Multilevel algorithmic approach for analysis of human interactome and development of 'CancerNet' tool

Thesis submitted to

Cochin University of Science and Technology

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Under the Faculty of Science

By

Arinnia Anto Manjaly

Reg. No. 4839

Under the Guidance of

Dr.Padma Nambisan

DEPARTMENT OF BIOTECHNOLOGY

COCHIN UNIVERSITY OF SCIENCE AND TECHNOLOGY

COCHIN - 682022, KERALA, INDIA.

OCTOBER 2016

Declaration

I hereby declare that the thesis entitled **"Multilevel algorithmic approach for analysis of human interactome and development of 'CancerNet' tool"** is the authentic record of research work carried out by me for my doctoral degree, under the supervision and guidance of Dr. Padma Nambisan, Professor, Department of Biotechnology, Cochin University of Science and Technology and that no part thereof has previously formed the basis for the award of any degree, diploma, associateship or other similar titles or recognition.

Cochin - 682022
27/09/ 2016

Arinnia Anto Manjaly

DEPARTMENT OF BIOTECHNOLOGY
COCHIN UNIVERSITY OF SCIENCE AND TECHNOLOGY
COCHIN - 682 022, KERALA, INDIA.
Ph: 91-484 – 2576267 |e-mail: padmanambisan@gmail.com,
padmanambisan@cusat.ac.in

Dr. PADMA NAMBISAN
Professor

Date: 03/10/2016

Certificate

This is to certify that the thesis entitled "**Multilevel algorithmic approach for analysis of human interactome and development of 'CancerNet' tool**" is a record of bonafide research work done by Mrs. Arinnia Anto Manjaly under my supervision and guidance, in partial fulfilment of the requirement for the degree of Doctor of Philosophy, under the Faculty of Sciences of Cochin University of Science and Technology.

I certify that all the suggestions made by the doctoral committee during her presynopsis are included in the thesis, and that no part thereof has been presented for the award of any degree.

Dr. Padma Nambisan
Supervising Guide

Acknowledgement

I thank God for all the blessings you showered on me throughout the journey of life making my dreams come true.

I express my heartfelt and profound sense of gratitude to my research guide and mentor, Dr. Padma Nambisan, for her encouragement, inspiring guidance, valuable suggestions and wonderful backing throughout the course of my doctoral research. I am taking this opportunity to thank you for accepting me as a Ph.D student. All your suggestions and ideas helped me a lot to finish my work on time. Thanking you once again for making my dream come true.

I express my profound gratitude to late Dr. C.S. Paulose for his encouragement and motivation throughout the research period.

I express my gratitude to my doctoral committee members, Dr. Sartitha G Batt for her valuable suggestions and support throughout my research work.

I am thankful to Department of Science and Technology for providing the financial support in the form of Women scientist of Fellowship. I am also thankful to Cochin University for providing all facility and support for my research work. I thank all the present and past office staffs of the Department of Biotechnology for their prompt support and help. My special thanks to the higher authorities and administrative staffs of Cochin University of Science & Technology for their help and co-operation.

Words fail to express my sincere gratitude to my colleagues at the Plant biotechnology Lab. I am deeply indebted to all my seniors in our lab for providing friendly and motivating environment. I would like to thank Dr. Jikku Jose, Dr. Jasmine Koshy and Sudha Hariharan for all the support, care and friendship you offered me.

I specially thank Soumya P S, lab mate for all the loving support and affection throughout my research. I sincerely acknowledge Kiran Lakshmi M S, Anala R, Nayana P, Aiswarya C and Anuja Dileep for their support and help. Thank you for being great friends and for all the joyful atmosphere created by you.

I express my truthful appreciation to Bindiya E.S, Sritha K.S, Nanditha M, Tina K. J and Anu M.A for their friendship and support.

I really thank Harisree Nair, Rekha Mol K.R, and Dr. Karthikeyan for their constant support throughout my research. My thanks to all the research scholars of the Neuroscience laboratory especially Dr. Anju T. R for the friendship, help and co-operation.

My heartfelt thanks to contract lecturers Dr. Sreekanth, Dr Anoop and Dr. Manjusha for their constant support and valuable suggestions during my research work.

My genuine thanks to Vidhya Menon, lab mate in IISC. Thank you for being a constant supporter since the time we met. You are good motivator and my greatest strength.

Words cannot express my gratitude to my dear parents for their love, motivation and prayers. I am deeply indebted to them for their persistent support for accomplishing my goals. Thank you Appachan and Amma for being always my strength. My best friend forever, my brother Lijo Anto M A. Special thanks to my elder brother Varghese Anto M A. Sincere thanks to my in-laws for supporting me.

Last things are always the very best, my daughter Jecintha Rose (Ponnu) and Bajis Jose (my bajiettan). The two people who had to adjust a lot and supported me throughout during the research work.

Arinnia Anto Manjaly

Contents

List of Tables

List of Figures

Abbreviations

%	-	Percentage
<	-	less than
>	-	greater than
3-D	-	Three-dimensional
API	-	Application programming interface
AP-MS	-	Affinity purification coupled to mass spectrometry
AUC	-	Analytical ultracentrifugation
BioPAX	-	Biological Pathway Exchange
BP	-	Biological process
CANC	-	Cancer and non-cancer complexes
cBio	-	Catalyst Biosciences
CC	-	Cellular component
CCA	-	Connected component algorithm
CIN	-	Cancer interaction network
COSMIC	-	Catalogue of Somatic Mutations in Cancer
CP	-	Cancer proteins
DB	-	Database
DDIs	-	Domain-domain interactions
DIP	-	Database of interacting proteins
DLS	-	Dynamic light scattering
DNA	-	Deoxyribonucleic acid
EBI	-	European bioinformatics institute
et al.	-	and others
FCS	-	Functional class scoring
Fig	-	Figure
FTP	-	File Transfer Protocol
GML	-	Geography Markup Language
GO	-	Gene ontology
HGM	-	Human Genome Meeting
HIN	-	Human interactome
HIV	-	Human immunodeficiency virus
HP	-	Human proteins
HPPI	-	Human protein- protein interaction
HPPIN	-	Human protein- protein interaction network
HPRD	-	Human protein reference database
HUPO	-	Human proteome organization
Ig	-	Immunoglobin
IntAct	-	Interaction database
KEGG	-	Kyoto Encyclopedia of Genes and Genomes
LCC	-	Largest connected component
MCL	-	Markov clustering
MCODE	-	Molecular complex detection

MF	-	Molecular function
MINT	-	Molecular INTeraction database
miRNAs	-	microRNAs
MIs	-	Molecular interactions
mRNA	-	messenger ribonucleic acid
NCBI RefSeq	-	National Center for Biotechnology Information reference sequence
NNF	-	Nested network format
NPL	-	National Physical Laboratory
OCG	-	Overlapping Cluster Generator
OJB	-	ObJectRelationalBridge
OMIM	-	Online Mendelian Inheritance in Man
ORA	-	Over-representation analysis
OWL	-	Web Ontology Language
PDB	-	Protein Data Bank
PHP	-	Hypertext Preprocessor
PIN	-	Protein interaction networks
PIR	-	Protein information resource
PPD	-	Plasma proteome database
PPI	-	Protein –Protein interaction
PT	-	Pathway topology
RAS	-	Rat Sarcoma
RNA	-	Ribonucleic acid
RNSC	-	Restricted neighborhood search clustering
SBML	-	Systems biology markup language
SCOPe	-	Structural Classification of Proteins — extended
SGD	-	Saccharomyces Genome Database
SIB	-	Swiss institute of bioinformatics
SIF	-	Simple interaction format
SNP	-	Single-nucleotide polymorphism
SPC	-	Super paramagnetic clustering
SQL	-	Structured query language
TF	-	Transcription factor
UniMES	-	UniProt Metagenomic and environmental sequences
UniParc	-	UniProt Archive
UniProt	-	Universal Protein Resource
UniProtKB	-	UniProt knowledgebase
UniRef	-	UniProt reference
UV	-	Ultraviolet
XGMML	-	eXtensible Graph Markup and Modeling Language
XLX	-	Excel
XML	-	EXtensible Markup Language
Y2H	-	Yeast two-hybrid screening

Chapter 1

INTRODUCTION

A cell is the most vital unit of life. A lot of functions occur inside a cell leading to the intact generation and development of the cell. The biological macromolecules such as DNA, RNA, and proteins contribute to all these functions. Tremendous efforts have been and are being made to understand the protein complexities and to recognize different functions of proteins (Marcotte *et al*., 1999). In a living organism, the proteins links with nucleic acids and other proteins to execute diverse purposes. For a valid function, the protein interactions with other proteins or biomolecules must be qualified (Axelrod, 2001). Some unqualified alterations in proteins results in unwanted functions which ultimately leads to diseases. All these resulting diseases are directly or indirectly associated with proteins (Rossin *et al*., 2011). Environmental fluctuations, hereditary factors, lifestyle, etc causes mutation in proteins. Mutated proteins play a chief role in causing sickness (Wang *et al*., 2010].

It is known fact that everyday a variety of diseases are affecting the human beings. Immunity related diseases, neurodegenerative diseases, cardiovascular diseases, Infectious diseases, cancers, etc are the chief kinds of disease in human body. With the increasing growth of diseases, the rate of experiments are also increasing for the discovery of drugs. For drug discovery, numerous combination of proteins are investigated to relate to the correct drug. When a drug reacts to the respective protein appropriately, the result is the expected cure to evade the disease. The rate of recovery

from a disease differs from disease to disease depending on various factors like age, lifestyle, complexity, immunity, etc. Among the huge list of diseases, cancer is one of the hazardous ailment as the recovery rate is low considering the probabilities of relapse. Many evidences confirm the fact that the abnormalities of cancer cells are commonly the consequence of mutations in protein- programming genes that regulate the division of cells. This in turn results in the uncontrolled activities of protein with other small molecules. The disease starts spreading when the cancer proteins link to other proteins in a cell.

The protein networks, protein complex interaction and molecular complexes plays a vital role in drug discovery as the drug affects the normal protein also and there are chances of altering the biological system too (Sevimoglu and Arga, 2014). Another aspect of drug discovery is the protein identification. Drug discovery step can be initiated only after the identification of target protein (Hughes *et al.*, 2011). For comparative investigations, the *in vivo* experimentations and their outcomes are stored in databanks and this also efficiently support the *in silico* researches. The *in silico* researches are economical, less time consuming and reliable when compared to the *in vivo* studies. Databases are qualified only when an appropriate analysis of these biological data are performed. This also helps in retrieving the hidden information in the data, which in turn leads to comprehend new inferences. The study of large-scale biological databases are supported by system biology or computational biology (Kroger and Bry, 2003). The usage of databanks by employing diverse algorithms is one of the prime advantage of System biology. In the biological system, the computational information study of proteins and its interactions guides a researcher to disclose numerous unknown information.

A new approach termed as network biology was hosted for the depiction of biological organization constructed on mathematical graphs. In this graph, molecules are embodied as node and interactions are embodied as edges. A foremost benefit of graph theory is the study of large networks with mammoth information. The current investigation illustrates that graph-theory is exceptionally implemented in large biological system exclusively with human interactome. In biological system, a diversity of interactions emerged such as momentary or constitutive protein-protein, RNA-RNA, protein-DNA, and protein-ligand interactions. The platform for computational analysis of molecular level interactome is provided by the PPI networks.

Graph theory also helps in the forecast of hub proteins. The collaboration of proteins can be robust, feeble, steady, temporary or conditional (Hartwel *et al*., 1999). When the proteins do not interact, it results in the deprivation of biological functions. Protein interaction network contains extremely linked as well as poorly linked proteins. Most of the proteins associate with just a few number of proteins. Comparatively, a few proteins associate with a large number of other proteins. This form of expansively binding proteins are categorized as hub proteins (Jeong *et al.*, 2004). The elimination of a hub protein is considered as terrifying when coordinated with a non-hub protein. The hub proteins facilitate signaling of oncogenes.

The physically non-linked proteins can be linked in many other ways too which can be derived by employing graph theory analysis. For example, structure linked, functionally linked, pathway linked, etc. Among them, pathway analysis is the major linking property.

Pathway analysis resulted to be the leading methodology to obtain knowledge of the primary biology of varied proteins and gene expression. It regulates complexity and has amplified the descriptive part. Currently, nearly all the bioinformatics researches pursue statistically significant pathways to authenticate one or the other computationally significant consequences or biological explanation. However generally accepted, the first generation pathway analysis methods that is, Over-Representation Analysis (ORA), decodes molecular computations from ingenious investigation and agree that pathways and genes are not reliant on each other. The second-generation methodologies like Functional Class Scoring (FCS) deciphers these limits. The methodologies based on Pathway Topology (PT) additionally improves the FCS methods by creating an allowance for the category and number of interactions amid genes. This is usually ignored by FCS.

Other than these applications, there are uncertain explanation and technical challenges. The resolution data is fewer, inadequate provisional and cell-specific information, and insufficient elucidations restrict the enlargement of the next-generation approaches for pathway analysis. The ineffectiveness to follow the vital biological organization in examination confines the effectiveness of predominant procedures. However, in spite of these hurdles, as the volume and category of worthwhile explanations amplify, collective with methodological advances and research tactics that convey improved regulation for strategic anticipating for consequential biological investigations, the expediency of pathway analysis and resilience in the inferences are anticipated to magnify.

The Gene Ontology or GO project at http://www.geneontology.org/ deals with strategic, terminologies and classifications that includes

abundant ranges of molecular and cellular biology. GO is spontaneously available for usage in the description of genes, products and sequences. A number of archetypal organism catalogues and genome elucidation sets utilizes the GO and add their set of interpretation to the GO reserve. The GO directory integrates the vocabularies and subsidized annotations. It affords comprehensive permission to access this data in copious formats. The associates of the GO Consortium persistently work jointly, consulting the external specialists as obligatory, to gain and acquaint the GO vocabularies.

The GO Web resource also promotes access to prevalent certification about the GO project and acquaintances to solicitations that employ GO information for scrutiny of functional traits.

The genomics era has witnessed assembling of immense sizes of biological information, conveyed by the broad propagation of biology-focused sequences. Miscellaneous categories of information from varied means are followed in ways that is rational to the biologists to mark the outstanding consumption of biological databases and the information they encompass. The principal component of the integration task is the strengthening and usage of explanation standards like ontologies. Ontologies convey conceptualizations of information realms, which accelerates communication among researchers and the norm of domain information by computers for multifarious usages.

The high-quality PPIs and relating information sets can be used in both large and small scale studies to devise a better perception of biological system. The phenomenal increase of the *in vitro*, *in vivo* and *in silico* experiments day by day have resulted in the immense growth of extensive information about proteins and its characterizations. The entire data mined

from these researches are stowed in an accredited location termed as database. There are countless biological databanks management scheme available to stock data associated to proteins (Mayer, 2009). The databases are assembled as structural and functional databases liable to the nature of the data. The information connected to protein structure such as secondary and tertiary structures are provided in structural database. For instance, SCOPE, PDB, etc. The information connected to the protein functions are stowed in the functional databases. For example, MINT, GO, HPRD, etc. Protein domains are engaged in investigations and also for structure-based drug design. The contemporary researches confirms key importance of PPI networks in disease-drug region such as cancer, neuro disorder, etc as an outcome of abnormal conduct of interactions amid various proteins.

Graph theory also helps in the comparative analysis of human cancer and non-cancer protein complexes. A molecular level study is imperative in drug discovery to understand the tightly packed human interactome and how the protein-protein interaction complexes are linked.

In this study, the main focus was on the usage of algorithms to extract hidden information from huge databases and the application of these information in the relevant areas.

The following section portrays the areas emphasized for this investigation:

- Integration of large-scale human protein interaction networks and analysis by using multilevel algorithms.
- Identify how cancer proteins are physiologically and pathologically served in a tightly packed molecular protein complexes in a human interactome.

- Develop a tool to retrieve tightly packed molecular protein complexes present in a human interactome.

OBJECTIVES

The following are the objectives of this study:

1. Identification of available human proteins from the public databases.
2. Identification and mapping of human protein-protein interaction network (HPPIN).
3. Identification of cancer and non-cancer interactions in HPPIN.
4. Identification of major cancer and non-cancer (CANC) complexes in the network.
5. Validation of interacted complexes.
6. Identification of the hubs from CANC complexes.
7. Characterization of CANC complexes.
8. Develop a tool for CANC complexes.

Chapter 2

REVIEW OF LITERATURE

2.1 Rudimentary unit of life

The most rudimentary unit of life is a cell. All living beings are composed of cells and forms the foundation for Cell Theory in biology. Prokaryotes and eukaryotes are the main two categories of cell. The prokaryotes are single-celled organisms and eukaryotes are extremely evolved multi-cellular organisms. All plants and animals are examples of eukaryotes. The prokaryotic cell lacks a nucleus and their DNA is encompassed in the same compartment similar to the cytoplasm. The eukaryotic cells are composed of membrane-bound parts where the specific metabolic activities occur. The nucleus houses the DNA of the eukaryotic cell and plays a significant role. Fig 2.1 depicts the components of a cell.

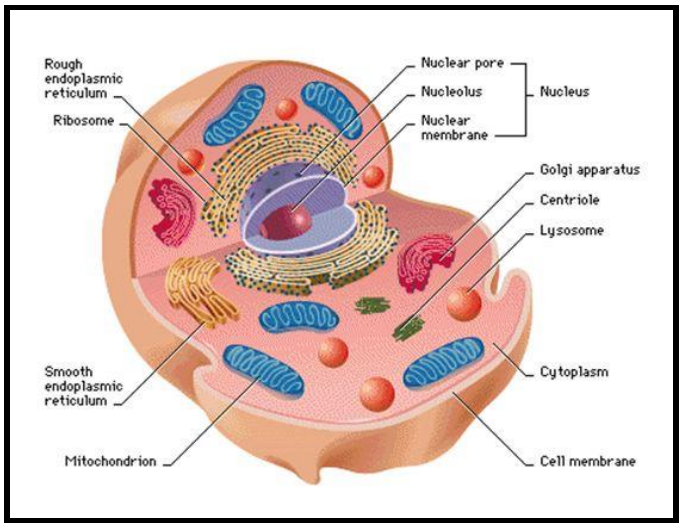

Fig 2.1: Cell components

This image displays the various components inside a cell (Adapted from Biologycorner).

2.1.1 Proteins

In a multicellular organism, the proteins are enormous and multifaceted molecules that perform numerous and critical roles. Proteins execute most of the task in a cell and are part of well-defined structural, regulation and functional aspects. Proteins are combination of 20 different types of amino acids. The specialty of a protein and the structure is determined by the sequence of amino acids. Different amino acids own dissimilar structures and chemical characteristic features. The amino acids with same charge tend to stay farther and with the opposite charge stays close to one another. The size of the amino acids also defines the configuration of a protein. Generally, a protein embraces a three-dimensional structure based on the sequence of amino acids. Table 2.1 portrays the functions of protein.

Table 2.1: Functions of protein

Function	Description	Example
Antibody	Protects the body from pathogens such as bacteria and viruses.	Immunoglobin (Ig)
Enzyme	Acts as catalyst in complex biochemical reactions occurring in a cell.	Protease
Messenger	Transmits signal to synchronize biological processes.	Hormone
Structural component	Provides support and structure.	Actin
Transport or storage	Carries and binds atom and small molecules in a cell.	Ferritin

Proteins showcase a chief character in cellular developments. Consequently, it is imperative to comprehend how they execute their functions. However, proteins do not work in isolation. They synchronize together or with other ligands to generate several biological developments in a ranked manner. Multiple proteins substantially bind with each other to formulate a stoichiometric established network. These protein complexes interact with one another to formulate functional units and pathways to execute almost all the cellular procedures. Though enormous quantities of proteomic information are existing, mining biological perceptions on proteins and protein complexes is a perplexing chore for the reason that the existing high throughput information are haphazard and unplanned. This leads to inadequately connect with the queried objective.

2.1.2 Central dogma

In molecular biology, one of the crucial mechanisms is the transfer of information from Deoxyribonucleic acid (DNA) to Ribonucleic acid (RNA) and then from RNA to proteins. This mechanism is widely known as the central dogma of molecular biology. Fig 2.2 depicts the central dogma of a cell:

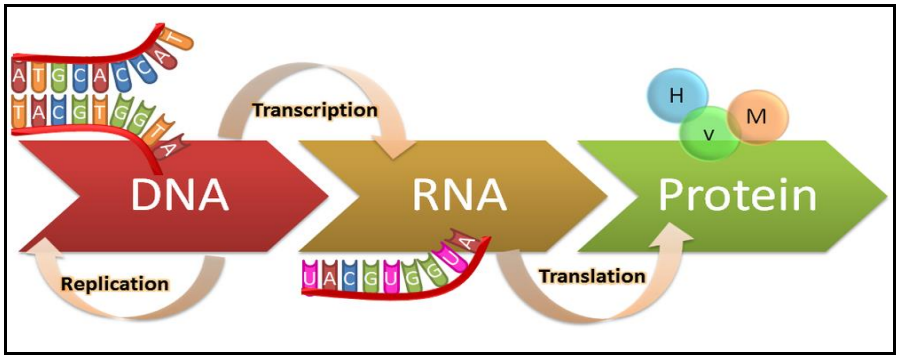

Fig 2.2: Central dogma of a cell

In molecular biology, the central dogma depicts the flow of genetic information from DNA to RNA in a biological system (Adapted from Biology Genius).

DNA or the genes are composed of the genetic instructions, which are used in the growth and operation of all living beings. Transcription is the process where the information from DNA is transferred to RNA. Several copies of messenger RNAs (mRNA) are created. Fig 2.3 is a representation of the transcription process.

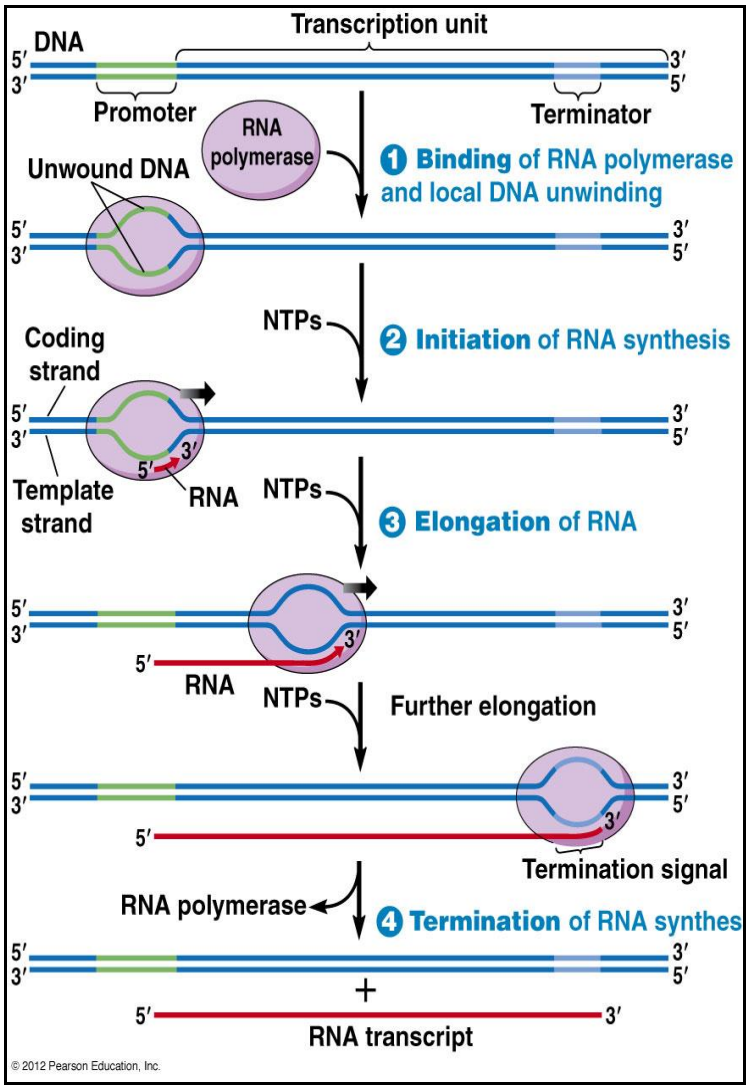

Fig 2.3: Transcription

(Adapted from Department of Biology, Memorial University of Newfoundland)

After that mRNAs are converted into proteins through translation. The mRNAs have codons which has the required information encoded in nucleotide triplets and are translated into proteins. Fig 2.4 represents the translation process.

