnoeg, C Hanig, E E Wanker, and M E Futschik, "UniHI 4: new tools for query, analysis and visualization of the human protein-protein interactome," *Nucleic Acids Res. (Database issue)*, vol. 37, pp. D657–D660, 2009.

[396] D R Rhodes, S Kalyana-Sundaram, V Mahavisno, R Varambally, and J Yu, "Oncomine 3.0: Genes, Pathways, and Networks in a Collection of 18,000 Cancer Gene Expression Profiles," *Neoplasia*, vol. 9, no. 2, pp. 166–180, 2007.

[397] J Lv, H Liu, J Su, X Wu, H Liu, B Li, X Xiao, F Wang, Q Wu, and Y Zhang, "DiseaseMeth: a human disease methylation database," *Nucleic Acids Res.*, vol. 40, pp. D1030–D1035, 2012.

[398] A Palleja, H Horn, S Eliasson, and L J Jensen, "DistiLD Database: diseases and traits in linkage disequilibrium blocks," *Nucleic Acids Res.*, vol. 40, pp. D1036–D1040, 2012.

[399] L M Schriml, C Arze, S Nadendla, Y W Chang, M Mazaitis, V Felix, G Feng, and W A Kibbe, "Disease Ontology: a backbone for disease semantic integration," *Nucleic Acids Res.*, vol. 40, pp. D940–D946, 2012.

[400] A S Syed, M DAntonio, and F D Ciccarelli, "Network of Cancer Genes: a web resource to analyze duplicability, orthology and network properties of cancer genes," *Nucleic Acids Res.*, vol. 38, pp. D670–D675, 2010.

[401] S Hatsell, T Rowlands, M Hiremath, and P Cowin, "β-catenin and Tcfs in Mammary Development and Cancer," *Journal of Mammary Gland Biology and Neoplasia*, vol. 8, no. 2, pp. 145–158, 2003.

[402] D Maglott, J Ostell, K D Pruitt, and T Tatusova, "Entrez Gene: gene-centered information at NCBI," *Nucleic Acids Res.*, vol. 39, pp. D52–D57, 2011.

[403] J Amberger, C Bocchini, and A Hamosh, "A new face and new challenges for Online Mendelian Inheritance in Man (OMIM)," *Hum Mutat.*, vol. 32, no. 5, pp. 564–567, 2011.

[404] R L Seal, S M Gordon, M J Lush, M W Wright, and E A Bruford, "genenames.org: the HGNC resources in 2011," *Nucleic Acids Res.*, vol. 39, pp. D514–D519, 2011.

[405] R Goel, B Muthusamy, A Pandey, and T S Prasad, "Human protein reference database and human proteinpedia as discovery resources for molecular biotechnology," *Mol Biotechnol.*, vol. 48, no. 1, pp. 87–95, 2011.

[406] P Flicek, M R Amode, D Barrell, K Beal, S Brent, D Carvalho-Silva, P Clapham, G Coates, S Fairley, S Fitzgerald, L Gil, L Gordon, M Hendrix, T Hourlier, N Johnson, A K Kahari, D Keefe, S Keenan, R Kinsella, M Komorowska, G Koscielny, E Kulesha, P Larsson, I Longden, W McLaren, M Muffato, B Overduin, M Pignatelli, B Pritchard, H S

Riat, G R Ritchie, M Ruffier, M Schuster, D Sobral, Y A Tang, K Taylor, S Trevanion, J Vandrovcova, S White, M Wilson, S P Wilder, B L Aken, E Birney, F Cunningham, I Dunham, R Durbin, X M Fernandez-Suarez, J Harrow, J Herrero, T J Hubbard, A Parker, G Proctor, G Spudich, J Vogel, A Yates, A Zadissa, and S M Searle, "Ensembl 2012," *Nucleic Acids Res.*, vol. 40, pp. D84–D90, 2012.

[407] A Bairoch, R Apweiler, C H Wu, W C Barker, B Boeckmann, S Ferro, E Gasteiger, H Huang, R Lopez, M Magrane, M J Martin, D A Natale, C ODonovan, N Redaschi, and L S Yeh, "The Universal Protein Resource (UniProt)," *Nucleic Acids Res.*, vol. 33, pp. D154–D159, 2005.

[408] D A Benson, I Karsch-Mizrachi, K Clark, D J Lipman, J Ostell, and E W Sayers, "GenBank," *Nucleic Acids Res.*, vol. 40, pp. D48–53, 2012.

[409] K D Pruitt, T Tatusova, G R Brown, and D R Maglott, "NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy," *Nucleic Acids Res.*, vol. 40, pp. D130–135, 2012.

[410] D A Benson, I Karsch-Mizrachi, D J Lipman, J Ostell, and D L Wheeler, "GenBank: update," *Nucleic Acids Res.*, vol. 32, pp. D23–D26, 2004.

[411] B Boeckmann, A Bairoch, R Apweiler, M C Blatter, A Estreicher, E Gasteiger, M J Martin, K Michoud, C ODonovan, I Phan, S Pilbout, and M Schneider, "The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003," *Nucleic Acids Res.*, vol. 31, no. 1, pp. 365–370, 2003.