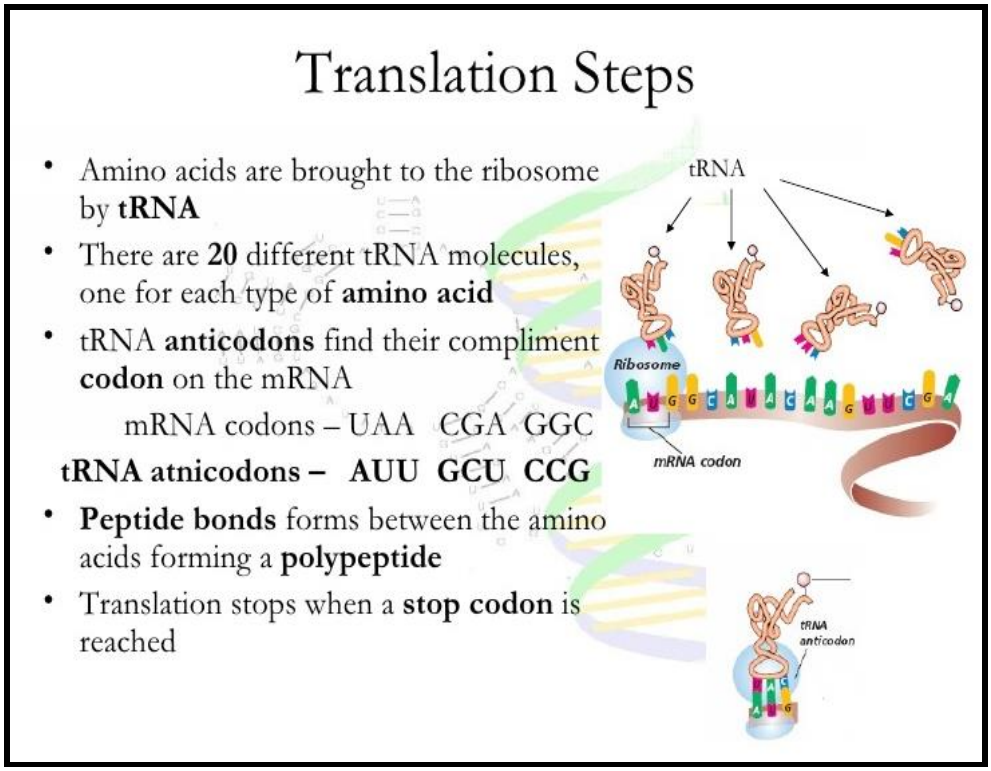

Fig 2.4: Translation

(Adapted from SlideShare)

2.1.3 Proteins as workforces

Proteins are the dynamic elements of cells. The proteins govern and arbitrate many of the biological functions and execute a cell to work. One of the significant and distinctive features of proteins is their ability to bond with other molecules and perform multiple functions. Proteins perform a wide variety or almost all the functions in a human body. For instance,

construction of structural complexes, signaling between cells, production, repair and reproduction of DNA, etc. The following section portrays the four exceptional hierarchical orders of protein structures, which are pivotal to perform a task:

- Primary structure: A linear sequence of amino acids.
- Secondary structure: A frequently reiterating resident assemblies steadied by bonds of hydrogen. Common instances are the alpha helix, beta sheet and turns.
- Tertiary structure: The three-dimensional assembly of the complete polypeptide chain, which performs protein function.
- Quaternary structure: The combination of numerous molecules of protein or single complex of proteins. The complexes of proteins are the groundwork for various cellular procedures. In addition, these protein complexes combine with other complexes and construct diverse molecular machinery, which are essential for biological functions. Fig 2.5 depicts the different structures of proteins:

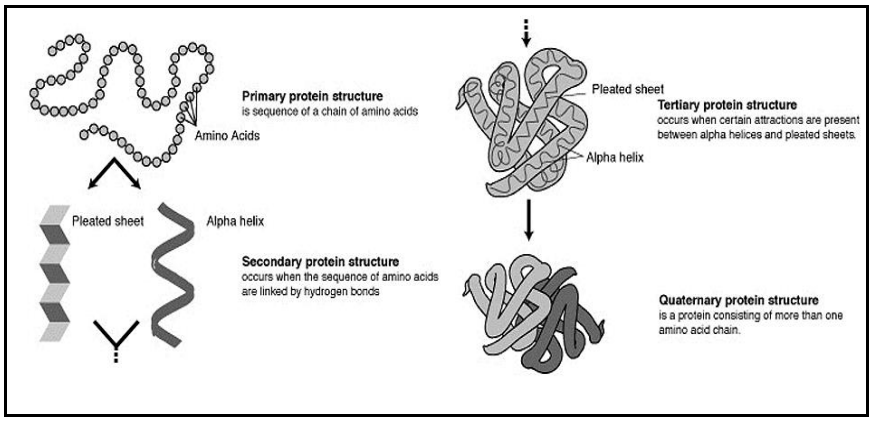

Fig 2.5: Structures of protein

(Adapted from Protein Structure in Wikipedia)

2.2 Protein-Protein Interaction (PPI)

Proteins interacts with biomolecules and other proteins to perform different functions in a living organism. When PPIs interact with other sets of PPIs, it leads to the development of protein complexes. All these complexes contribute to the formation of a protein interaction network (Ofran and Rost, 2003). In a macromolecular organization, the protein complexes are the central functional components. The widespread analysis of PPIs delivers a valuable basis for understanding the roles of protein that are mandatory for countless biological developments in living organisms (Phizicky and Fields, 1995). In addition, they also provide some useful hints on protein with unidentified function (Kemmeren *et al*., 2002). In system biology, the extensive PPI network is an essential context to study the multifaceted cellular processes and a precondition for precise prototypes.

Extensive records of PPIs focusses to establish a framework for widespread prototypes of molecular procedures. Abundantly sequenced genomes aids currently as foundation for genetics. Similarly, whole maps of PPIs is anticipated to function as a compact source for a methodical modeling line of cellular procedures. Compared to the extremely prosperous mapping genome schemes, the advancement in enlightening interactomes are very slow, especially about the human interactome. It is only in recent times that there was a tremendous effort to advance in investigational and computational efforts to expand the systematic maps of human protein interactome. Though, these maps are assumed to offer an enhanced comprehension of human biology, cautious validation of these maps are necessary. This is due to the fact that the network extracting methods owns some strong points and flaws. This may give rise to

investigational prejudices and increased degree of false positives interactions.

2.3 Generation of large-scale PPI

There are incredible methodologies accessible today to identify the interaction between proteins. These methodologies are classified according to the type of investigation background as follows:

- *In vitro*
- *In vivo*
- *In silico*

2.3.1 In vitro

The *in vitro* methods are implemented external to the living organism in a controlled environment. The following section portrays the various *in vitro* methods:

2.3.1.1 Affinity purification coupled to mass spectrometry or AP-MS

Proteins do not act in isolation. They usually arbitrate their biological roles by networking with other proteins (Charbonnier *et al*., 2008). Several methodologies are developed to analyze the multiple aspects of PPI due to the prime impact of PPIs in biology (Meyerkord and Fu, 2015). To understand the various functional aspects of protein like the affinity of proteins to bind or to understand the associated disease causing proteins we need a thorough knowledge of the interacting collaborates. However, there is no sufficient data to support this parameter. There are certain *in vitro* methodologies like Y2H to assess this. But, the comparison between *in vitro* and *in vivo* investigation results becomes a challenge.

AP-MS is an exemplary approach to study the PPI mapping (Gingras *et al*., 2007). AP-MS is the confinement of biological material

through specific fortification with a ligand combined to a solid support. The ligands can be DNA, RNA, proteins, etc. In this technique, a protein to be tested is combined with a medium. Further, the protein combination is passed through the medium. The proteins that bind with the tester protein are reserved in the medium. The proteins that do not bind are redundant (Kemmeren *et al*., 2002). Fig 2.6 is a schematic representation of AP-MS.

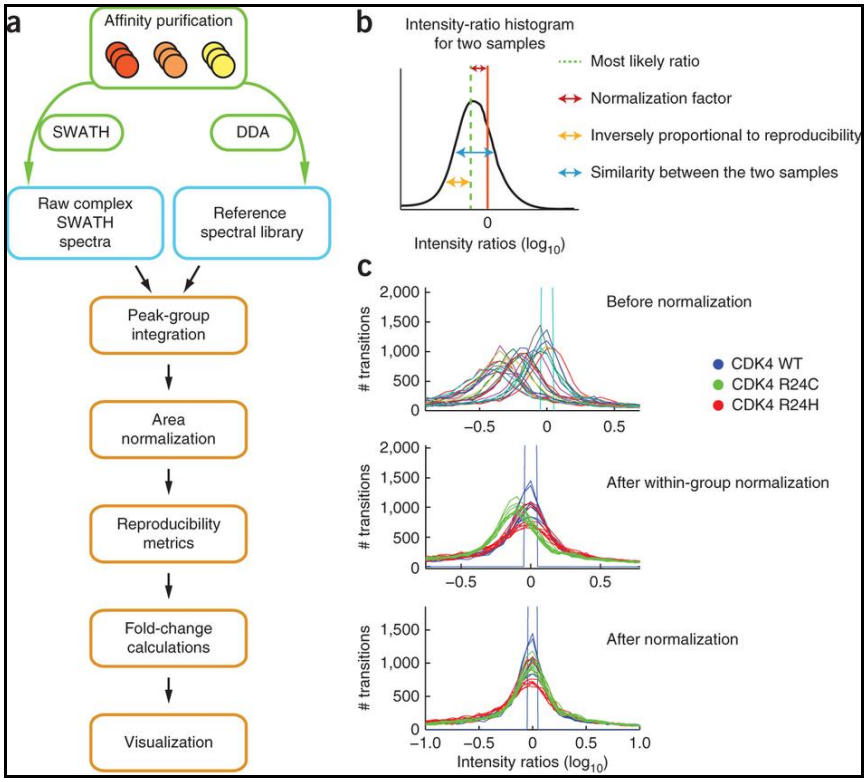

Fig 2.6: Schematic representation of AP-MS

(Figure adapted from Lambert et al., 2013)

2.3.1.2 Analytical centrifugation

The analytical ultracentrifugation (AUC) method is a multipurpose and potent technique for the quantitative investigation of macromolecules in a medium. The chief application of AUC is to investigate the

macromolecules from a variety of solvents. Absorbance, interference and fluorescence are the three optical systems employed in AUC. These parameters helps to provide apt and specific sedimentation surveillance in real time. The velocity of sedimentation uses hydrodynamic theory to describe the macromolecular size, shape and interactions. One of the advantage is that these investigations are performed in free medium. Hence, the impediments due to matrices or surfaces interactions are nil. Consequently, the sample is processed for supplementary trials followed by AUC. AUC is employed to classify the purified protein interactions. Fig 2.7 is a schematic representation of AUC.

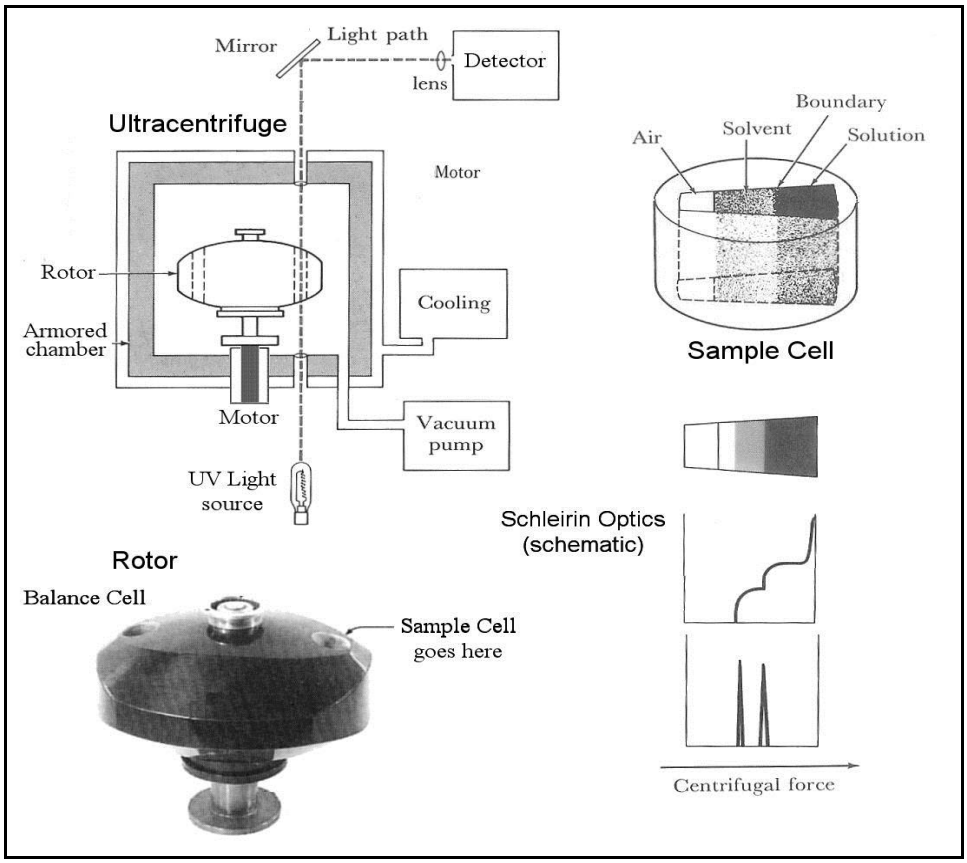

Fig 2.7: Schematic representation of AUC

(Figure adapted from Freifelder (1983))

2.3.1.4 Dynamic light scattering

Light scattering and its numerous forms contribute to an intense analysis of macromolecular collaborations in a medium. It delivers a simple clue for the association or dissociation of complexes by calculating the nonconformities in the average molecular mass. This method is based on thermodynamics, and various quantitative analysis such as reaction rate parameters and stoichiometry.

In the conventional methods, the investigations on light scattering to study protein interactions in a medium was vexing and rigorous. There was a requirement of large quantities of sample, which impeded the protein researchers. The modern technologies address these issues because of progress in instrumentation and procedures (Some and Kenrick, 2012). This also resulted in the usage of lower quantities of sample and simplified and automated investigations. Attri and Minton, 2005a, 2005b; Kameyama & Minton, 2006 were the pioneers applying these investigations to study the protein-protein interactions (Chu, 1974).

Dynamic light scattering (DLS) is a one such automated method grounded on estimating the oscillations in the intensity of light disseminated from atoms in a medium without tormenting the structure (Zhu *et al*., 2001). This is the result of Brownian motion of the disseminating components. It is employed to govern the coefficients of dissemination and the particle's dimensions in a medium quickly and precisely. DLS allows strong investigations even when placed in a microtiter plate in lower capacities with free surfaces. Fig 2.8 is a schematic representation of DLS.

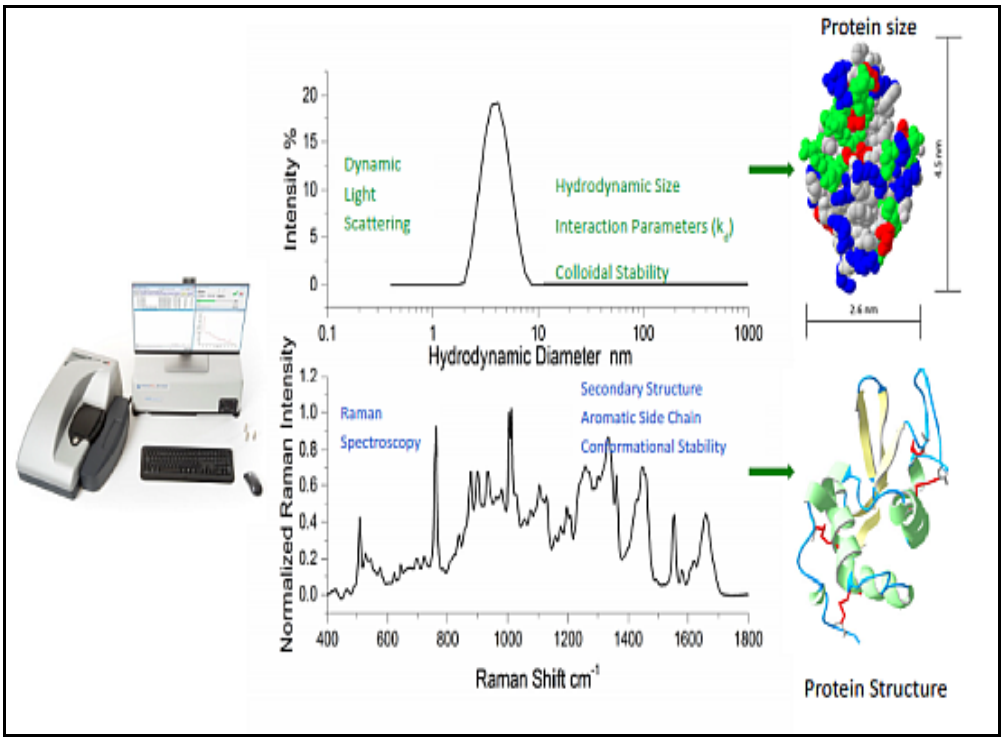

Fig 2.8: Schematic representation of DLS

(Figure adapted from Lewis et al., 2014)

2.3.1.4 Fluorescence spectroscopy

Numerous researches are conducted in the functional proteomics field to validate how the dimensions of a particular protein compound is mapped to a particular function and a particular cellular reaction. A number of methodologies are recommended to attain this objective, together with high-throughput determination of PPIs present in a proteome (Jameson *et al*., 2003). High-throughput investigation of PPIs is generally restricted to complexes of protein with dissociation coefficients of 10^{5}–10^{12} M. The bonding computation executed in living cells include lower than 104 protein particles. Subsequently, the procedures employed to enumerate PPIs falls under this range of concentration. There are more than 30,000 proteins in a living cell. Therefore, these methodologies must be qualified

to determine a specific protein interaction. Fluorescence spectroscopy is a method qualified for this requirement (Axelrod, 2003). The Fluorescence spectroscopy is generally employed to enumerate the interactions and subtleties of a single protein component present in living cells. Fig 2.9 is portrays the fluorescence spectroscopy:

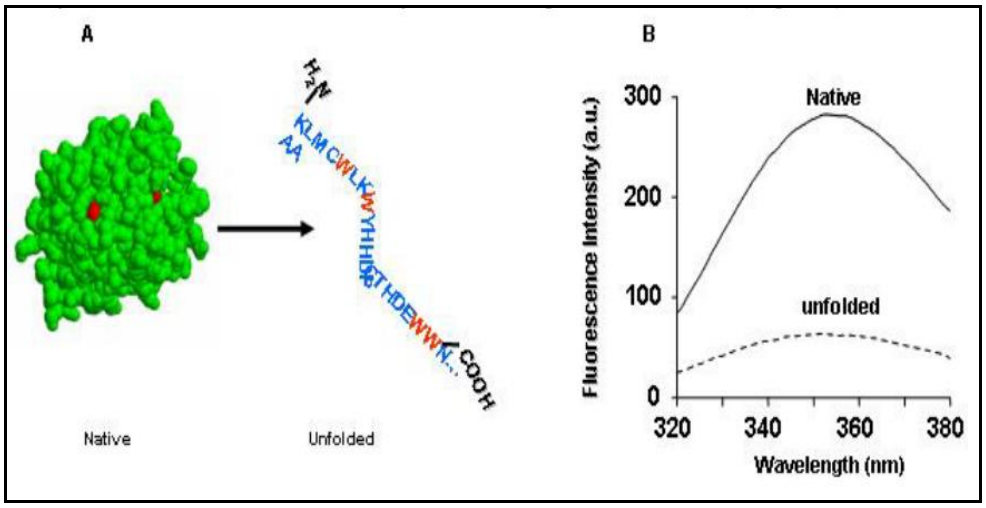

Fig 2.9: Fluorescence spectroscopy

2.3.2 *In vivo*

In vivo refers to experimentations performed inside a living being. Investigations on creatures and scientific trials are the two well-known forms of *in vivo* investigations. *In vivo* techniques can offer the protein interaction list. In this circumstance, only a miniscule portion of the potential complexes is compliant to lab investigation (Tarassov *et al*,. 2008). Yeast two-hybrid screening is employed here to study the PPIs *in vivo*.

2.3.2.1 Yeast two-hybrid screening or Y2H

The complete set of genome sequences for innumerable mock-up organisms are available today. This astonishing development in the field of

genomics has led to novel methodologies in genetic investigation to accompany conservative genetics. Subsequently, advanced investigations have to be employed in proteomics to complement developed genetics (Uetz *et al.*, 2000; Schwikowski *et al.*, 2000; Ito *et al.*, 2001; Giot *et al.*, 2003; Li *et al.*, 2004).. Enumerate procedures are established to illustrate proteins and their respective functions on a large scale (Fields and Song, 1989). Fig 2.10 represents Y2H.

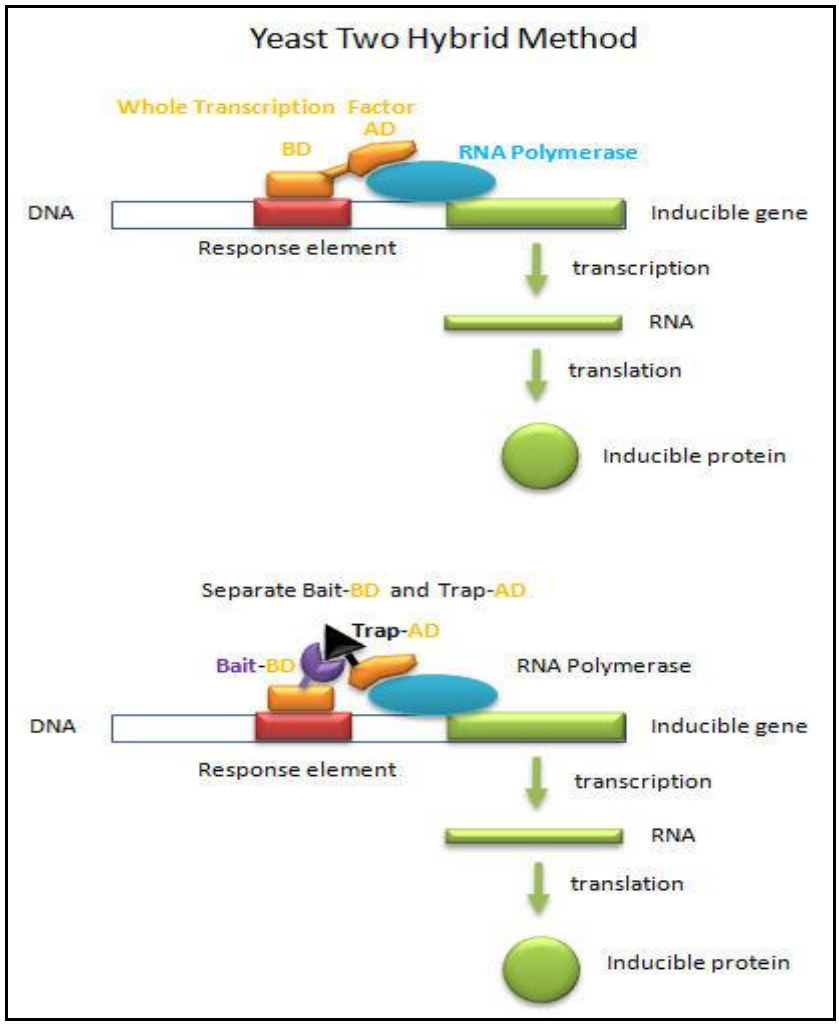

Fig 2.10: Yeast two hybrid method

(Figure adapted from BIOwiki)

One of the most widely employed procedure to investigate PPI is Yeast two-hybrid screening (Y2H). This procedure is successful to identify the PPIs on a large scale. The prime aspect of examination is the physical communications or binding among the proteins (Hishigaki *et al*., 2001). Y2H is based on the restructuring of an effectual transcription factor (TF) after the communication amid manifold proteins. This type of reformation befalls in genetically metamorphosed yeast strains, where a certain phenotype develops as an outcome of transcription of the reporter gene such as HIS3. This is established from the modification of colour in the yeast colonies bred on a medium without histidine and LacZ (Eisenberg *et al*., 2000).

2.3.3 *In silico*

In silico methods are implemented by employing computer simulation with mock-ups reminiscent of a real world or on a computer. The sequence and structure based approach is employed to investigate PPIs *in silico*. The *in vivo* methods and *in vitro* methods directed to the advance of extremely momentous tools to study PPIs. However, *in silico* ways and means substantiated to be of sophisticated accuracy, enhanced support of data-intensive exploration, precise replications through sophisticated models and supplementary adeptness. Fig 2.11 portrays the *in silico* environment.

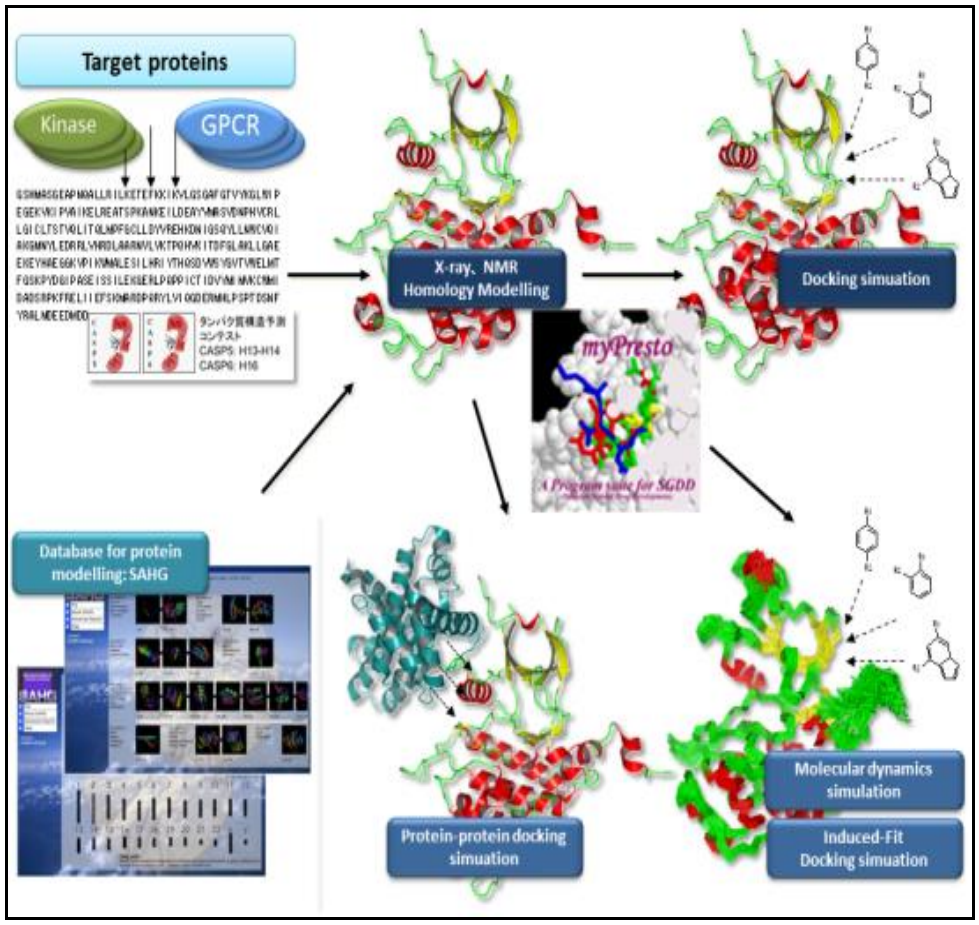

Fig 2.11: *In silico* environment

2.3.3.1 Structure-based prediction approaches

The swift accretion of novel genes in the databases emphasizes on challenges of recognizing the governing DNA structures constructed on the sequence information. Most prognostic algorithms are grounded on manifold arrangements of recognized binding locations. The structural based approach is novel to scrutinize the protein-DNA network. The notion of structure-based approach is to forecast PPIs, in cases where two proteins have an analogous structure. For instance, if two proteins A and B interacts with each other, then there are chances of two other proteins A1 and B1,

which may be similar to A, and B. The resultant inference is that there are probabilities that A1 and B1 interacts with each other. The PDM database validates resources and tools to researches for fostering the structure of a probe protein.

The sequence of transcription start points owns a chief character in gene expression control. They are predictable by controlling proteins which perform by governing the degree of transcription origination and binding as transcription activators or repressors. The recognition of such sequences from a particular gene is a challenge and is vital for comprehending its transcription at regulation level.

2.3.3.2 Sequence-Based prediction approaches

The requirement to transmute the mounting quantity of biological data into facts involved a number of disciplines. This is acquired by employing investigational and computational methodologies. This approach aims to decrypt the functional aspect of linkages and interfaces flanked by proteins (Sharan *et al*., 2007). The available computational methodologies for forecasting PPIs provide very less information compared to the information available for genomic sequences. Some of the aspects like biological function, gene expression and essentiality are moderately explored in a small number of living beings.

An inference was arrived at that a distinctive modus operandi to detect PPIs in an interaction network based on the genome sequences favorably supports the forecast of interaction networks. The sequence-based approach is grounded on the notion that an interface detected in one species can be used as an allusion for interaction in another species. Furthermore, orthologous-based approach and domain-pairs based

approach are the widely used categories of sequence-based prediction approaches.

2.3.3.3 Orthologous-based approach

The rapid progression of PPIs facts directed to the initiation of analysis of PPI network. In spite of progress in most advanced procedures, the information on various model organism's interactomes are incomplete. Orthologous-based approach is beneficial to explore these interactomes.

Orthologues are genes present in dissimilar species but originated from a mutual ancestor because of speciation (Lee *et al.,* 2008). The notion behind the orthologous-based methodology is to transmission annotation from a functionally demarcated protein sequence to the target sequence grounded on the uniformity. This was accomplished by means of the pairwise indigenous sequence algorithm (Memisevic and Przulj, 2012).

The extension of the entire orthologous sets facilitates the interpretation for innumerable eukaryotes. The same interpretation technique is also realistic to the forecast of inter-species interaction.

2.3.3.4 Domain-pairs based approach

The requirement to understand which protein is present in a particular organism and to decipher PPI are essential to comprehend the various biological processes at cellular level. As a result, the prediction of PPI networks was the objective for research in the field of proteomics.

A domain is an exclusive organizational entity of protein that is not dependent on other entities. Protein domain take part in an significant role in the forecast of protein structural class, sub-cellular location of proteins, the type of protein membrane, the class and subclass of enzymes.

Protein domains are used in researches and for structure-based drug designing (Memišević *et al.,* 2013). Furthermore, domains are involved in

the PPI at intermolecular level and therefore, are rudimentary to PPI. It is manifested from manifold studies that domain-domain interactions (DDIs) from miscellaneous experimentations are stronger than their equivalent PPIs (Rivas *et al*., 2010).

2.4 Graph theory

A major challenge confronted by proteomics is the understanding of enormous number of PPIs. When the protein networks were investigated, it was found that the protein binding characteristic has the capability to provide valuable awareness about the internal mechanism of cells as well as deducing complex maladies. In computation biology, the graph theory shares a pivotal role after the cascade of PPI data from the recent investigations of PPI. The PPI network are represented as graphs on a general note. In this case, nodes and PPIs signify the proteins by edges. The networks are of varying length and fixed edges in general. Conversely, many of the contemporary PPI networks signify only binary bindings and are not directed.

In system biology, one objective is to elucidate structural, functional and regulation of networks from the available resources. Graph theory is the focal point for this objective as it facilitates analysis of structural properties of PPI networks, which in turn is linked to the functional aspect. This approach leads to competent and actual planning of investigations as graph theory aids in theory generation by constructing predictive models.

2.5 Interaction amid complexes

A pathway generally contains a group of stoichiometric complexes that interact to accomplish a particular biological task. In this procedure, the complexes interact with one another to synchronize their activities for

varied causes. As a result of interaction, two complexes comes physically close to one other. This allows them to perform together on a substrate. In certain cases, one complex orients the substrate to generate some intermediate product and some other complex governs the intermediate product to deliver the final product. This two-step process is executed competently by interacting and staying in vicinity. Two complexes are also brought closer because of interaction, which in turn alters the other complex. The alteration either triggers or constrains the other complex by changing its 3-D alignment.

The interactions occurs depending on the requirement for a particular biological job like environmental changes. Hence, they are momentary by nature. This momentary nature leads to trouble in detecting the complex-complex interactions. In addition, computational investigation of interactions amid complexes are inadequate due to the deficiency of an ample group of recognized complexes.

It is evident from several investigations that interacting complexes are susceptible to share a common functional group. This is partially owing to the circumstance that complexes interact with one another in the similar pathway to attain a definite biological mission. Hence, as an alternative to forecast interactions amid complexes, trials can be carried out to forecast complexes belonging to the same pathway. Nevertheless, compared to the identification of complexes, the categorized information and superior evaluation of complexes in a common pathway are deficient.

2.6 Protein- protein interaction network

A protein interaction network is the consequence of infinite binary interactions. A miscellaneous network of proteins combined by interactions as edges represents a PPI network. The PPI network provides a

comprehensive outlook of cellular function and biological procedures (Kar *et al*., 2009). The forecast of the protein function is the paramount intent of the PPI network. The expansion of a reliable and stable PPI network is significant for offering an understanding at the elementary level of a sickness mechanism (Zhang, 2009). The PPI network can be employed as a definitive prototype for assimilation of data and investigation (Lin *et al.*, 2007). PPIs proved to be a prevailing source to investigate illness mechanisms, associated diseases, etc in many studies. The progression of a comprehensive data set of PPI network turns out to be a massive advantage to scholars. The accumulation of the entire identified PPIs for a particular cell or organism is designated as interactome. The interactome network functions as a mean to interpret the graph theory, which in turn supports understanding of manifold biological developments. There are few concerns in the availability of interactome data (Tuba and Kazim, 2014). The data inadequacy related to the PPI network prompts a danger to any disease genes investigation.

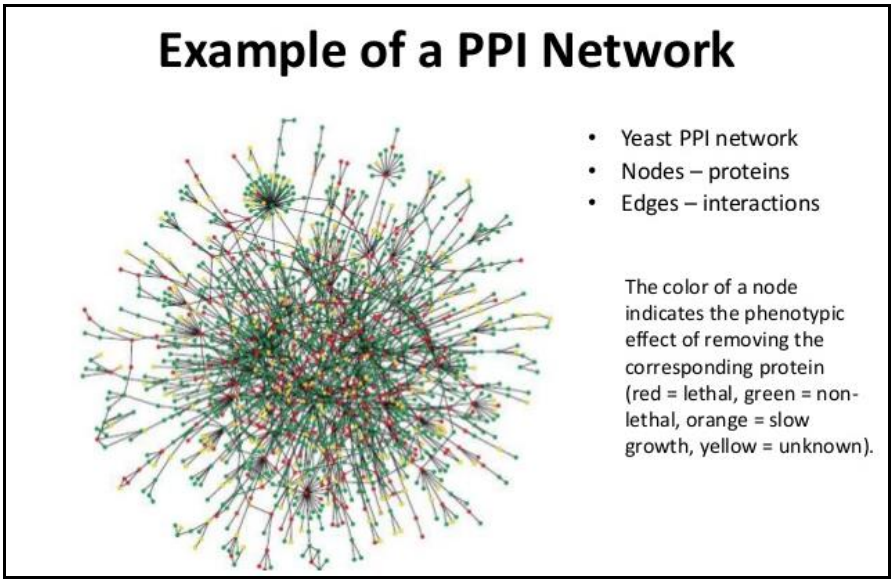

Fig 2.12: Yeast PPI network

(Figure adapted from SlideShare).

The quality of PPI networks must be high for the researches related to the biomedical field. Nevertheless, the existing extensive human PPI networks are relatively unsystematic and also speculated as false positives. Several assurance increasing patterns by means of omics were established to resolve this concern. Moreover, a small number of existing PPI networks deliver their particular assurance counting patterns. Incorporation of PPI maps with such assurance score aids the scholars to evaluate the excellence of interactions that exist in the databases.

An actual concern is the biologically meaningful explanation of the PPI maps. Though improvements in contemporary genome-wide interactome schemes prompted a massive PPI information, this also leads to new challenges for scholars primarily owing to the complexity of interaction networks. In order to comprehend this complexity, it is essential to achieve meaningful data in the background of biological systems. This not only necessitates identification of the functional aspect of single proteins but also the biological procedures and somatic interactions in which they participate. To achieve this, the PPI networks are supposed to be integrated with other functional information to develop the extensive data. Earlier investigations portray that coupling PPI networks with pathway or expression information leads to illustrate prospective transformers of various diseases or biological developments.

The resolution demands complete incorporation of the existing human interaction maps, executing enquiries at the network level, assessing information with high quality, updating the integrated database regularly, and integrating the PPI networks with additional genomic and functional information.

2.7 Human protein atlas

The Human protein atlas is a massive store of the proteomic data position (Newberg *et al.*, 2009). Highly-throughput automated tools are required to explain and categorize proteins and their functions across different cells as the database's size makes progression (Uhlén *et al.*, 2005). The classifiers are mentored to identify different patterns of proteins by including the mentored learning method. This approach is successful to analyze the subcellular patterns (Glory and Murphy, 2007). The contemporary researches prove that the categorization can be extracted to analyze patterns of proteomes (Chen *et al.*, 2007). One of the exceptional characteristics of the Human Protein Atlas is the availability of proteins that is imaged in innumerable diverse cells and tissues. The addition of confocal microscopic images of various antibodies to the atlas is a considerable enhancement in recent times (Barbe et *al.*, 2008).

2.8 NCBI's Reference Sequence

NCBI's reference sequence or NCBI RefSeq is a publically available databank associated with naturally existing protein sequences, RNA and DNA. The uniqueness of this databank is with respect to the huge, curated sequence databank demonstrating distinct but unambiguously allied records ranging from appropriate genomes to translation products at the multiple species level. Compared to similar other databases, NCBI's RefSeq is preferred due to the non-redundant, comprehensively cross-linked, and opulently explained records on proteins and nucleic acid (NCBI Handbook).

2.9 Plasma Proteome Database

In 2005, Plasma Proteome Database or PPD was coined as part of Human Proteome Organizations (HUPO). PPD led way in augmentation of

proteomics and aided in identifying numerous new plasma proteins. There was a vast amount of data accommodated in this regard and hence resulted in the enhancement of proteomics data in the PPD. The latest version has included the information on mass spectrometry-resultant information as well as data related to manifold reaction observing analyses after verifying through experiments. The plasma proteins are a key drive in the area of biomarkers. So, PPD has facilitated a batch-based inquiry labelled Plasma Proteome Explorer. This allows the users in comparing a proteins list with identified plasma proteins to evaluate originality of the database. In humans, PPD expedites researches by supporting as an all-inclusive plasma proteins reference, facilitate discovery of biomarkers and tasks related to translation. http://www.plasmaproteomedatabase.org/ provides more information on PPD.

2.10 Human interactome

The products of gene mediate their respective function inside intricate complexes of interrelated macromolecules. Investigations in prototypical entities propose that compound macromolecular systems own lively and topological features that replicate biological occurrences (Vidal *et al*., 2011). The interactome network is the comprehensive assortment of the entire PPIs that occurs inside a cell. Therefore, a complete understanding of relationship between genotype and phenotype in human beings requires explanations of how interactome networks are disturbed because of hereditary and somatic disease susceptibilities. In turn, this necessitate high quality and widespread proteome and genome-scale records of macromolecular interfaces such as PPIs, protein-nucleic acid interactions, etc.

In the initial researches, the human PPI interactome maps delivered network based elucidations for few of the genotype and phenotype associations. These implications persist as inadequate and of deficient feature to develop precise worldwide elucidations. Consequently, there is a necessity for empirically governed superior proteome-scale interactome reference records.

The challenges are multifarious to produce a wide-ranging twofold reference PPI map. It is indeterminate to remark now whether such a wide-ranging network can be mapped ever by the cooperative exertions of modest investigations. The forecasts using computational approaches of PPIs can generate data at proteome scale.

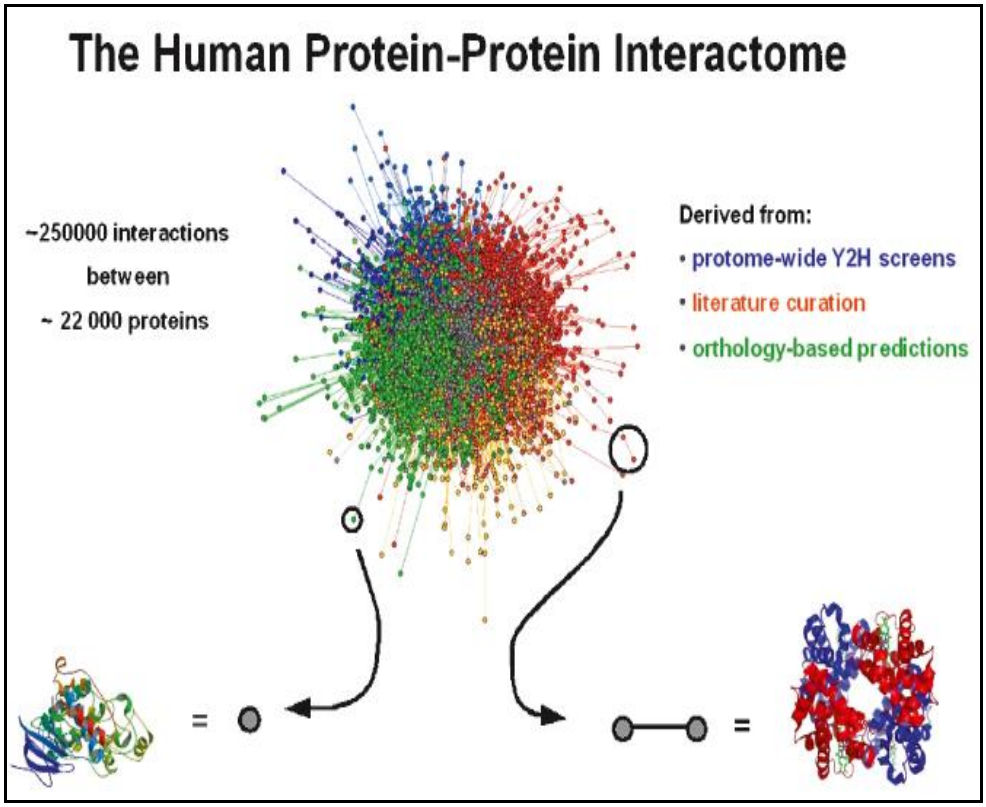

Fig 2.13: Human protein-protein interactome with 22000 proteins

2.11 PPIs and diseases

PPI is the result of association of two or more proteins with each other by countless ways. The PPIs provide a key to understand the biological processes that transpire within and among cells. The inference from various studies confirm the fact that biological processes are fundamentally interactions amid manifold proteins (Zhang *et al*., 2011). In addition, the PPI networks govern the course of data inside and amid all biological processes.

The commotions in PPI networks leads to many diseases. It can be monogenic diseases like sickle cell anemia as a result of disruption of a specific biochemical pathway or a further complicated ailments like cancer, where numerous signaling pathways (Sam *et al*., 2007) are involved. On the contrary, the commotion of a group of PPI results in a specific ailment or the group of PPIs are shared between numerous networks and in turn to several diseases. There is a surplus of information available to infer the association of protein and ailments that are integrated to PPI databases. The major challenge is mapping the PPIs to the specific human ailments (Ideker & Sharan, 2008).

The methodologies to support the high-precision forecasts that are useful to determine the effective prevention of ailments, analysis and diagnosis must be the objective of PPI related studies. The engendered substantiations must be investigated to govern their significance.

2.12 Cancer

A multicellular organism survives when its entire cells perform according to the instructions defined for cell development and replication. In some circumstances, a regular cell turn out to be rebellious, divide franticly, attack other cells, seizing sources, and ultimately slaying the host

body. Fig 2.14 depicts the difference between a normal call and a cancer cell.