[412] J Westbrook, Z Feng, S Jain, T N Bhat, N Thanki, V Ravichandran, G L Gilliland, W Bluhm, H Weissig, D S Greer, P E Bourne, and H M Berman, "The Protein Data Bank: unifying the archive," *Nucleic Acids Res.*, vol. 30, no. 1, pp. 245–248, 2002.

[413] Y Assenov, F Ramirez, S E Schelhorn, T Lengauer, and M Albrecht, "Computing topological parameters of biological networks," *Bioinformatics*, vol. 24, no. 2, pp. 282–284, 2008.

[414] M E Smoot, K Ono, J Ruscheinski, P L Wang, and T Ideker, "Cytoscape 2.8: new features for data integration and network visualization," *Bioinformatics*, vol. 27, no. 3, pp. 431–432, 2011.

[415] L A Profenno, A P Porsteinsson, and S V Faraone, "Meta-analysis of Alzheimer's disease risk with obesity, diabetes, and related disorders," *Biol Psychiatry.*, vol. 67, no. 5, pp. 505–512, 2010.

[416] L Citrome and B Vreeland, "Schizophrenia, obesity, and antipsychotic medications: what can we do?," *Postgrad Med.*, vol. 120, no. 2, pp. 18–33, 2008.

[417] R L Kolotkin, P K Corey-Lisle, R D Crosby, J M Swanson, A V Tuomari, G J L'italien, and J E Mitchell, "Impact of Obesity on health-related Quality of Life in Schizophrenia and Bipolar Disorder," *Obesity*, vol. 16, no. 4, pp. 749–754, 2008.

[418] P E Keck and S L McElroy, "Bipolar disorder, obesity, and pharmacotherapy-associated weight gain," *J. Clin. Psychiatry*, vol. 64, no. 12, pp. 1426–1435, 2003.

[419] L Liu, L Yu, L Kubatko, D K Pearl, and S V Edwards, "Coalescent methods for estimating phylogenetic trees," *Mol. Phylogenet. Evol.*, vol. 53, no. 1, pp. 320–328, 2009.

[420] L S Kubatko and J H Degnan, "Inconsistency of phylogenetic estimates from concatenated data under coalescence," *Syst. Biol.*, vol. 56, no. 1, pp. 17–24, 2007.

[421] N Rosenberg and R Tao, "Discordance of species trees with their most likely gene trees: the case of five taxa," *Syst. Biol.*, vol. 57, no. 1, pp. 131–140, 2008.

[422] E Mossel and S Roch, "Incomplete lineage sorting: consistent phylogeny estimation from multiple loci," *IEEE/ACM Trans, Comput. Biol. Bioinform.*, vol. 7, no. 1, pp. 166–171, 2010.

[423] E M Jewett and N A Rosenberg, "iGLASS: an improvement to the GLASS method for estimating species trees from gene trees," *J. Comput. Biol.*, vol. 19, no. 3, pp. 293–315, 2012.

List of Publications

1. L. Nayak and R. K. De, “An algorithm for modularization of MAPK and Calcium signalling pathways: Comparative analysis among different species”, *Journal of Biomedical Informatics*, vol. 40, pp. 726-749, 2007.

2. L. Nayak and R. K. De, “Modularized Study of Human Calcium Signaling Pathway”, *Journal of Biosciences*, vol. 32, no. 5, pp. 1009-1017, 2007.

3. N. Tomar, L. Nayak and R. K. De, “Comparative analysis of various modularization algorithms and species specific study of VEGF signaling pathways”, *Journal of Biomedical Science and Engineering*, vol. 3, no. 10, pp. 931-942, 2010.

4. L. Nayak and R. K. De, “Disease Comorbidity and the Human Wnt signaling Pathway: A network-wise Study”, *OMICS: A Journal of Integrative Biology*, 2012. (Communicated)

5. L. Nayak and R. K. De, “Signal Transduction Pathway Resources for Everyday Research”, *Everyman’s Science*, 2012. (Revision communicated)

6. L. Nayak, N. Tomar and R. K. De, “Computational Phylogeneticity of Biological Pathways: A developmental study of TCA cycle over a set of organisms”, *in book Recent Trends in Computational Biology and Computational Statistics Applied in Biotechnology and Bioinformatics*, chapter 14, pp. 337-369, 2011, New India Publishing Agency (NIPA), New Delhi, India. (In Press)

7. R. K. De and L. Nayak, “MAPK signaling pathways and their recursive modularization”, *in Proc. 15th International Conference on Computing (CIC'06)*, Mexico City, Mexico, IEEE Press, 2006, pp. 203-208.

8. L. Nayak and R. K. De, “Developmental trend derived from modules of Wnt Signaling Pathways”, *in Proc. 4th International Conference on Pattern Recognition and Machine Intelligence (PReMI'11)*, Moscow, Russia, LNCS, vol. 6744, pp. 400-405, 2011.

9. L. Nayak and R. K. De, “Finding better partitions and conserved modules in Wnt signaling pathways”, *in Proc. The 2012 International Conference on Bioinformatics and Computational Biology (BIOCOMP'12)*, Nevada, Las Vegas, USA, 2012. (Accepted)

10. L. Nayak, N. P. Bhattacharya and R. K. De, “Deriving a phylogenetic tree from Wnt Signaling Pathways”, 2012. (Submitted to be considered for a Young Scientist Award by Indian Science Congress)