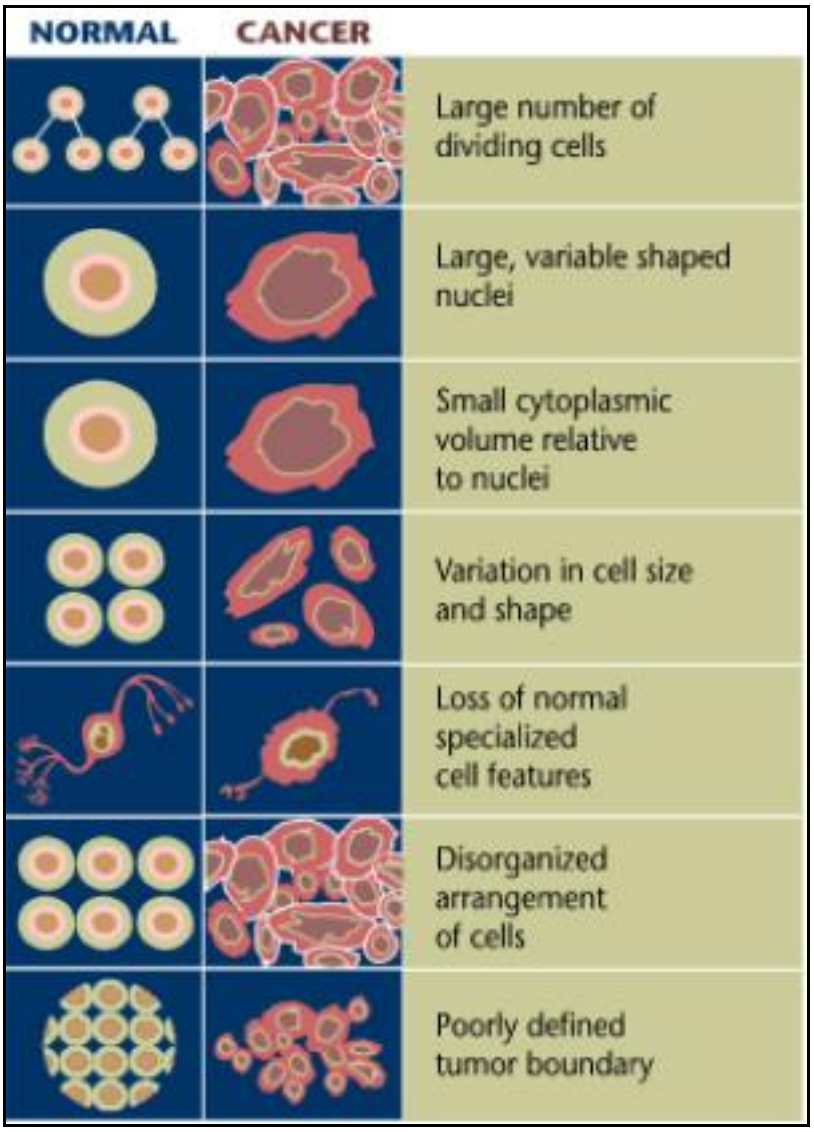

Fig 2.14: Normal and cancerous cell

(Figure adapted from Viewyer).

In order to comprehend the reason behind the rebellious nature of a cell, it is essential to understand the regular process of cell development and replication. The researches performed cell biology, molecular biology, biochemistry unveiled the astoundingly meticulous data on molecules, and procedures that consents cell division, development, segregate, and execute indispensable roles. Such rudimentary information of cell biology steered the real-time detections of the process of cancer. There are specific molecules, which are responsible for the development of a cell by means of cell cycle and in turn control cell development. This comprehension of regular cell cycle procedures and the processes that are skewed decipher the respective mechanisms that elicit cancer. One of the major reasons prominent to the development of cancer is the irregularity in the cell cycle. Cancer encompasses minimum of 100 diverse maladies. However, the entire cancer cells own one common characteristic. The anomalous cells are identified to having disrupted multiplication of cells. To be precise, cancer are the result of alterations that leads the normal cells to procure anomalous functions. The alterations are the outcome of hereditary transmutations or are prompted by environmental aspects such as X-rays, UV light, tobacco products, viruses, and certain chemicals. Most of the proof recommends that most cancers are the outcome of various events or factors. In other words, maximum four to seven actions are typically the prerequisite for a regular cell to develop through a sequence of malicious phases to aggressive cancer. It requires years to lapse from the initial occurrence to the growth of cancer. The expansion of biological procedures aids in the initial phase analysis of plausible cancers even before the cancerous cells are observable.

Cancer is the result if a sequential molecular procedures that primarily vary the regular features of a normal cell. The regular governing system that averts overgrowth of cells and the foray of other tissues are incapacitated in cancer. The transformed cells multiply and develop with the aid of signals that generally impede cell development. Consequently, there is no requirement of dedicated signals to prompt development and multiplication of cells. These cells acquire new features, lower linkage of cells, and manufacture of novel enzymes, and alterations in structure of cells. The transmissible alterations permit the cell and its descendants to multiply and develop among regular cells that usually impede the progression of neighboring cells. Consequently, the cancerous cells proliferate and attack other normal cells.

The anomalies of cancer cells are generally the outcome of transmutations in protein- programming genes that control division of cells. More genes are mutated over a period. The result of mutation of these genes is that the proteins, which usually repair DNA damages due to mutation. Subsequently, mutations proliferate in the cell and affects the normal functioning of cells and their progenies. Certain mutated cells perish but other variations leads to multiplication at rapid level of the abnormal compared to the regular cells. This superior progress labels most cancer cells with the functions restricted in the regular and vigorous cells. These cells are designated as benign if they reside in their original site and as malignant when they become hostile. In malignant types, the cancer cells frequently metastasize and propel cancer cells to distant locations in the body and leads to novel cancerous cells.

The proto-oncogenes encrypt the proteins that conduct a signal to the nucleus to kindle the division of cells in normal cells. The signalling proteins perform in a sequence of phases designates as signal transduction cascade or pathway. This flow comprises a membrane receptor for the molecular signal, intermediate proteins that mediate the signal across the cytoplasm, and transcription aspects in the nucleus that trigger division of cells in the genes. One protein triggers the next in every phase of the cascade. Nevertheless, some aspects or factors triggers more than one protein in the cell. Proto-oncogenes are a set of genes that lead normal cells to convert into cancerous cells because of mutation (Adamson, 1987; Weinstein and Joe, 2006). In proto-oncogenes, mutations are stereotypically prevailing in nature. The mutated variety of a proto-oncogene is designated as an oncogene. The oncogenes triggers the signaling pathway incessantly and the outcome is the augmented manufacture of factors stimulating expansion of cancerous cells. The proto-oncogenes program proteins that works to trigger division, impedes differentiation, and terminates death of cells. The entire procedure is significant for regular growth of human and for the preservation of organs and tissues. However, oncogenes usually demonstrate amplified manufacture of these proteins and leads to augmented division of cell, diminished differentiation of cell, and hindrance of cell demise. In total, these phenotypes describe cancer cells. Therefore, oncogenes are presently a chief molecular objective for designing anti-cancer drugs. For instance, RAS is an oncogene that usually plays the role of a switch in the signaling pathway, that is, it switches on or off the signaling pathway. When a mutation occurs in in RAS, it switches on the signaling pathway interminable ensuing in an uninhibited growth of cells. It is evident from various studies that about 30 percent of cancers are

the upshot of mutation in RAS. For instance, lung, thyroid, pancreatic and colon carcinomas.

There are innumerable aspects, which lead to the formation of oncogene. These aspects include the redisposition of genes in the chromosome that transfers the proto-oncogene to a new site, transformation of a proto-oncogene to an oncogene by transmutation of the proto-oncogene, a proliferation in the quantity of normal proto-oncogene replicas or a virus induced in the DNA or approximately a proto-oncogene. The outcome of any of these aspects is a transformed form of the gene leading to cancer. Most of the oncogenes are prevailing mutations. A single replica of this gene is adequate for manifestation of the progress trait. The cells inheriting the mutant form of the protein acquires a new function that is not existent in cells with the regular gene.

The proteins generated by tumor suppressor genes generally impedes growth of cells and prevents formation of tumorous cells. Transmutations in these genes produces cells that no longer displays irregular development and multiplication of cells. The tumor suppressor genes products perform in the cytoplasm, at the cell membrane, or in the nucleus. Mutations causes forfeiture of function and turns them to recessive. This implies that the attribute is not articulated unless the mutation of both the normal gene replicas.

2.12.1 PPIs and cancer

The manifestation of cancer genomics, focused treatments, and system oncology have ominously extended the background of PPI networks in cancer for healing detection. Widespread medical and biological research steered the prediction of protein interface hubs and nodes essential for the procurement and preservation of cancer features indispensable for

transformation of the cell (Garraway *et al.*, 2013). This type of cancer facilitating PPIs are promising targets for healing. The PPI interfaces targeting as an anticancer stratagem has turned into a realism by means of progressions in technology related to detection and substantiation of PPI and PPI-targeting proxies in experimental backgrounds. The imminent research fixated at genomics-based PPI focused detection, categorization of PPI networks, PPI targeted biochemical library proposal, and use cases based on cancer affected infirm persons speed up the expansion of PPI grounded anticancer agents which helps in personalized medicines for future generation (Hennessy *et al.*, 2005). Fig 2.15 depicts the PPIs and cancer association.

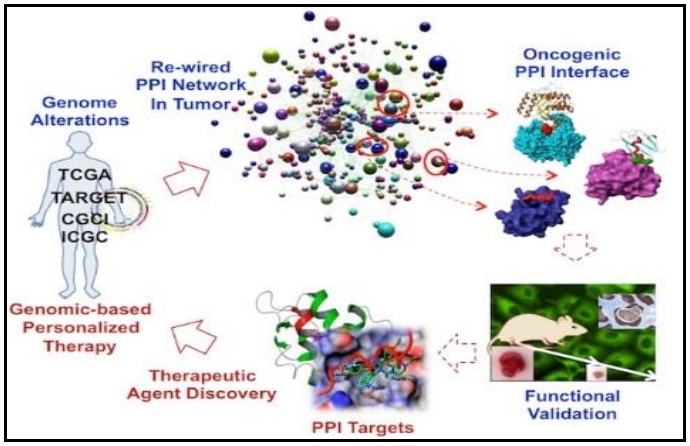

Fig 2.15: PPIs and cancer association

(Figure adapted from National cancer institute, 2013)

2.13 Mounting curiosity in PPI targets

PPI networks epitomize a vastly capable and challenging set of prospective targets for healing progressions (Wells and McClendon, 2007). In cancer, PPIs develop signaling hubs and nodes that conduct pathophysiological signals besides molecular interfaces to attain a cohesive biological harvest. This ultimately promotes origin of tumor, advancement

of tumor, incursion, and metastasis. As a result, trepidation of pathways by the PPI commotion perilous for cancer, propositions an innovative and operative stratagem for curbing the spread of cancerous signals. As oncology has advanced ominously advanced in recent period, the curiosity in PPI targeting as anticancer stratagems also augmented.

2.14 PPI networking and cancerous signalling networks

The generation of tumor cells are the result of diversity of ecological, epigenetic, and inherited factors that prompt the cancer instigating cells and the procurement of molecular and physical features. The features include circumvention of tumor suppressors and constant multiplying signaling allows the expansion and advancement of malignancy and are renowned as distinguished trademarks of malignant cells (Hanahan and Weinberg, 2011). These trademarks are responsible for a molecular agenda to comprehend factors related to cancer, associating signaling procedures to pathological aftermaths. A mixture of hereditary and epigenetic amendments using highly coordinated signaling complexes identifies the capacity of cancerous cells. It is notable that PPIs depicts the rudimentary entities inside such dynamic complexes.

PPIs perform indispensable roles in associating complexes that transmit cancerous signals, permit the procurement of trademark characteristics of cancer, and extend assorted roles in motivating and upholding the development of cancerous cells on oncogenic prompt (Imielinski *et al.,* 2012). PPIs initiates a series of reactions to endorse unrestrained cell division whether it is involving receptors with non-regulated tumor factors or dimerization of receptor elicited by gene augmentation or transformations (Hennessy *et al*., 2005). The consequence of cancerous complex rescheduling is that few PPIs promotes distinctive

features of cancer and in contrast, other PPIs are imperative for multiple features of cancer. Consequently, the assumption is that the intervention of few grave PPIs can incapacitate manifold processes that cancerous cells depends on or endurance. Enormous number of PPIs are active in executing tumor generation by the ordinance of oncogenic complexes. Hence, these PPI networks signify productive base for the detection of anticancer treatments. Fig 2.16 depicts the properties of cancer signaling network.

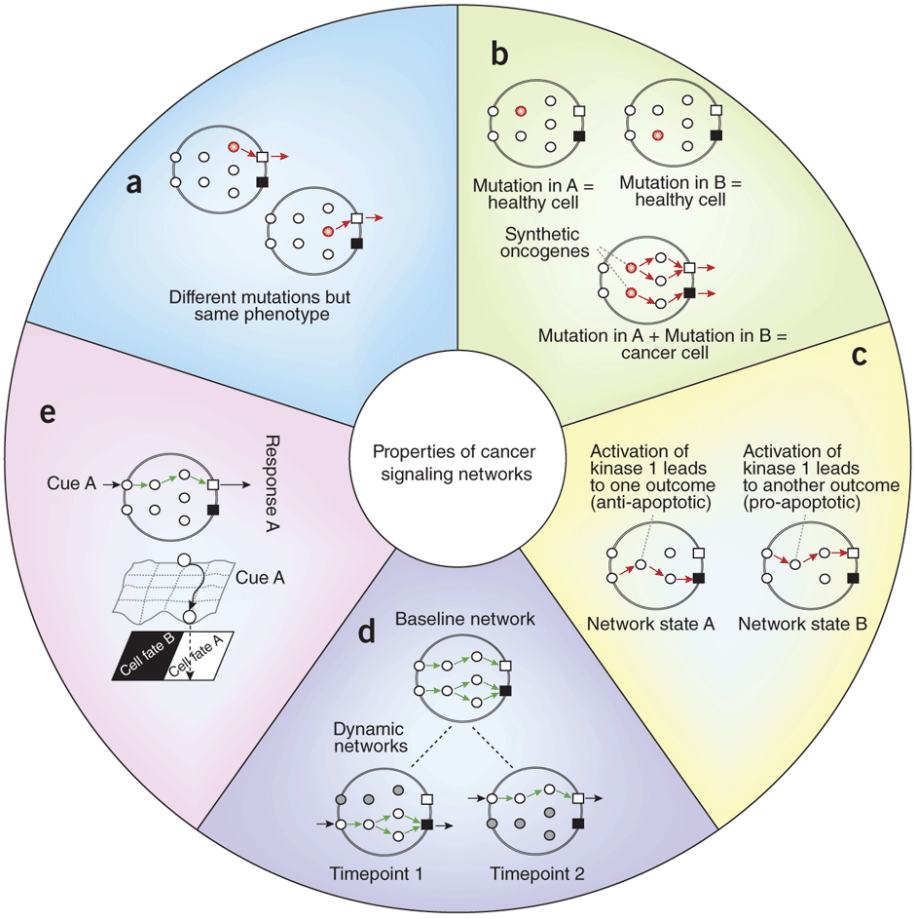

Fig 2.16: Properties of cancer signalling network

(Figure adapted from Creixell et al., 2012)

There are immense prospects for targeting PPIs. There are prevailing authenticated PPIs which are dynamic focus for the development of therapeutic procedures. Furthermore, novel conceptions and capable PPIs are emerging for anticancer medicine development.

The augmented information of cancer genomics and PPI arbitrated epigenetic procedures and interpretation of unambiguous cancer onco-fusion proteins revealed an enormous quantity of novel PPIs that are allied with cancer pathology. Contemporary perception of the significances of several cancer treatments offers unexpected PPIs as possible cancer objectives to augment the efficiency of treatments.

The Cancer Genome Atlas and The International Cancer Genome Consortium are large-scale genomics initiatives, which steered the discovery of surfeit genomic vacillations that initiate generation and progression of tumor (Vogelstein *et al.,* 2013). Large-scale experiments were executed to investigate methodically the transformations in cancer to determine cancer-related PPI complex maps (Cancer Genome Atlas Research, N. 2008). These studies and the prophesied novel PPIs unveiled new PPIs that perform as most important navigators of cancer. Hence, they are the prospective objectives for therapeutic investigation (Zhang *et al*., 2012).

The cancer genomics authenticated not only the prominence of conventional trademarks of cancer but also unveiled new features that are complexly allied to cancer, such as RNA splicing and epigenetic dysregulation (Garraway *et al.,* 2013). Novel opportunities are offered by modern developments in investigations delineating the support of dysregulated epigenetic mechanisms to cancer for PPI targeting. For example, the dysregulated histone acetylation and methylation are allied

with generation of tumor cells. This alteration in turn instructs particular identification of altered histones by methyl lysine-coupling proteins and by acetyl lysine- coupling domains (Kelly *et al*., 2010). Cancer related transmutations in the RNA-splicing technology specifies the significance of PPIs in the regimenting RNA processing in cancer.

PPIs are also imperative for the catalytic functions of many enzymes comprising epigenetic- transforming enzymes susceptible to targets. Fusion proteins deals with cancer sensitive targets. Hence, there is a requirement of aiming at onco-fusion-protein explicit PPIs (Daigle *et al*., 2011).

2.15 PPIs and protein complexes

PPIs frequently encompass multi-protein networks for hub proteins that arbitrate signaling of oncogenes. The major challenge lies in the discerning restriction of a specific PPI in the network for an anticipated healing effect. One more challenge is the investigational recognition of discerning agents. The identification of the particular modulators using advanced methodologies hasten the expansion of selective PPI inhibitors.

One of the mounting prospect for target PPI in cancer is revamped PPIs in signaling networks of oncogenes initiated by therapeutic means. The therapeutic means prompted PPIs harvest novel cancer dependency and assist as novel objectives to win over the pharmacologically convinced medicine resistance. PPI variation is anticipated to own a significant role in forthcoming mechanism-based combination treatments.

2.16 Hub proteins

All the proteins interact with other proteins more or less. The interaction can be strong, weak, stable, momentary or provisional (Hartwell *et al*., 1999). There occurs a dearth of biological function if the proteins do

not interact. Protein interaction network encompasses highly connected a well as poorly connected proteins (Batada *et al.*, 2006). Most proteins bind with just a fewer number of proteins. In contrast, a few proteins bind with a large number of other proteins. This type of comprehensively binding proteins are labelled as hub proteins (Jeong *et al.*, 2000). The confiscation of a hub protein is deliberated as perilous when matched with a non-hub protein. This phenomenon is labelled as the centrality-lethality rule (Jeong *et al.*, 2000). This rule is grounded on the architecture of the interaction network and is focal to decipher the network function. In a network, the centrality-lethality rule postulates that hub proteins incline to tally proteins that are vital and strong. There are investigations, which elucidate that hub proteins are physiologically significant, and evolution wise well-preserved compared non-hub proteins (Wuchty and Almaas, 2005). The degree of connectedness for hub proteins was delineated subjectively based on researcher's prerequisite regardless of the exceptional topological and functional implications. Hub proteins display eight or more interactions on a common note. Recent several sovereign studies elucidated that hub proteins are allied to disease instigating genes, including cancer.

The appropriate graph and superfluous graph are employed to exemplify the relationship of a node to the core graph to unravel the dilemma of distinguishing a protein system from the PPI networks. This technique enables the symbolization of tightly or loosely coupled node to a core graph. The Relevancy Judgment algorithm enables forecast of protein complexes from PPI networks. This algorithm also probes whether a node fits into a protein system by looking at the implication of core graph and nodes. The high-throughput accuracy of this algorithm is apparent from manifold investigations.

2.17 PPI databases

An imperative concern for the interpretation of the functional association of the human genome is the mining of data about the development of a protein complex and the associated role of a related PPI network. A database based on biological system is the assortment of information that is systematized which makes the contents effortlessly controllable, reachable, and rationalised. The activities related to the construction of a database are assemblage of information, which is effortlessly retrieved, and providing the customer constantly obtainable (Peri *et al.*, 2003). Currently, there are numerous databases available for information mining of these PPIs (Titz et al., 2004). The databases are categorized into the following (Rivas and Fontanillo (2010) :

- Primary: Primary database gathers information on the existing PPIs, which are discovered based on laboratory, that is, *In vivo* and *In vitro* techniques. These include the nucleotide sequences and three-dimensional structures.
- Secondary: Secondary database gathers data extracted from the analysis of primary data like the secondary structures and domain.
- Prediction: Prediction database includes all the predicted PPIs based on numerous modus operandi. For instance, the bio-molecular interaction network database (BIND) is developed on an extensive system that allows an elaborate depiction of the approach. In BIND, the PPI data were measured experimentally and included links directing to the concluding substantiation from the literature (Bader *et al.*, 2001).

In the existing information about PPI data, the focal emphasis is on the binding partners of proteins or in other words binary protein interactions. The scarcity of information on how the proteins form complexes, the networking among the complexes or the interconnectivity of protein complexes cannot be suspended. Even though the number of databases pertaining to the PPI are more, the comparative study on such databases are moderately low. Consequently, the partially available information on the human interactome restricts the application of usability in system biology. Therefore, it is important to integrate the various data available to fill the ambiguities in the human interactome. Few resources are available to study this kind of integration but the explanation on the protein complexes is feeble.

2.17.1 HPRD

Human Protein Reference Database or HPRD is an entity database that assimilates a mammon of information pertinent to the roles of human proteins in well-being and ailment. Information concerning numerous PPIs, variations in post-translation, associations of enzyme and substrate, relations of ailment, manifestation of tissue, and localization of subcellular components were mined from the collected work for an essential set of human proteins. Most of the data was attained manually by researchers who read and construed available documentation in the course of elucidation procedure.

HPRD owns a spontaneous enquiry network consenting apparent admittance to the entire characteristic of proteins. This is created by means of open source technologies and is liberally accessible at http://www.hprd.org to the scholarly community. This combined

bioinformatics platform is beneficial in labelling and extracting the huge amount of PPIs and variations.

HPRD supplements other PPI databases as well. Thousands of proteins were interpreted comprising numerous PPIs. Several PubMed links were provided to various arenas that orient a customer to the pertinent principal literature. However, an accurately mistake-free and wide-ranging databank is not possible deprived of the participation of the biomedical community. A comment button is provided for each molecule that enables to collect responses from users. These annotations from a user permit to resolve any inaccuracies and to apprise available information regarding interpreted proteins other than the enduring exertions to augment and apprise the information. Fig 2.17 depicts the HPRD database.

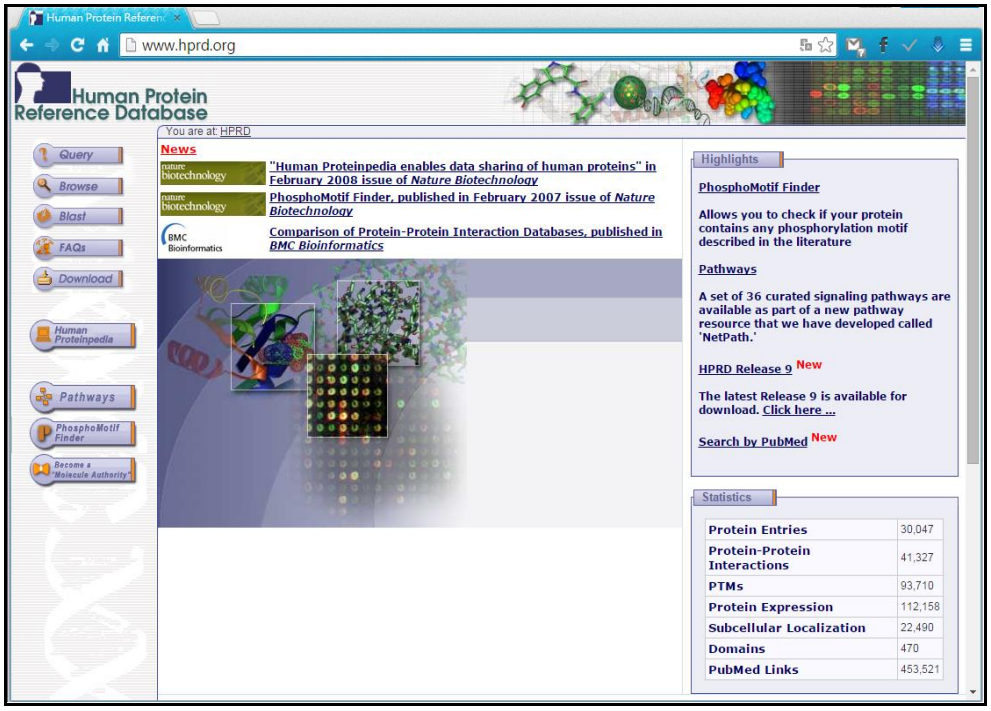

Fig 2.17: Schematic representation of HPRD database

(Figure adapted from BIOINFORMATICS.FR)

The integration of the widely accessible microarray information into HPRD enables a gene-centric interpretation to govern whether the mRNA expression array of a particular gene is conveyed to be transformed by any available investigation. Microarray users benefit from this comprehensive explanation to categorize proteins in numerous means to produce new theories or to filter the possible aspirants convoluted in a biological procedure. This database is a valued reserve for the proteomic community as a consequence of mainstream amendments of post translation which has a static molecular mass that permits accurate pursuits of the protein database.

2.17.2 IntAct

IntAct delivers a toolkit and an open source database to store, present and investigate PPIs. The web network offers documented and graphical depictions of PPIs, and consents discovery of the interaction networks. A web amenity countenances undeviating computational admittance to recover interaction complexes in XML layout. IntAct comprises numerous twofold and multifaceted interactions introduced from the theories and assist in association with the Swiss-Prot team and makes use of severely meticulous vocabularies to safeguard data reliability. All IntAct software, information and meticulous terminologies are accessible at http://www.ebi.ac.uk/intact.

Protein interactions deliver a valued reserve for the interpretation of cellular function, and PPI investigations to focus on the research for biomolecular components (Hermjakob *et al.,* 2004). Investigational methodologies like Y2H or tandem affinity purification permits the production of enormous amounts of information on PPIs. However, basically most of the ventures improve their individual organizations to

store, represent and analyze PPI information. Other than the replication of effort, this outcomes in a great amount of inconsistency amid diverse PPI sets of data. IntAct delivers a widespread, open source databank and investigation scheme for PPIs, which are locally mounted and amended to the requirements of the indigenous association, thus dropping expansion time, and endorsing reliability of data sets interaction by the usage of the equivalent substructure and explanation system.

The IntAct information prototype has the following three chief constituents:

- Experiment: An experiment assembles a number of interactions. This is generally from one journal, and catalogues the investigational circumstances in which these interactions are engendered. An experiment contains only one interaction, or numerous interactions in the incident of comprehensive experiments.
- Interactor: An interactor is a genetic object contributing to an interaction, which is generally a protein, but hypothetically also a sequence of DNA, or a trivial fragment.
- Interaction: An interaction encompasses one or more interactors contributing to the interaction. The depiction of interactions is not restricted to twofold interactions or information on multi-protein interactions.

Fig 2.18 depicts the IntAct database.

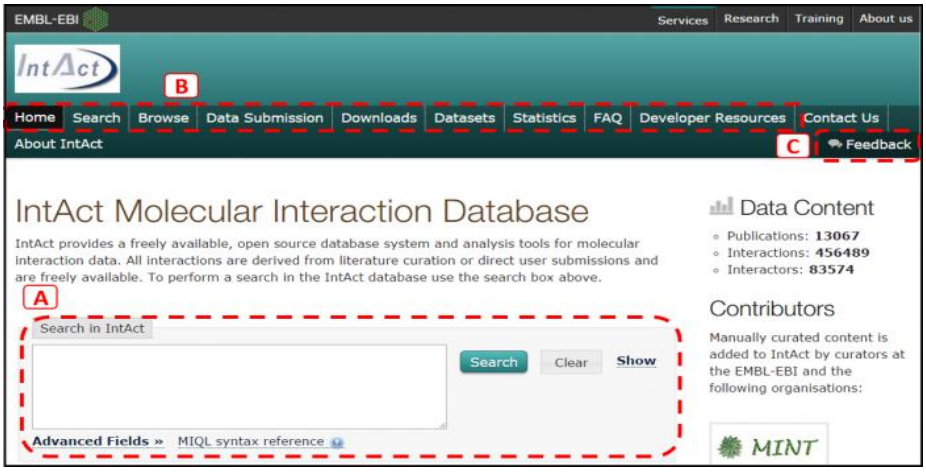

Fig 2.18: IntAct database

(Figure adapted from European Bioinformatics Institute, 2014)

IntAct presently delivers a modest search interface that pursuits the databank by IntAct accession number, by name, or by identifiers of external databases. For example GO.

The data recovered are demonstrated in two views, that is, an experiment view and a binary view. The binary view demonstrates the entire identified interaction associates and the nominal documented explanation for a particular protein. This swiftly provide suggestions of probable useful roles for non-characterized proteins. The number of subsidiary investigation is designated for every pair of interactions, and a linkage permits interchanging to the research view, which displays a comprehensive view of all the experimentations associated to the particular twofold interaction. Both the views provide the entire terms from hyperlinked meticulous vocabularies, providing direct admittance to their delineations. In addition, both views permit the choice of particular proteins and their respective demonstration in the graphical view. This view shows proteins in the context of their local interaction networks. Only the local interface neighborhood up to a specified distance is revealed for lucidity.

Nonetheless, there is option to enlarge the network globally, or only around particular proteins of concern. Any of the demonstrated proteins can be designated as the novel center of the interaction network. A distinct section of the graphical view demonstrates all the GO terms which are interpreted to proteins in the exposed interaction network. The entire proteins which have the annotated GO term or any of its child terms are highlighted by nominating any of these GO terms. This feature offers a rapid technique to discover the functional background of proteins interactively.

2.17.3 DIP

The Database of Interacting Proteins (DIP; http://dip.doe-mbi.ucla.edu) is a databank that documents experimentally generated PPIs. This databank is envisioned to afford the community of scientists with a complete and cohesive tool for surfing and competently mining data about PPIs and interaction networks involved in various biological processes. Other than registering the details of PPIs, DIP is also beneficial for comprehending the functions of protein and PPI associations, investigating the characteristic properties of networks of PPIs, benchmarking forecasts of PPIs, and analyzing the progression of PPIs.

DIP targets to assimilate the miscellaneous form of investigational data on networking proteins into a common and effortlessly retrieved databank. Countless scientific periodicals and documentations encompass biological data about PPIs. Even though these documentations are referred by the scientific communal day-to-day, recovering specific information from such reserves necessitates additional exertion compared to DIP. DIP syndicates data from manifold interpretations and investigational methods and also provides data on protein interacting networks.

The extraction and integration of the plethora of data about PPIs to a manageable milieu is the paramount intention of DIP. The databank of specific organisms such as EcoCyc for *Escherichia coli*, YPD (1) for yeast, and pathway databases like CNSB and KEGG, frequently encompass information data concerning the protein pathways and complexes. Similarly, DIP was found to complement the present database, and to consent researches to develop and complement the annotations of PPI of one organism with other organisms.

The data on PPI was stored in the DIP as a single text file in its novel commencement (Marcotte *et al.*, 1999). The DIP is now executed as an interpersonal database written in the SQL, explicitly mySQL (TcX Sweden) to tackle the growing capacity of data efficiently. SQL proficiently manages miscellaneous categories of information and permits fast cataloguing and investigation. The databank has the option to expediently extend as per the requirement, without shifting the content of the current database, by the addition of new tables and fields to the assembly of information.

A table of protein information, table describing details of experiments detecting the PPIs and a table of PPIs are the three linked tables of DIP.

2.17.6 MINT

The Molecular INTeraction database (MINT) at http://mint.bio.uniroma2.it/mint/ targets the storage of data in an organized format. The priority is the data on molecular interactions (MIs) by mining investigational particulars from peer-reviewed journals which are published. Currently, the MINT group emphases the effort on physical PPIs. Computationally or genetically concluded PPIs are not involved in

the databank. MINT has endured an intense restructuring of the database structure and information model and has vividly augmented the quantity of stowed PPIs in the past four years. The novel variety of MINT is grounded on an entirely modernized database structure, which deals with better competent information investigation and scrutiny, and is categorized by records with a more affluent explanation. The number of physical PPIs rose to over 95 000 in the past few years. The complete database serves free access online in both collaborating and batch modes over an FTP server and web-based interfaces. In addition, MINT also includes a dataset of human PPIs concluded from experimentations with orthologue proteins in prototypical organisms branded as HomoMINT (http://mint.bio.uniroma2.it/mint/).

Cells are multifaceted systems whose functioning is administered by a sophisticated network of molecular interactions (MIs) and has a pertinent subcategory are PPIs. Transcriptional regulation and signal transduction pathways and are the archetypal instances of biological procedures arbitrated by PPI. The MI database (MINT, http://mint.bio.uniroma2.it/mint/) was premeditated to assemble investigation substantiated PPIs in a binary or multifaceted demonstration.

MINT promoted the IntAct interpersonal prototype in January 2006. IntAct is an open source databank particularly intended for the storing, presenting and examining the MIs. The schema is available at http://intact.sourceforge.net/uml/intactCore.gif. The key benefit of espousing the IntAct prototype is its capability to embody protein complexes and the other categories of molecules as interaction partakers. Also, the easiness with which novel characteristic features and toolkits for storing, representing and analysis of data are added. In addition, MINT is

also attuned with the entire tools and upgrades developed by the IntAct consortium. MINT is grounded on the open source PostgreSQL database management system found at http://www.postgresql.org.

The entire data can be retrieved as Java objects using the IntAct API by means of OJB found at http://db.apache.org/ojb/ as the object-relational mapping tool. The web application is grounded on the Struts framework found at http://jakarta.apache.org/struts/ running on the Tomcat servlet container found at http://tomcat.apache.org/ and the Apache server at http://www.apache.org/.

The latest web-based interface has provided the researches with lot of expansions and augmentations permitting a better proficient databank assessment. One of the notable feature is that the query can be grounded on gene or protein names, identifiers of external databases or UniProt keywords like PDB ((Berman *et al*., 2007), UniProtKB, SGD, Ensembl, Reactome, PubMed, etc. There is option to enquire about explicit species datasets like mammalian, viruses or Drosophila melanogaster. BLAST, a sequence similarity search can also be accomplished for the pursuit of proteins which are homologous to the enquiry protein.

The inquiry results are shown in a table appearing in the left edge an outline of the protein highlights reported in Uni-Prot and, in the right edge, the rundown of the connection accomplices curated in MINT. The connections can be shown graphically by an upgraded variant of the 'MINT viewer', a Java applet resulting from the applet Graph (http://java.sun.com).

The viewer speaks to the associations by lines (edges) and nodes (proteins), and allots the nodes a size corresponding to the protein's atomic weight and a shading which relies on upon the species. The diagram showed by the viewer can be extended and altered intelligently by moving

or erasing nodes. Proteins connected to OMIM are presently highlighted in red. Keeping in mind the end goal to ascribe an unwavering quality list to the reported collaborations, we have likewise doled out every cooperation a certainty level, in view of the test recognition strategy and test conditions. The aftereffects of the investigation performed in the MINT viewer can be caught in various arrangements prepared for export, that is, PSI2.5-XML, PSI1.0-XML, Osprey, or flatfile.

MINT is currently supplemented by HomoMINT, a surmised human protein collaboration system where communications found in model living beings and gathered in MINT are mapped onto the relating human orthologues. Through the MINT website pages it is additionally conceivable to hunt the HomoMINT dataset down gathered associations.

2.17.6 Gene ontology

The Gene Ontology or GO project at http://www.geneontology.org/ offers planned, organized vocabularies and categorizations that includes numerous areas of molecular and cellular biology. GO is freely accessible for unrestricted usage in the explanation of genes, products and sequences. Several prototypical organism databanks and genome explanation sets uses the GO and add their explanation sets to the GO reserve. The GO databank assimilates the vocabularies and contributed annotations. It offers complete admittance to this data in numerous formats. The members of the GO Consortium recurrently work communally, including external specialists as required, to increase and apprise the GO vocabularies.

The GO Web resource also offers admittance to widespread certification about the GO project and associates to solicitations that employ GO information for analysis of functional aspects.

The genomics era has witnessed the accretion of massive quantities of biological information, convoyed by the extensive propagation of biology-focused databanks. Diverse types of data from diverse resources are assimilated in ways that is logical to the biologists to mark the superlative usage of biological databases and the information they comprise. The foremost constituent of the integration work is the enlargement and usage of explanation criteria like ontologies. Ontologies deliver conceptualizations of information realms, which expedites communication among scholars and the usage of domain data by computers for manifold uses.

2.17.6 UNIPROT ID

The UniProt Consortium forms UniProt. This consortium is an association of the Swiss Institute of Bioinformatics (SIB), the European Bioinformatics Institute (EBI), and the Protein Information Resource (PIR). UniProt encompasses four constituents:

- The UniProt Knowledgebase (UniProtKB): This is a proficient and opulent protein database consisting of two sections:
    - UniProtKB/Swiss-Prot: Encompasses superior, manually explained and essential protein sequence records. Manual explanation entails investigation, evaluation and assimilation of all existing sequences for a particular protein. Also, a comprehensive review of concomitant investigational and prophesied information. UniProt team collect biological data from various records and execute various computational examinations. This maintains

in a distinct record comprising the diverse protein products arising from a specific gene. The Protein and the respective families or groups are reviewed on a regular basis to match with the latest scientific discoveries.

- UniProtKB/TrEMBL: Encompasses superior computationally investigated records augmented with programmed explanation and cataloguing.

- UniProt Reference Clusters (UniRef): UniRef100, UniRef90 and UniRef50 databases unify sequences spontaneously from different species. UniRef100 is grounded on the entire UniProtKB records. UniRef100 is created by grouping all these records based on identity of the sequence. UniRef90 and UniRef50 are constructed from UniRef100 to offer records with reciprocal sequence identity with ≥ 90% or ≥ 50% respectively. These databases are linked to the analogous UniProtKB records. In each cluster, the entire sequences are ranked to enable the choice of a typical sequence.
- UniProt Archive (UniParc): Captures all the protein sequence data available in public. This is the reason why UniParc is the best ample widely reachable essential database of protein sequence. Redundant data are the result of the presence of a protein sequence information more than once in a given database. UniParc addresses this problem. This stores the information only once and allocates an exclusive UniParc identifier.

- UniProt Metagenomic and Environmental Sequences (UniMES): The sequences derived directly from the ecological samples are stored in UniMES. The derived information is subjected to further analysis after conjoining with InterPro. This is a cohesive resource for protein families, domains and functional sites.

All the above-mentioned databases are augmented for different requirement. The UniProtKB/Swiss-Prot provides access to functional information of proteins. The UniProtKB contains the sequence of amino acid, name of the proteins, taxonomic data, etc.

The UniRef databases deliver grouped sequences from UniProtKB and particular UniParc records to offer a comprehensive analysis of sequence at numerous tenacities. UniRef90 and UniRef50 extends reduction of database size by 40% and 65% respectively. This provides a swift sequence searches. UniParc is the supreme publicly manageable essential protein sequence database, which provides link to all fundamental sources and varieties of these sequences. This helps to comprehend whether the queried sequence is already present in the public domain or any closest relative is present. UniMES is a storehouse explicitly designed for metagenomics and ecological information.

There are different options to retrieve information from the UniProt database. Browsing allows browsing and investigating information from www.uniprot.org. Download the entire databases, that is, the UniProtKB, UniRef and UniMES databases from www.uniprot.org/downloads. The complete set of UniProt Knowledgebase releases are circulated on CD-ROM.

2.18 Global characteristics of PPI complexes

The local and global characteristics of a network can be investigated using various contemporary methodologies. The global properties offers a synopsis of a specified network. However, the convoluted modifications between the networks is not facilitated. In contrast, the local properties computes trivial, local forms or sub-structures termed as graphlets or motifs. The foremost benefit of investigating the local properties is apparent in case of networks with lacking node and edge groups. The cause is that the resident structures are almost complete compared to the biased global properties.

2.18.1 Degree of centrality

In modern bioinformatics, the most substantial challenge is to develop computation tools to comprehend and heal complex ailments. For instance, cancer. Manifold methodologies are employed so far to determine the candidate of cancer genes. It is evident from various investigations that PPIs are responsible for almost all the biological processes occurring a cell. Therefore, an algorithm based on graph centrality values of the human PPI network will be a highlight to identify genes causing cancer. The precise and accurate inference obtained from this algorithm can turn out to be an in effect prototype for detecting novel cancer protein.

PPIs are ultimate to almost each cellular process. These proteins perform many roles including inactivating, altering the kinetic properties and formatting a novel interacting site of proteins. Numerous paramount milestones are marked in the past few decades to comprehend the PPIs and thus discovering more facts on the multifaceted biological system. Protein complexes executing a particular biological role regularly comprises of extremely linked protein modules. Investigations about these protein

modules portrays a vital part in comprehending the pathophysiological features of compound ailments like cancer.

Cancer is the result of unrestrained development of anomalous cells in the body. Over 200 varieties of cancer is identified so far. About 9 million cancer cases are detected every year. Over 4.5 million individuals expire from cancer each year globally. Initial finding of cancer increases the likelihoods of an efficacious cure and persistence to a superior extend. It is an exceptionally multifarious hereditary disorder and practically 5–10% of humanoid genes subsidize the beginning of cancer. Only 1% is recognized so far. A methodical investigation of the proteins coding cancer genes in a PPI network aids to ascertain novel candidate genes. The algorithm generally emphases on certain graph centrality values of a PPI network. For instance, degree, shortest path distance between two proteins, betweenness centrality, clustering coefficient, etc.

Generally, for the algorithm based studies, PPI data are accessed from public databases. The size of the PPI data are outsized. Hence, the data is composed randomly by selecting n cancer proteins and n non-cancer proteins from the PPI data. A subclass of interactions is engendered by choosing the entire interactions of the nominated proteins from available PPI data. The resulting data are embodied as an adjacency milieu and the measurement of innumerable graph centrality values are executed. The generation of ranks for all these proteins are carried out based on the centrality parameter. The final score of each protein is arrived from the discrete ranks of each protein. This procedure is iterated on mock-ups produced from another n cancer proteins and n non-cancer proteins. The final ranking is updated after every repetition. This procedure is iterated until the final score require no further update. In a graph, diverse centrality

values are proposed to determine the significance of a node. The nodes are arranged by computing five centrality processes.

Fig 2.19 depicts the schematic representation of the procedure with respect to a single repetition.

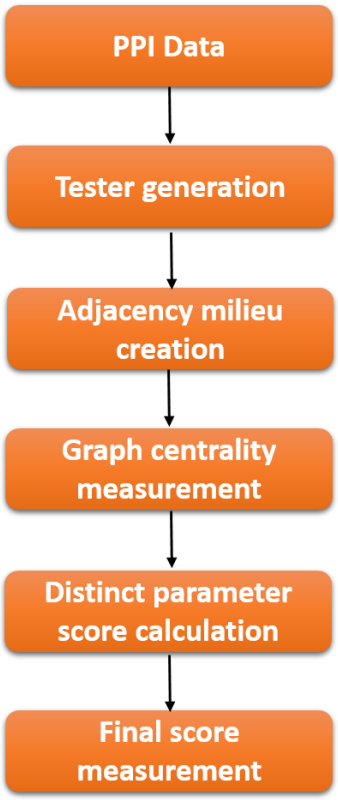

Fig 2.19: Schematic representation of the procedure with respect to a single repetition

2.18.2 Clustering coefficient

Various intricate data cliques own general depictions as networks. The briefing, associating, cataloguing and sculpting of these data sets across an extensive assortment of disciplines are imperative deeds occurring concurrently (Newman, 2003). Several amounts are calculated to

illustrate a network. For instance, the notions of path length, clustering coefficient, and degree proves to be tremendously beneficial.

Watts and Strogatz (1998) devised the catchphrase small world network to designate the frequently stirring condition where a meagre network is vastly grouped, that is, resembling a common lattice and hitherto owns shorter path lengths like that of a random graph. Many complex networks are investigated and categorized as small worlds. Likewise, the well-known scale free characteristic of the degree distribution is acknowledged as an assurance for several real data groups (Barabasi and Albert, 1999; Newman, 2003). The scale free and small world properties are extensively investigated for unidirectional or binary networks.

2.18.3 Neighbourhood connectivity

The node connectivity depends on the number of its neighbors to which a particular protein is connected. The connectivity of a neighborhood node n is well defined as the average connectivity of all its neighbors (Maslov *et al*., 2002). The distribution of neighborhood connectivity offers the average of the connectivity of all neighborhood nodes with neighbors. Fig 2.20 depicts the distribution of neighborhood connectivity for a network.

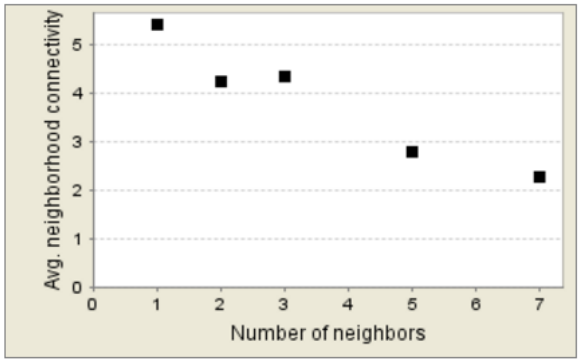

Fig 2.20: Distribution of neighbourhood connectivity for a network

(Figure adapted from Network Analyzer online help)

In equivalence to the in- and out-degree, each node n in a focused network is mapped to an in- and out-connectivity. Therefore, in focused networks, the following types of neighborhood connectivity in a node is identified:

- Only out: The average in-connectivity of all out-neighbors of n
- Only in: The average out-connectivity of all in-neighbors of n
- In and out: The average connectivity of all neighbors of n.

Three neighborhood connectivity distributions, that is, only out, only in, and in/out are identified grounded on the three definitions mentioned above. The edges between highly connected and little connected nodes are dominant in the network if the neighborhood connectivity distribution is a diminishing function in k.

2.18.4 Shortest path length

In a large scale-free network, calculating the average shortest-path length requires abundant memory space and calculation time. Therefore, analogous calculation is applied. In a network, the shortest paths computes the length of the entire shortest paths from or to the vertices. The length of the shortest path between two nodes a and b is L(a,b). The shortest path length distribution gives the number of node pairs (a,b) with L(a,b) = k, where k = 1,2,….

The network diameter is depicted as the maximum length of shortest paths between two nodes. In case of a disconnected network, the diameter is depicted as the maximum of all the diameters of its linked constituents.

The network diameter and the shortest path length dissemination indicates small-world properties of the analyzed network (Watts and Strogatz, 1998).

2.18.5 Pathway analysis

The genome-wide connotation investigations are extensively employed for detecting mutual genetic variations that subsidize human multifaceted characters with the expansion of high-throughput genotyping expertise. Concentrating on the utmost substantial single-nucleotide polymorphisms (SNP), genome-wide connotation investigations after analysis of specific SNPs efficaciously identified numerous SNPs concomitant with compound human ailments (Manolio *et al*., 2009).

On the other hand, in almost all the cases the recognized SNPs only mutually elucidate a trivial fragment of heritability. This leads to a challenge for identifying genetic variations with trivial or reasonable discrete reasons for human compound ailments. Furthermore, these procedures incline to own a petite reproducibility. Consequently, it results in a diminutive overlay among verdicts of diverse investigation sets examining the similar biological system.

The contemporary research validates that assessing gene expression disparities associated to predefined group of interrelated genes or pathways frequently upsurges the arithmetical supremacy and harvests further strong upshots (Virtaneva *et al.*, 2001). In recent times, quite a few pathway-based investigation procedures were suggested precisely for genome-wide connotation investigations (Wang *et al.*, 2010).These approaches vary from each other in several facets. This includes the methodology to assess the statistical implication, requirement of individual-level SNP genotypes, computation of gene-level summary statistics, etc. The emphasis of investigation from specific SNPs to pathways steered the recognition of several eloquent pathways in biological system.

Pathway analysis has turn out to be the foremost selection for acquisition of awareness of the principal biology of diverse gene expression and proteins. It diminishes intricacy and has amplified illustrative supremacy.

Nowadays, nearly all the bioinformatics investigations seek statistically substantial pathways to validate either computationally consequent outcomes or biological elucidation. Though broadly accepted, the first generation pathway analysis approaches that is, Over-Representation Analysis (ORA), decouples molecular computations from well-designed investigation and accept that pathways and genes are not dependent on each other. The second-generation methodologies like Functional Class Scoring (FCS) resolves these precincts. The methodologies based on Pathway Topology (PT) further enhances the FCS methods by making an allowance for the type and number of interactions between genes. This is generally overlooked by FCS.

Conversely, in spite of these exertions, there are unresolved explanation and procedural defies. The resolution data is less, incomplete conditional and cell-specific data, and inadequate explanations limit expansion of the next-generation methodologies for pathway analysis. The incompetence to assimilate the vibrant environment of a biological organization in investigation restricts the efficacy of prevailing techniques. Nevertheless, in spite of these steeplechases, as the amount and kind of useful observations upsurge, combined with technical developments and investigation approaches that deliver enhanced supervision for tactical forecasting for consequent biological experimentations, the usefulness of pathway analysis and buoyancy in the inferences are expected to expand.

2.19 PPI network topology analysis

The other set of method available for investigating PPI networks is to appraise the network topology. This method is rarely employed to investigate the low-density networks or hub proteins. This method is mostly useful for high-density networks. In this case, the proteins might be important fraction of a proteome, or to investigate a specific function in a cell or a process. PPI network topology is usually demarcated by dimensions as follows:

2.19.1 Node degree

This represents the number of nodes that interact with a particular node. In other words, the degree or connectivity of a node is the total number of edges incident with a particular node.

2.19.2 Degree distributions

The rudimentary topological property of a protein is the degree or connectivity. The degree is referred to as the number of links or connections a particular protein has with respect to other proteins. The proteins with the higher degree are designated as hubs. They are obligatory for the endurance of a cell. The degree distribution proposes the number of interacting partners with respect to a selected protein.

The degree of a protein node signifies the sum of the direct links of this node in a protein network. P(n) is the possibility that a node has P links in a group of proteins grounded on the dissemination. The number of links on a node specifies that the network firmness is comparatively high when compared to the lower number of links (Zhu *et al*., 2001).

For example, in the network displayed by the figure, hypothetical node A has a degree of 4 (kA = 4). The average degree of nodes for the

whole network (<k>) is used as an index to describe the 'density' of a network.

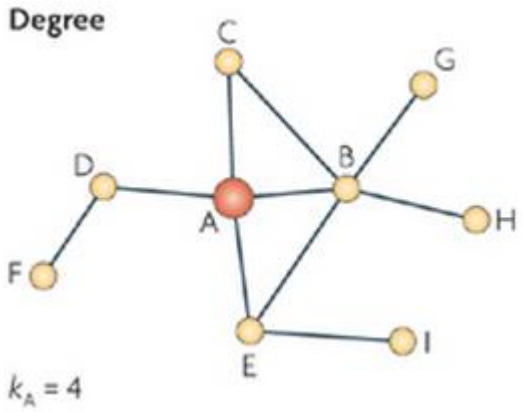

Fig 2.21: Degree distributions

2.19.3 Correlations

The real PPI networks are regarded as correlations in the degrees of a node. This is accomplished by the comparatively short paths between any two nodes. Also, by the manifestation of an enormous number of short cycles.

2.20 Human cancer and non-cancer interaction

A regular epithelial cell controls the discharge of autocrine and paracrine elements that inhibit abnormal development of adjacent cells paying way to vigorous expansion and standard metabolic rate. One factor accountable for the commencement of cancer is deliberated as the downfall of this homeostatic cell competitive system (Vogelstein *et al.*, 2013). The cancer- suppressive microRNAs (miRNAs) are veiled by regular cells as anti-extensive signal units. The result of varied analysis at global level indicate that secretory tumor-suppressive miRNAs endorse as a demise signal in a cell competitive procedure.

2.20.1 cBio

The framework of a genome sequence delivered by the project of human genome permits accurate mapping of variations in human genome to investigate their relationship with ailment and the respective effects on the functional aspect of genes. Somatic mutations and germ-line variations were recognized as imperative in cancer with respect to the portrayal of oncogenes and tumor suppressor genes some decades ago in terms of jeopardy, forecast, reaction and oncogenesis to therapy. The portal for cancer genomics cBio at http://cbioportal.org is publically available reserve for collaborative investigation of multidimensional cancer genomics information set. At present, 5000 tumor tester information is accessible from 20 malignancy investigations. The cBio portal ominously depresses the obstacles among compound genomic information and cancer scholars who require swift, spontaneous, and superior right of entry to profiles at molecular level and medical traits from comprehensive cancer genomics ventures and endows scholars to decode the available opulent information groups into biologic acumens and medical solicitations (Cancer Genome Atlas Research Network, 2008).

All types of gene level data are stored and then joined with existing unknown medical facts like complete existence and aseptic survival intermissions. The information is systematized as a patient and gene function. The rudimentary notion of the portal is the perception of transformed genes. Explicitly, a gene is classified as transformed in a particular patient with respect to deletion, mutation, amplification, or the comparative expression of mRNA is lower than or superior than a user-defined inception. The concept of transformed genes is an authoritative streamlining conception that empowers consumers to scrutinize compound

information groups and to improve the biologic assumptions concerning persistently transformed gene groups and pathways in a biologic system.

A strategic distinguishing aspect of cBio is the comfort of use. All characteristics of cBio are thus accessible over a rationalized 4-step interface on web. In detail, the customers are channeled to select the following:

- A study on concerned cancer.
- More than one or just one profile of the genome.
- A case study of the effected patient.
- A concerned set of gene.

Customers have the preference to calculate spontaneously the common exceptionality and co-existence among the entire gene duos. Also, the choice to accomplish cross-cancer enquiries by using a modest 2-step inquiry after selecting All Cancer Studies and entering the interested group of genes. Fig 2.22 depicts cBio cancer genomics portal.

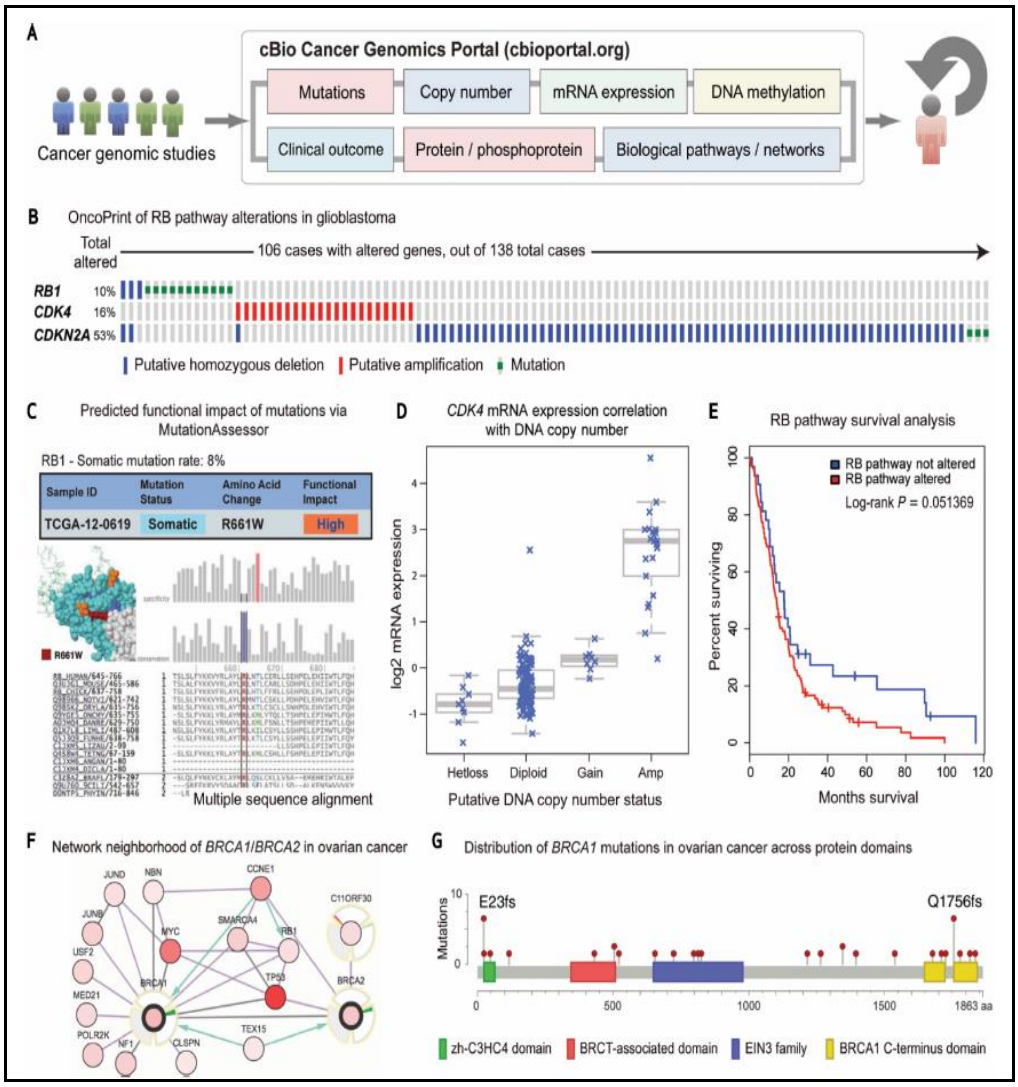

Fig 2.22: cBio Cancer Genomics Portal

A, the cBio Cancer Genomics Portal is an open platform for interactively exploring multidimensional cancer genomics data sets in the context of clinical data and biologic pathways. B, OncoPrint of RB pathway alterations in GBM. Genomic alterations of different members in the RB pathway are mutually exclusive. The OncoPrint provides an overview of genomic alterations (legend) in particular genes (rows) affecting particular individual samples (columns). C, mutation details for RB1. The predicted functional impact of the RB1 missense

mutations in GBM can be assessed through Mutation Assessor. This includes a predicted functional impact score, a multiple sequence alignment of different family members, and a 3-dimensional structure view, when available. D, correlation plot for CDK4. GBM samples with CDK4 amplification have markedly increased CDK4 mRNA expression. E, survival analysis. GBM cases with an RB pathway alteration have worse overall survival than cases without an RB pathway alteration. F, network view of the BRCA1/BRCA2 neighbourhood in serous ovarian cancer. BRCA1 and BRCA2 are seed genes (indicated with thick border), and all other genes are automatically identified as altered in ovarian cancer. Multidimensional genomic details are shown for BRCA2 and C11orf30/EMSY. Darker red indicates increased frequency of alteration (defined by mutation, copy number amplification, or homozygous deletion) in ovarian cancer. G, distribution of BRCA1 mutations in ovarian cancer across protein domains. The 2 hot spots (p.E23fs and p.Q1756fs) represent the common founder mutations 185delAG and 5382insC frequently observed in BRCA1. (Figure adapted from Cancer Discovery, 2012).

2.20.2 Sanger

Frederick Sanger established a technique to regulate the residue of amino acid situated on the N-terminal position of a polypeptide chain by with the chemical agent fluorodinitrobenzene. Initially, it was anticipated that this technique only delivers the sequences situated on the N-terminal. Sanger took the investigation further by employing fractional hydrolysis, numerous proteolytic enzymes, and primary variety of chromatography. In molecular biology, a lot of exhilaration resulted when Sanger revealed the technique to categorize proteins initially. The preliminary curiosity in Bioinformatics thrusted by the inevitability to construct databanks of the biological sequences. Sanger Institute Catalogue of Somatic Mutations in Cancer (COSMIC) is an enduring exertion supported by the Sanger

Institute to gather transmutation information from the systematic works (Forbes *et al.*, 2005). As part of this literature, it comprises information on thousands of discrete transmutations over different genes from capacious samples.

2.20.3 Cytoscape software

Cytoscape is employed to visualize biological cascades and molecular interaction networks. In addition, to integrate these networks with interpretations, gene expression profiles, etc. Cytoscape was initially intended for biological research but now it serves as a common place for analysis of complex network and visualization. The fundamental dissemination of Cytoscape delivers an elementary group of features for integration, investigation and conception of data. The other surplus characteristic features are accessible as Apps, which are formerly designated as Plugins. The Apps are existing for investigates of molecular and complex profiling, novel designs, maintenance of supplementary file format, scripting, etc. Anyone can develop by employing the Cytoscape open API. Almost, all the Apps are freely accessible from Cytoscape App Store.

Cytoscape supports the following standard and annotation file formats:

- Excel Workbook (.xls, .xlsx)
- Delimited text
- GraphML
- PSI-MI Level 1 and 2.5
- BioPAX
- SBML

- Extensible graph markup and modelling language (XGMML)
- Graph Markup Language (GML or .gml format)
- Nested network format (NNF or .nnf format)
- Simple interaction file (SIF or .sif format)

It also supports the MS Excel™ Workbook and Delimited text files. There are options to import data files including the GO annotations or expression profiles extracted from other applications or programs. There is a provision to load and save random traits on nodes, edges, and networks. For instance, generate a group of confidence values for particular PPIs by adding a set of annotation terms for a query protein.

2.20.3.1 Excel Workbook (.xls, .xlsx) and Delimited text

Cytoscape has innate provision for Microsoft Excel files, that is, .xls, and .xlsx. In addition, the delimited text files. These files encompass tables with information on network and qualities related to edges. The customer has the rights to stipulate columns containing target nodes, source nodes, edge attributes and interaction types in the course of file import. Igraph at http://cneurocvs.rmki.kfki.hu/igraph/ is one of the network analysis tools, which has the ability to export graph as a modest text file. Cytoscape in turn has the ability to read these text files and build networks from that data.

2.20.3.2 GraphML

GraphML is a wide-ranging and simple to handle file format for graphs. It is centered on XML. The comprehensive group of documents in this format is accessible at http://graphml.graphdrawing.org/

2.20.3.3 PSI-MI Level 1 and 2.5

The PSI-MI format is an information altercation layout for PPIs. It is based on XML format and employed to designate PPI and the related data. PSI-MI XML format design is accessible at http://psidev.sourceforge.net/mi/xml/doc/user/

2.20.3.4 Biological Pathways eXchange (BioPAX)

BioPAX is a Web Ontology Language or OWL document. This is aimed to exchange the information related to biological pathways. The comprehensive set of documents in this format is accessible at http://www.biopax.org/.

2.20.3.5 Systems Biology Markup Language (SBML)

The Systems Biology Markup Language or SBML is based on XML format and is employed to illustrate the biochemical networks. The SBML file format description is accessible at http://sbml.org/documents/.

2.20.3.6 Extensible graph markup and modelling language (XGMML)

Extensible graph markup and modelling language or XGMML is the XML advancement of GML. The GML delineates this file format. Other than the network data, XGMML comprises the aspects related to node, edge, and network. The XGMML file format design is accessible at http://cgi5.cs.rpi.edu/research/groups/pb/punin/public_html/XGMML/.

The advantage of using XGMML compared to GML is that it provides the manipulability allied with all XML document kinds. If you are unsure about which format to select, choose XGMML.

A java system property “cytoscape.xgmml.repair.bare.ampersands” must be set to "true" if you experience any trouble in reading older files.

2.20.3.7 Graph Markup Language (GML or .gml format)

GML is an opulent graph format language compared to SIF. Several other visualization packages enhances this file format. The GML file format design is available at http://www.infosun.fmi.uni-passau.de/Graphlet/GML/.

2.20.3.8 Nested network format (NNF or .nnf format)

The NNF format is the simplest format. Compared to SIF, NNF permits the discretionary consignment of single nested network for one node. It is impossible to specify any other qualities of a node. NNF supports only the following two line formats:

- A node confined to a network: **network node.**
- 2 nodes linked together in a network: **network node1 interaction node2.**

If a network name is observed as a node name formerly in any other columns, then the network is nested in the node with the same name. If a name is formerly demarcated as a network and later appears again as a node name, then the previously defined network is nested in the node with the same name. Conclusion is that, whenever a name is used in a network name and a node name, this infers that the network is nested in the node of the same name.

2.20.3.9 Simple interaction file (SIF or .sif format)

A simple interaction format is suitable for constructing a graph from a group of interactions. It also enables less effort to syndicate diverse interaction groups into a bigger network. Also, add novel interactions to a prevailing information set. The foremost drawback is that this format do

not comprise any data on the layout. This empowers Cytoscape to reproduce a novel layout of the network every time it is uploaded.

2.21 Associated work

A protein complex or module is well-defined as a group of proteins which are linked by single or multiple genetic or cellular interactions. Or modules can be labelled as a set of cellular constituents and their interactions contributes to a particular biological function (Hartwell *et al*., 1999). Therefore, it is essential to recognize these complexes of PPIs to augment the existing knowledge of human PPI networks organization. Specially, this can lead to the comprehension of the functional aspect of an unknown protein, by relating it with the known protein function. Till now, several investigations were executed to detect the segmental association in numerous biological networks.

PPIs are vital for the immense mainstream cellular procedures. This fundamental aspect has steered demanding investigations of extensive matchings of PPI networks. In preceding period, quite a few stratagems to label the humanoid interactome were projected and chased (Lehner and Fraser *et al*., 2004). These approaches can be allocated to literature-based, high-results yielding yeast-two-hybrid-based (Y2H), orthology-based, or mass spectrometry-based interaction maps type. All approaches have their own pros and cons. Nevertheless, the information on the consequential interaction maps are affected is not very distinct. Simultaneously, pioneer challenges in therapeutic and biological investigation to methodically use interaction information groups were carried out (Goh *et al*., 2007). Even though the consequences were encouraging, the fact cannot be repudiated that the amount of efficacious exertions to explore the PPI maps are

inadequate. An imperative motive for this state could be the mislaid assimilation of segregated maps. Evidently, it directed instantaneously initiate the confederation of previously detached interaction maps. On the other hand, there were many evidences to prove that dependability and excellence of miscellaneous PPI maps must be evaluated rigidly, particularly if their approaches of extraction were different. A judgement was hence appropriate as exertions in the direction of mutual tuning and apprising of presently detached PPI databases were as anticipated. Although such assimilation enables the information admittance for scholars, it also had chances of probable prejudices of the different mapping methods in distinct databanks.

The study on relative evaluations of PPI maps were previously executed for S. cerevisiae concerning the overlay, exposure and dependability (Bader and Hogue, 2002). Mering *et al.*, executed a relative investigation to assess the exactitude and to detect preconceptions, pros and cons of all the techniques employed for engendering yeast PPI information. Their investigation specified that existing PPI information are extremely conflicting, primarily owing to the incidence of vast false-positive frequency, and quite a few approaches had collection and recognition prejudices complementing the approaches. Bader *et al.*, 2002 investigated that the diminutive overlay could be the result of a vast negative-discovery frequency or vast false-negative rate.

A modest comparison of the outcomes from studies related to yeast-human maps are ambiguous concerning the diverse fundamental mapping and biological methodologies. Hence, a methodical assessment of existing human PPIs maps is necessary to get an enhanced perception into the topological structure and functional configuration.

An ensuing delinquent of human PPI networks is their distribution over manifold sites. Researchers must execute recurring examinations in numerous databanks to explore ample data on human proteins of choice. These exertions are palpably timewasting as many enquiry set-ups and identifiers must be employed in diverse databanks. One of the chief restriction in existing PPI databank is the usage of single protein interactions at a time for querying. Conversely, contemporary system biology necessitates compound network-oriented exploration for PPIs of manifold proteins.

Superior PPI networks are indispensable for the biomedical investigation (de Silva *et al*., 2006). But, the existing extensive human PPI networks are relatively inept (Chaurasia *et al*., 2006). Numerous assurance counting patterns were established to confront this difficulty (Li *et al*., 2008). In addition, some of the existing PPI networks offer individual assurance notching patterns (Rual *et al*., 2005). Incorporation of PPI maps with these assurance notch helps researchers to evaluate the superiority of PPIs present in the databanks.

Additional defies are the consistent apprises and augmentation of PPI databases. The data will increase unremittingly as the study on human interactome is still incomplete. Hence, it becomes vital to design a flexible architecture that constantly updates the existing interaction data, and also allows easy annexation of freshly exposed PPIs yet to be revealed.

Concluding, nonetheless an imperative concern is the logical elucidation of PPI maps. As a result of the complexity of PPI networks, researchers face defies even though there were advances in new genome-wide interactome projects. In order to comprehend the complexity, it is essential to understand the biological processes and physical interaction to

achieve eloquent data in the framework of physiological systems. PPI networks must be incorporated with other functional information to develop the extensive data from them. Earlier investigations also led to the incorporation of PPI networks with pathway or manifestation information which in turn directs to the categorization of biological processes or prospective ailment transformers (Oti *et al*., 2006).

Current progresses in result-oriented methods allowing the complete investigations of PPI networks caused outsized, extremely linked networks. Nevertheless, these networks are just the stagnant depiction of the compound networks befalling inside the cell. This approach do not offer the prospect to investigate the intricacies and subtleties of pathways related to ailments (Barabasi and Oltvai, 2004). One way to identify the pertinent local networks and decode the functional complexes is to assimilate the PPI network with extra data such as localization, manifestation, or hereditary information (de Lichtenberg *et al*., 2005). Calvano and associates assimilated the transcription sketching information with the PPI networks to depict the endotoxin reactions in human blood leukocytes depending on time (Calvano *et al*., 2005). In recent times, Pujana *et al.*, executed an analogous stratagem (Pujana *et al*., 2007). In humans, they joined the gene expression profiling with the functional proteomic and genomic data from numerous species to engender breast cancer allied network. Baranzini and co-workers incorporated the PPI maps with genome-wide single node polymorphism (SNP) markers information to identify the sub-networks associated with multiple sclerosis (Baranzini *et al*., 2009).

The information on PPI are of unlimited prospect in the field of biomedical research. New developments using high-throughput techniques caused a swift accretion of the human PPI networks on a genome-wide

level globally. There are manifold challenges to overcome before these PPI networks turns out to be a keystone in therapeutic investigation.

2.22 Existing challenges

One of the intimidating responsibilities of proteomics is to register the whole PPI networks occurring inside a cell. Even though, the existence of manifold outsized genomic structures and developments in result yielding techniques provides a base to build PPI maps, the interactomes of several entities are unfinished. A key delinquent of existing methodologies is that they are not capable to depict interactions in an inclusive mode (Hart *et al*., 2006). In one contemporary study, the researchers appealed to generate an enhanced variety of the existing high-throughput techniques to recognize the yeast PPI network, designated as generation-2. Hitherto, the reportage of all the probable interactions in S. cerevisae reached 20% approximately (Yu *et al*., 2008).

Analogous hitches are also present in human PPI. Modern investigations depicted that the existing human PPI maps are inadequate and extremely imperfect. For example, HPRD, a literature based databank for human PPI maps (Prasad *et al*., 2009) reports only around 5% of the entire interactome (Stumpf *et al*., 2008).

Another grave concern is the quality of the existing PPI data. The information generated comprises increased frequency of negative false positive interactions, though a voluminous information set exists with respect to PPI as a result of manifold researches. In addition, these techniques also has many investigational prejudices concerning few types of proteins and cellular localizations. All these challenges stresses on the augmentation in high throughput techniques. The extent to which the interaction maps are inclined by the selection of mapping stratagem is

imprecise. Therefore, it is critical to evaluate and relate the eminence and dependability of the generated maps. The proportional investigation of PPI maps in lower eukaryotes displayed an unexpected deviation among diverse interaction maps (Mrowka *et al*., 2001). Human interaction maps are also not any exclusion. The evaluation is still deficient for human protein in spite of their anticipated prominence in the field of biomedical research. Therefore, grave assessment of the existing human PPI maps are obligatory concerning the technique elected to generate a protein network.

Chapter 3

MATERIALS AND METHODS

3.1 Development and mapping of human protein-protein interaction network (HPPIN)

3.1.1 Assortment of human proteins

The human proteins were collected from various reliable sources like NCBI literature survey, Uniprot database, Human protein atlas, and Plasma proteome database (Lane *et al*., 2011). All these databases employ standardized methods to identify proteins. The extracted proteins were assigned with a unique format to maintain consistency. The redundant proteins were removed from the dataset to align with noise removal.

The protein integration was carried by manually coded programs and validated by inbuilt excel programs. The final data set named as Human Proteins (HP) was derived.

All the extracted proteins were in dissimilar format as diverse databases use different formats based on their nomenclature standards. There were also a number of redundant and recurrent. A far reaching protein-driven ID mapping was executed to assign a unique ID to the proteins and to remove the redundant and recurrent. Table 3.1 displays the overview of human protein data sets used for the work execution.

Table 3.1: Overview of human protein data sets used for the work execution

DATA SETS
Uniprot
Human Protein Atlas
NCBI literature survey
Plasma proteome database

3.1.2 Assortment of PPI maps

PPIs are essential for the massive mainstream of cellular progressions. This fundamental part steered this rigorous research of extensive mappings of PPINs. In preceding era, quite a few approaches to record the human interactome were anticipated and chased (Prasad *et al.*, 2009; Ewing *et al.*, 2007; Stelzl *et al.*, 2005; Rual *et al.*, 2005; and Persico *et al.*, 2005).

Post the development of HP dataset, the primary binary PPIs were compared using HPRD, IntAct, DIP and MINT. Similar to the HP dataset extraction, human protein – protein interaction (HPPI) dataset extraction also had different format and standards in different databases. So, all the extracted PPIs were assigned a unique ID using ID mapping technology. Perl programme was developed to automatically change each protein with the specific Uniprot ID.

The PPI maps are organized by proteins and interactions. So, evaluations can either be executed with respect to proteins or to PPIs. The pair-wise evaluation of PPI maps (A, B) in respect to proteins encompassed, common proteins within the two maps were recognized. This describes the overlap $XAB = XA \cap XB$ where XA, B are the collections

of proteins in map A or B correspondingly. Consequently, the overlap is standardized with regard to the quantity of proteins in A and B ($XAAB = \frac{XAB}{XA}$; $XBAB = \frac{XAB}{PB}$). The average of $XAAB$ and $XBAB$ is denoted as the relative protein intersection among A and$B$. Table 3.2 displays the overview of human protein interaction data sets used for the work execution.

Table 3.2: Overview of human protein interaction data sets used for the work execution

Protein interaction data sets
HPRD
IntAct
DIP
MINT

All these methods of PPI intersection have a disadvantage. These processes evaluate coincidence of the pragmatic PPIs and lack information on missing interactions (Chaurasia *et al*., 2012). These unidentified interaction was validated by employing orthology-based approaches. The first orthology-based map was assembled from interactions predicted for human proteins (Lehner and Fraser, 2004).

Noise data removal was also employed to remove the redundant, recurrent and self-interaction proteins as they lack any relevant biological function. After assigning the integrated datasets with Uniprot ID, the unique model of the human interactome named as HPPN was derived.

The human interactome was plotted as a network using Cytoscape. Cytoscape is a tool grounded on graph theory. Graph theory is fundamentally used to distinguish communities in a network. This aids in assuming the size, hierarchies and the shortest pathway in a system. Graph theory is named so as it can be represented graphically. The graphical representation makes it easy to understand the properties of the network. Every vertex is a node and the edges are interactions. Numerous topological evaluations like connectivity, cluster coefficient, degree-distribution, and hub proteins were calculated and matched to inspect the network characteristic features of each map. Self-interactions were omitted in the graph-theoretical examination to circumvent artefacts and all evaluations were accomplished based on the prevalent linked graph for every map. The consequence of the outcomes was measured by comparing the two related network replicas. First was arbitrary graphs with the similar number of nodes and PPIs, but lacking preservation of the degree distribution. Second, the arbitrary graphs with maintenance of number of nodes and PPIs and also the degree distribution. The second method was employed in this study. These graphs were created employing the unique networks and reiterated arbitrary exchange of PPIs. The edge among node X and Y $(X-Y)$ and between A and B $(A-B)$ is transformed to $X-B$ and $Y-B$ (Maslov and Sneppen, 2002).

Human interactome network (HIN) was created by incorporating double protein interaction information from the HPPI dataset. Consequently, the proteins were represented as nodes and interactions were represented as edges. Fig 3.1 depicts the interactome development.

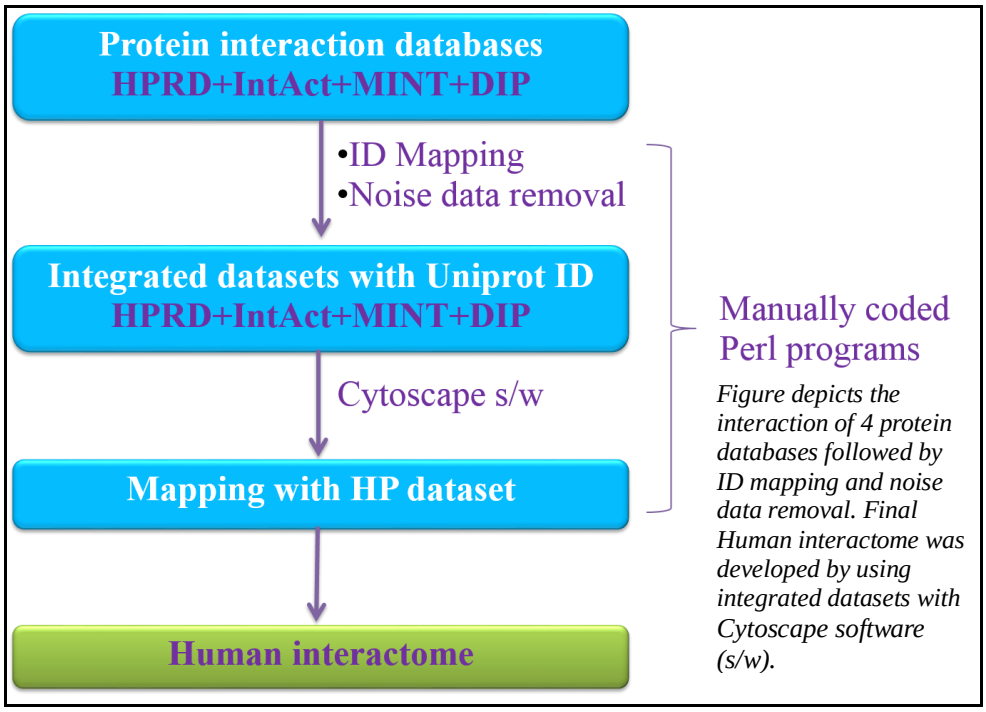

Fig 3.1: Flow chart for the interactome development

(Figure adapted from Anto et al., 2014)

3.2 Evaluation of cancer and non-cancer complexes in HIN

After extracting the unique HIN, cancer and non-cancer protein interaction were distinguished from HIN by commissioning the CBIO and Sanger database. As in the case of extracting protein dataset and PPI data set, the noise data removal and ID mapping of the subsequent information sets were accomplished by means of custom-made Perl programs (Yildirim *et al*., 2007]. The unique model of human interactome highlighted with cancer interaction network (CIN) was derived. In a binary interaction, if any one or both partner belongs to the cancer network then it is inferred to undergo cancer interaction. Table 3.3 displays the overview of cancer protein data sets used for the work execution.

Table 3.3: Overview of cancer protein data sets used for the work execution

Cancer data set
Sanger
cBio

Fig 3.2 depicts the mapping of cancer proteins in HIN.

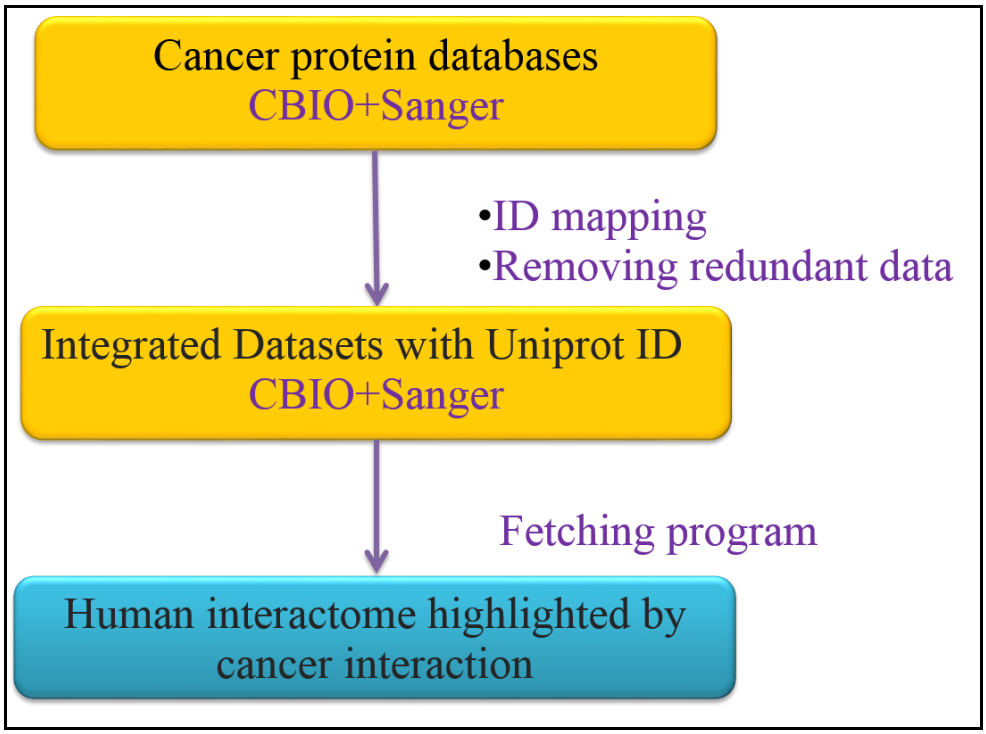

Fig 3.2: Mapping of cancer proteins in HIN

3.3 Mapping of major cancer and non-cancer (CANC) complexes in HIN

The human interactome from all the above mentioned investigation were assessed using the Molecular Complex Detection (MCODE) (Anto and Nambisan, 2014), and connected component algorithm (CCA) (Anto and Nambisan, 2014), which gave path to the expansion of cancer and non-

cancer complexes (CANC). CCA lists all its connected components of a disconnected network. The size or number of nodes of all the components are conveyed. The user can choose a particular connected component and can export the same as a distinct network in Cytoscape. This feature consents users to accomplish topological analysis on the prime connected component of a network. The network partitioning was observed by using CCA algorithm based on the similarity or distant values. The human interactome network compassing the cancer interactions was investigated using the CCA to remove all distracted interactions. The result was a highly connected human protein network in the human interactome.

Further, the MCODE was used to identify highly interconnected sub network as molecular complexes. The MCODE algorithm is a renowned automatic technique to discover extremely interrelated subgraphs as clusters or molecular complexes in a large PPI network. The selected complexes based on the MCODE value was then taken to analyze the protein interactions causing cancer.

MCODE instead uses a vertex-weighting scheme based on the clustering coefficient, Ci, which measures 'cliquishness' of the neighborhood of a vertex. Ci = 2n/ki(ki-1) where ki is the vertex size of the neighborhood of vertex i and n is the number of edges in the neighborhood (the immediate neighborhood density of v not including v). A clique is defined as a maximally connected graph. There is no standard graph theory definition of density, but definitions are normally based on the connectivity level of a graph. Density of a graph, G = (V,E), with number of vertices, |V|, and number of edges, |E|, is defined here as |E|; divided by the theoretical maximum number of edges possible for the graph, |E|max. For a graph with loops (an edge connecting back to its originating vertex), |E|max

= |V| (|V|+1)/2 and for a graph with no loops, |E|max = |V| (|V|-1)/2. So, density of G, DG = |E|/|E|max and is thus a real number ranging from 0.0 to 1.0.

The extremely unified and biologically important areas in a network is recognized using MCODE. The human interactome from both the analysis were evaluated using the CCA and MCODE algorithms (Anto and Nambisan, 2014), which culminated in the development of cancer and non-cancer complexes (CANC). Fig 3.3 depicts the CANC complexes in HIN.

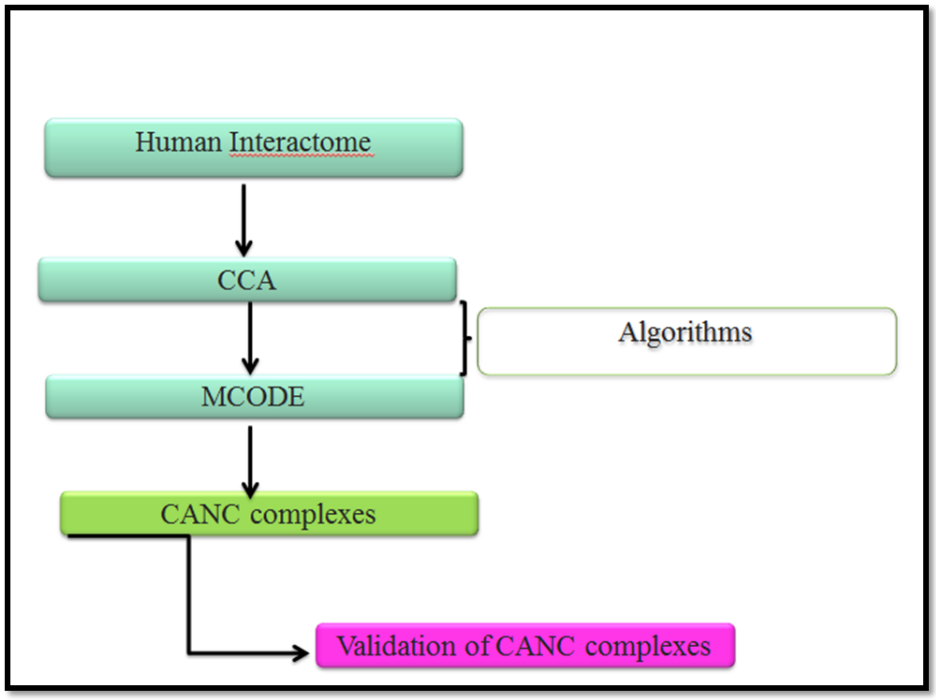

Figure 3.3: Flow chart for the CANC complexes in HIN

3.4 Validation of MCODE complexes

Gene Ontology was used for the validation of MCODE output. The data resulting from the GO was then passed through the statistical analysis for further validation.

3.4.1 Gene ontology analysis

GO was utilized to examine the useful conformation and consistency of PPI maps because currently it delivers the utmost complete efficient annotation for humanoid genetic composition (Ashburner *et al.*, 2000). GO comprises gene annotations concerning biological process (BP), molecular function (MF), and cellular component (CC) by employing a well-defined hierarchical ontology. Fig 3.4 depicts the validation of MCODE complexes with GO annotation.

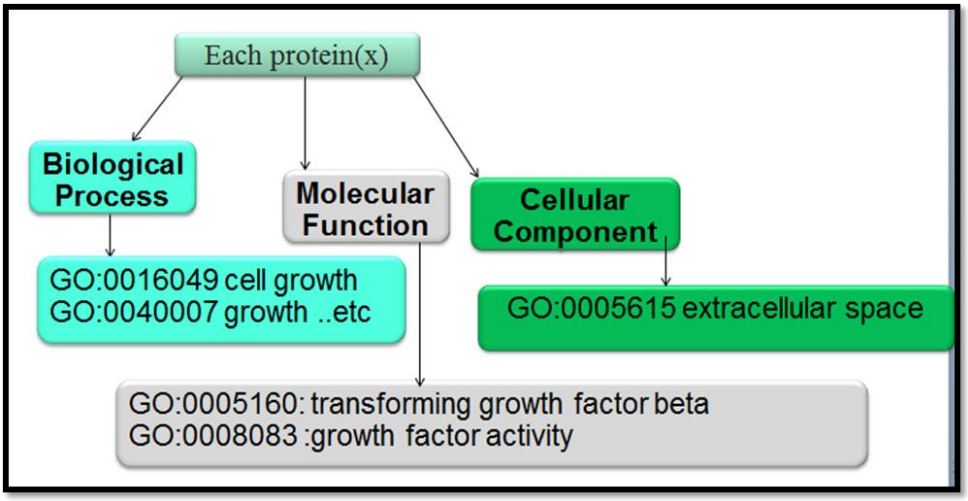

Fig 3.4: Validation of MCODE complexes with GO annotation

3.4.2 Hyper geometric distribution

Hyper geometric distribution assessment was employed to define the statistical implication which is specific to GO category that are over presented in a map. It will be ambiguous if you count just the proteins that share a common interpretation because the primary allocation of genes is not homogeneous. Therefore, P-values are employed to compute the statistical and biological implication of a collection of proteins (Asur *et al.*, 2007).

In the below equation, the size of the protein cluster is represented as n and m are the proteins sharing a specific interpretation. There are N proteins in the database with M proteins having same interpretation. The probability of m or more proteins out of n proteins associated with the equivalent GO is:

$$p - value = \sum_{i=m}^{n} \frac{(\binom{M}{i})(\binom{N-M}{n-i})}{\binom{N}{n}}$$

The proteins with lower P-values validates that the proteins are not grouped randomly and biologically significant compared to the proteins with higher$p - value$. There is a cut-off parameter employed to discern the relevant cluster of proteins from the irrelevant proteins. The protein cluster is irrelevant, if the $p - value$ is greater than the cut-off parameter.

The clustering score to compute the entire protein cluster is as represented by the below mentioned equation:

Clustering score = 1- $\frac{\sum_{i=1}^{ns} \min(pi) + (n1 * cutoff)}{(n_s + n_1) * cutoff}$

In the above mentioned equation, n_s and n_i denote the number of relevant and irrelevant protein clusters, and $min_{(pi)}$ denotes the lowest P-value of the relevant protein cluster represented by$i$. Therefore, each protein cluster is linked with one clustering score for all the three ontologies.

Link among the PPI maps and GO annotation was scrutinized based on the likeness of GO terms allocated to interacting proteins (Jansen *et al.*, 2003). The comparison of GO terms was computed by measuring their common path lengths contained by the GO tree. Analogous GO terms are anticipated to possess vast common paths. Arbitrary graphs of preserved

degree distribution were created to assess the implication. The dissemination acquired for the novel network was successively compared to the arbitrary networks acquired and computed log odds.

3.5 Hubs in the MCODE complex

Former investigation based on graph theory displayed that hubs, that is, nodes of high degree are mostly of critical prominence for scale-free network assembly (Albert *et al.*, 2000). The hubs are specified by extremely connected proteins in such PPI network. This has led to the inference that proteins rich with interaction are vital for accurate operation of cellular networks. First, it was examined whether hubs incline to be allocated to unambiguous functions, processes or locations using GO annotations to understand the prospective part played by hubs in human PPI networks. If the number of interaction was contained by the top 20% when compared to the analogous network, those Proteins was demarcated as hubs. Common inclinations were detected for orthology- and literature-derived networks.

There are two natural measures of difference of composition of nodes of a certain type (here, essential proteins) between two sets of nodes (here, hubs and non-hubs). One well known measure is the P-value for the Kolmogorov-Smirnov test for difference in distributions. If e1 is the fraction of essential proteins in the hub set and e2 is the fraction of essential proteins among non-hubs, then the Kolmogorov-Smirnov test is a test for inequality between distributions p1 ≡ { e1, 1−e1} and p2 ≡ { e2, 1−e2}. The measures are plotted in fig 3.5 as a function of number of high degree nodes included in the hub set.

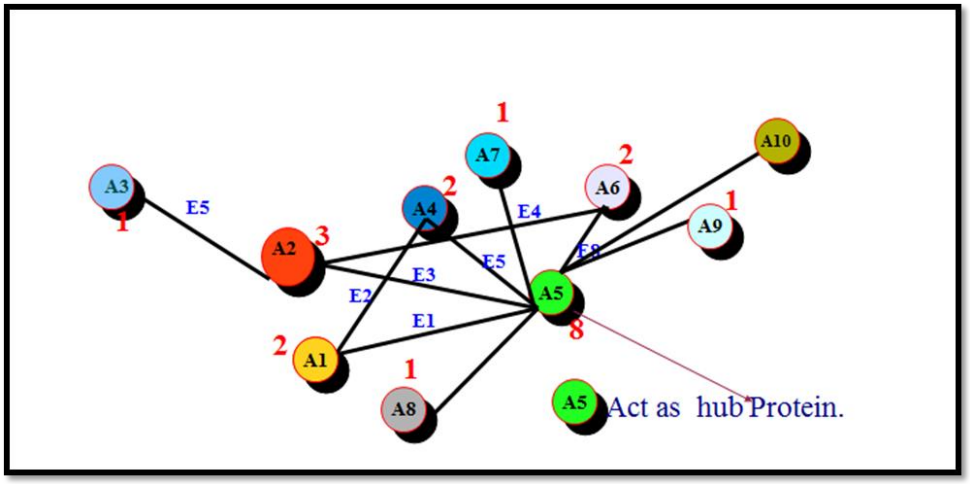

Fig 3.5: Hub in a network

3.6 Characterization of CANC complexes

The global and local topological properties of CANC complexes like degree distribution, assortativity, betweenness of centrality, clustering coefficient, shortest path length and pathway analysis were studied to validate the characteristics of hubs CANC complexes. The role of each node in a group of protein was evaluated.

3.6.1 Degree distribution

The degree of a protein node denoted the count of the direct links of this node in a protein network. $P(n)$ is the probability that a node has P links in the set of proteins based on the distribution. The number of links on a node indicated that the network stability is relatively high compared to the lower number of links (Albert *et al*., 2000).

For example, in the network shown in the fig 2.6, hypothetical node A has a degree of 8 (kA5 = 8). The average degree of nodes for the whole network (<k>) is used as an index to describe the 'density' of a network.

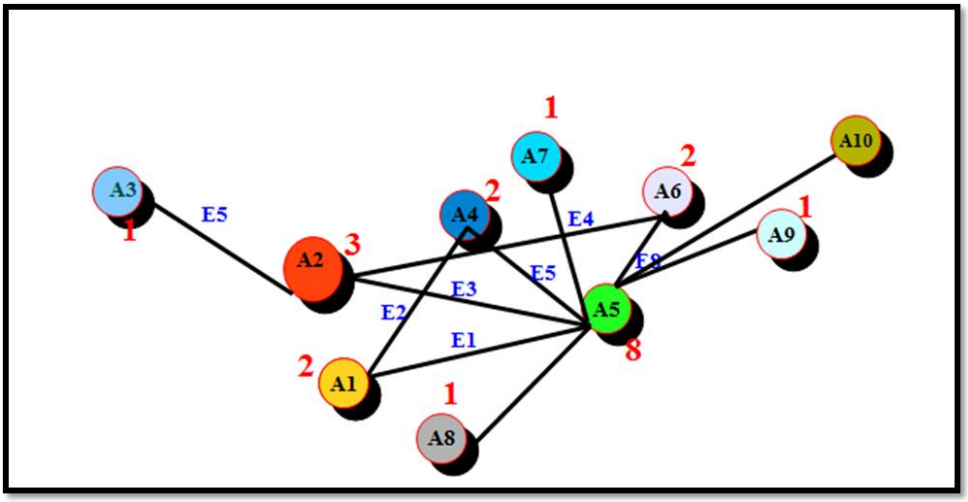

Fig 3.6: Degree distribution of node A5 =8

3.6.2 Betweenness of centrality

The betweenness of a node is the count of shortest paths between all probable pairs of nodes in the network that pass through the node. Betweenness computed the ways in which signals passed through the interaction network (Albert *et al*., 2000). From the below diagram betweenness(b) calculated as follows.

$Cb(n) = \sum s \neq n \neq t \ (\sigma st\ (n) / \sigma st)$,

where s and t are nodes in the network different from n, σst denotes the number of shortest paths from s to t, and σst (n) is the number of shortest paths from s to t that n lies on. Fig 3.7 depicts the betweenness for b.

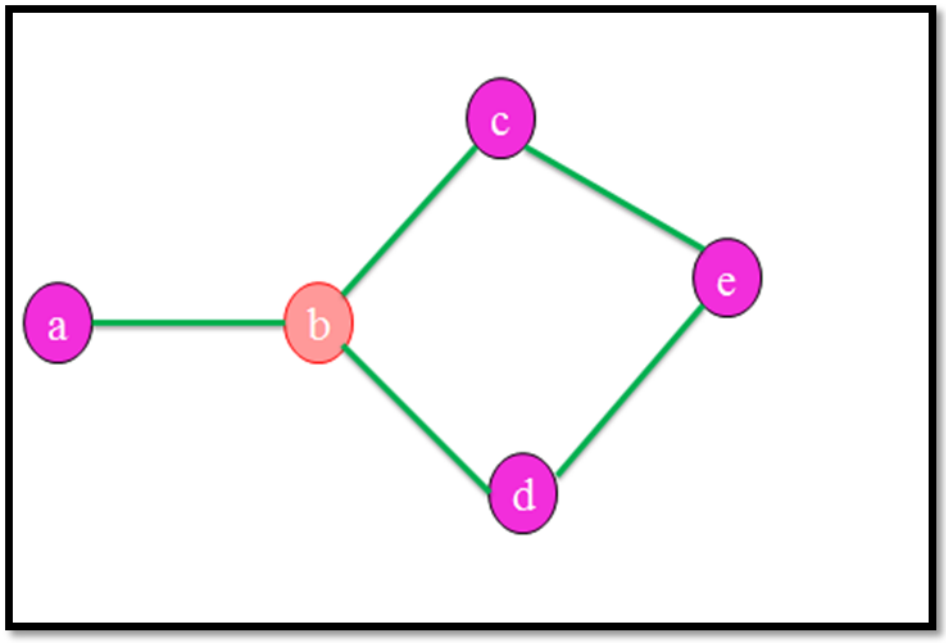

Fig 3.7: Betweenness for b

$$g\,(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

Cb(b) = ((σac(b) / σac) + (σad(b) / σad) + (σae(b) / σae) + (σcd(b) / σcd) + (σce(b) / σce) + (σde(b) / σde)) / 6 = ((1 / 1) + (1 / 1) + (2 / 2) + (1 / 2) + 0 + 0) / 6 = 3.5 / 6 ≈ 0.583

3.6.3 Clustering coefficient

The clustering coefficient is a ratio $\frac{N}{M}$. N is the number of edges between the neighbors of a network and M is the maximum number of edges existing between the neighbors of a network. The clustering coefficient of a node is always a number flanked by 0 and 1. The network clustering coefficient was inferred as the average of the clustering coefficients for all nodes in the network. If the clustering coefficient is zero, then the nodes are supposed to have less than two neighbors.

Cn = 2en/(kn(kn-1)), where kn is the number of neighbors of n and en is the number of connected pairs between all neighbors of n.

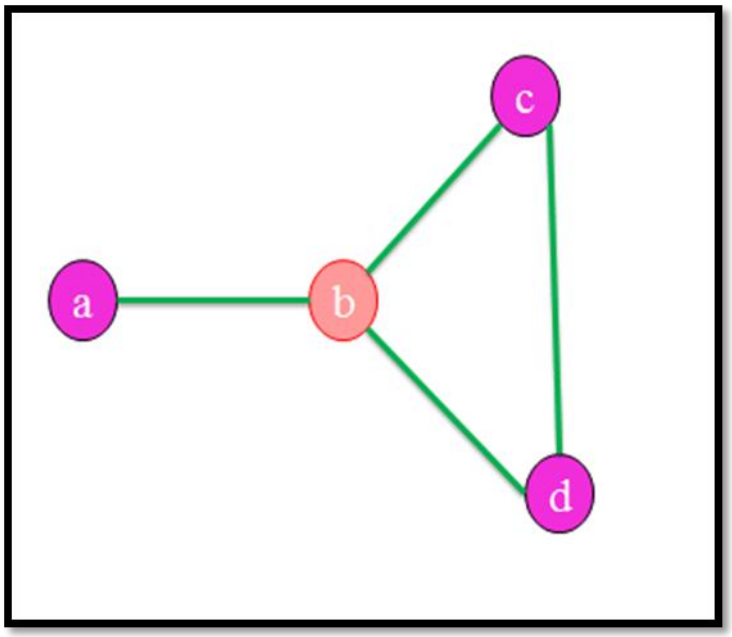

Fig 3.8: Clustering coefficient for b

In fig 3.8, there is one triangle that passes through node b (the triangle bcd). The maximum number of triangles that could pass through b is three (in this case, the pairs (a, c) and (a, d) would be connected additionally). This yields a clustering coefficient of Cb = 1 / 3.

3.6.4 Shortest path length

There are many alternative paths for a pair of nodes in the network. The path with the nominal number of links indicated the shortest path. The number of links transitory through in the shortest path indicated the shortest-path distance. The shortest length of the path provided the count of node pairs in a network. The fig 3.9 depicts the shortest path distance between nodes F to node H.

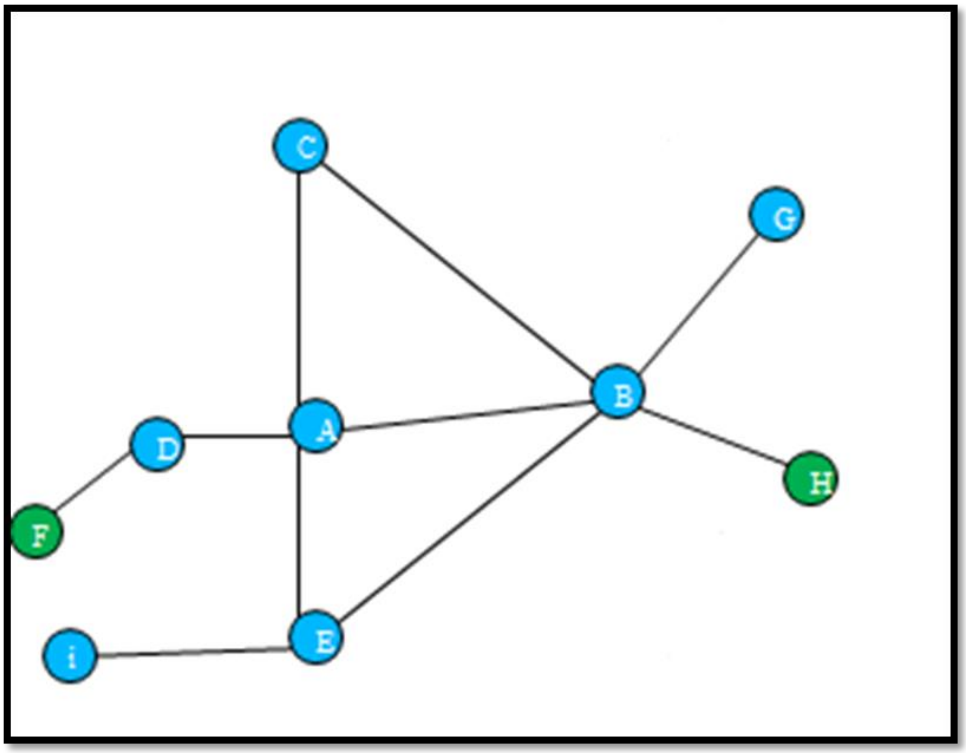

Fig 3.9: Shortest-path distance between nodes F to node H

$$(SP_{FH}) = (F, D, A, B, H)_{=5}$$

3.6.5 Pathway analysis

The pathway analysis contributes to data integration like the integration of diverse biological data. The datasets was integrated to the KEGG database using the self-programmed Perl programs. KEGG is the Kyoto Encyclopedia of Genes and Genomes. KEGG is a database reserve for comprehending high-level rationales and applicability of the biological system like the large-scale molecular datasets produced by genome sequencing and other high-throughput experiments. This integration culminated in deducing the pathway information of the PPIs and the

ranking of the pathway using HGM. Pathway analysis led to the identification of the linking cancer pathway in predicted complexes. The phenotype classification divulges the protein participants in major cancer disease. Fig 3.10 depicts the work flow of pathway analysis.

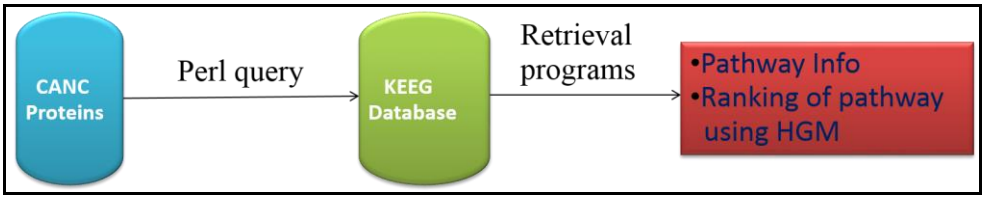

Fig 3.10: Work flow of pathway analysis

3.7 CancerNet

CancerNet tool was developed to retrieve fundamental information about proteins such as interaction, chemical component, etc. This tool works in coordination with other databases such as Uniprot as well as MCODE dataset. CancerNet is a web-based tool. The front end uses Perl and PHP, the backend uses mysql and is developed in the background of Ubuntu operating system. The data captured from the tightly packed human protein complex is integrated to the tool. A user just has to enter the name of the protein, Uniprot id, and gene number. CancerNet retrieves all relevant information from the protein symbol to interacting partner. The query passes through the perl-cgi. The query is processed using the SQL language and the required information is fetched from the database. No data is retrieved if the required information is not stored in the database. Fig 3.11 represents the CancerNet tool.

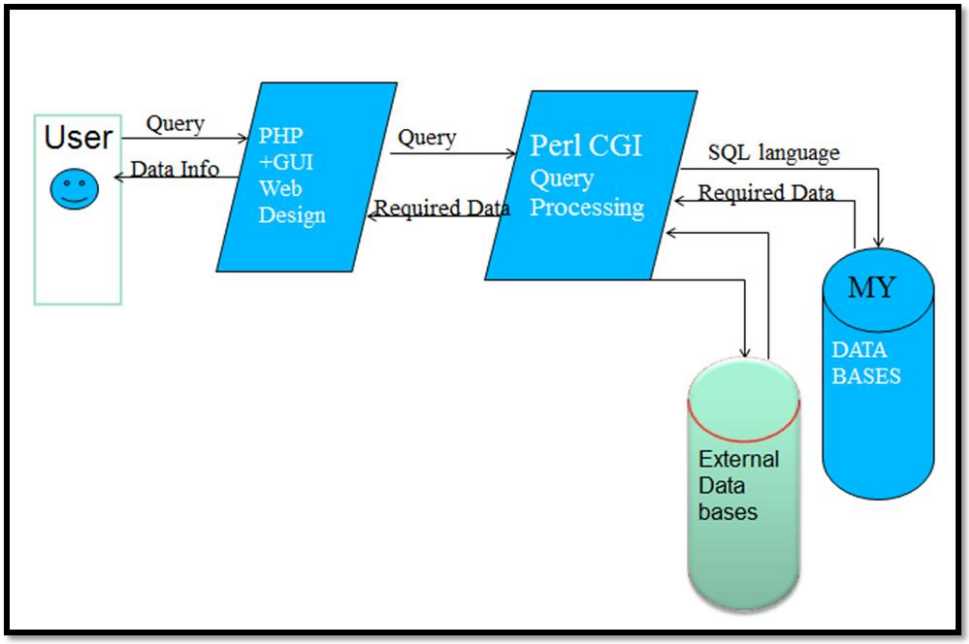

Figure 3.11: CancerNet execution with external datasets

Chapter 4

RESULTS

4.1 Development and mapping of human protein-protein interaction network (HPPIN)

4.1.1 Assortment of human proteins

The very first step to detect the participation of cancer proteins in major complexes was to identify all the available human proteins from legitimate databases. 20199, 10579, 10546, and 3695 proteins were integrated from Uniprot, NCBI Reference sequence, Plasma proteome and human protein atlas respectively.

A total of 45019 proteins were identified. It was inferred from the comparative study that only 3422 protein were common in the four database used for this investigation and 32119 protein were recurrent. The Uniprot database predominantly subsidized 9432 unique proteins.

Table 4.1 portrays the assortment of proteins from Uniprot, NCBI Reference sequence, Plasma proteome and Human protein atlas.

Table 4.1: Assortment of human protein from various databases

Data base	Proteins
Uniprot	20199
NCBI Refseq	10579
Plasma proteome	10546
Human protein atlas	3695

Total	Common	Redundant	Unique
45019	3422	32119	10078

In the initial stage, all the proteins possessed the database format. A manually coded Perl program was employed for altering the formats, integration, and identification of common proteins as well as unique protein. The extracted datasets was labelled as Human Protein (HP). Fig 4.1 portrays the identification of human proteins from various databases.

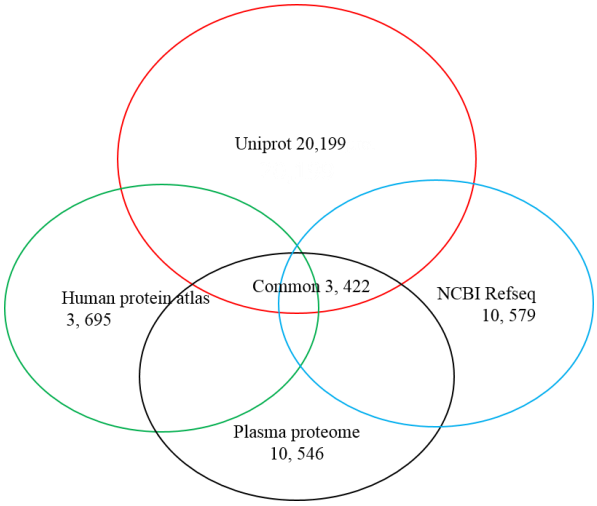

Fig 4.1: Identification of human proteins

Red circle indicates Uniprot database with 20199 proteins, green indicates human protein atlas with 3695 proteins, blue circle indicates NCBI refseq database with 10579 proteins and black circle indicates plasma proteome database with 10546 proteins. 3422 are the common proteins among the four databases.

4.1.2 Assortment of PPI maps

To develop the human interactome 41327, 19404, 8412, and 10807 binary protein interaction data proteins were integrated from the HPRD, IntAct, MINT, and DIP databases respectively. The databases or the datasets used in the investigation followed different standards and formats for the naming of PPIs owned. So, in order to maintain consistency in naming convention, the resultant PPIs were assigned a unique ID using personalized Perl programs.

A total of 422 self-interaction and 3456 duplicate interactions were removed from the final 76072 data set and was labelled as Human Protein-Protein Interactions (HPPIs). Fig 4.2 portrays the mapping of HP and PPI respectively.

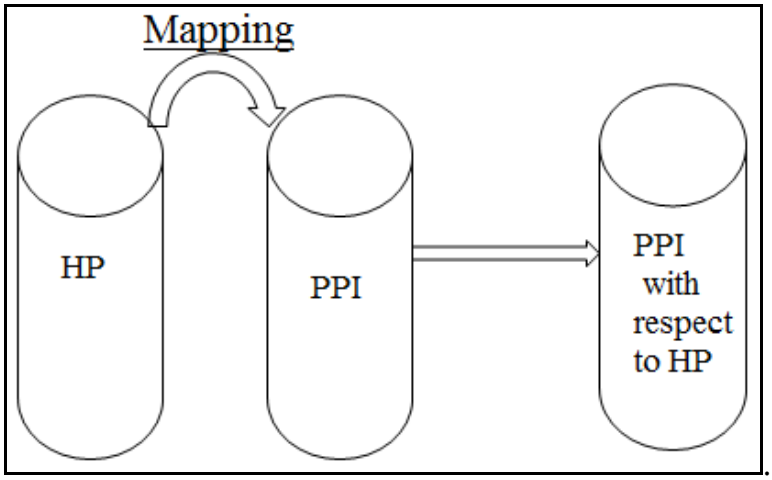

Fig 4.2: Identification of PPI maps

This image indicates the mapping of HP proteins in PPI using Perl programs.

Table 4.2a: Assortment of PPIs

Data base	Human Protein-Protein Interactions
HPRD	41327
IntAct	19404
DIP	8412
MINT	10807
Total	79950

Table 4.2b: Percentage of PPIs

	HPRD	**IntAct**	**DIP**	**MINT**
HPRD		21%	2%	21%
IntAct	42%		1%	37%
DIP	73%	34%		33%
MINT	66%	59%	2%	

Table 4.2c: Total interactions

Total	**Self-interaction**	**Redundant**	**Unique**
79,950	422	3456	76072

4.1.3 Mapping HP to PPI

A HP dataset and HPPI dataset was developed from the investigation. These datasets were mapped with each other. For mapping, each protein from the HP datasets was mapped to the HPPI dataset based on the analogous interaction. From a total of 10078 proteins, only 9951 proteins were inferred for interactions in the HPPI data set. 127 proteins interactions were missing. The residual 127 protein interactions were detected by employing the orthology-based approach. At the end of mapping, a total of 10078 proteins and its 58674 interactions were recognized. The dataset developed from this investigation was labelled as

Human Interactome (HIN). Fig 4.3 portrays the graphical view and analysis of HIN.

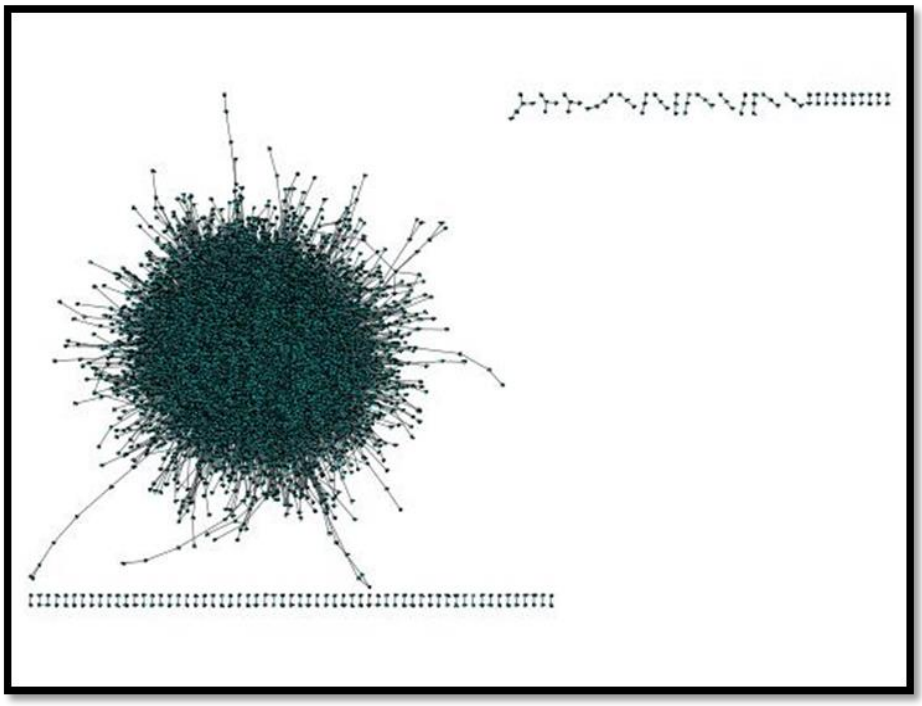

Fig 4.3: Graphical representation of human interactome

The circle indicates proteins and the links indicates the interaction. The proteins at the top corner and bottom of the image are the detached proteins and its interactions in the human interactome.

The entire extracted human binary protein interactions from this investigation was loaded into a highly authenticated tool Cytoscape 3.0.0 to visualize biological pathways and molecular interaction networks. These networks were integrated with gene expression profiles, other state data and annotations by means of Cytoscape. The exceptional tool Network analyzer was employed for the analysis of HIN. Fig 4.4 portrays the representation of an analysis of clustering coefficient using network analyzer.

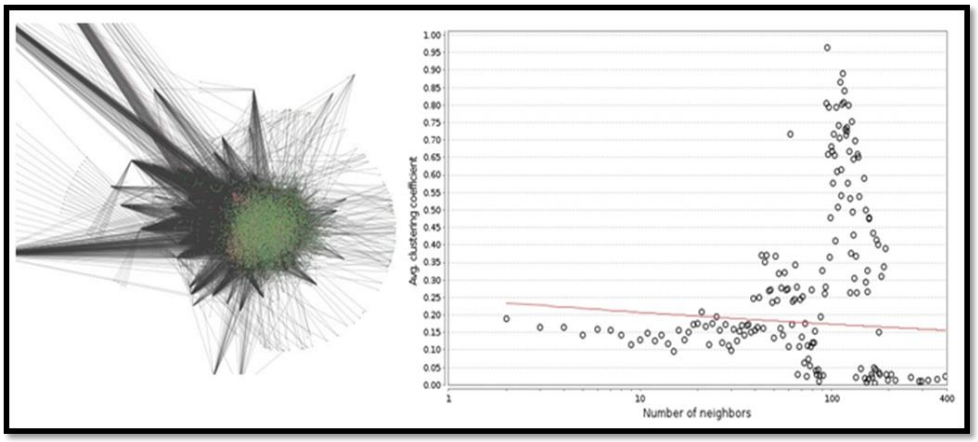

Fig 4.4: Analysis of clustering coefficient using network analyser

The left side of the image indicates the human interactome and on the right side is its statistical analysis of clustering coefficient, which is low.

Table 4.3 illustrates the statistics derived using a network analyzer.

Table 4.3: Analysis derived from Network Analyzer

Global characteristics	**Values**
Clustering Coefficient	0.135
Isolated network	87
Network diameter	15
Network radius	1
Network centralization	0.038
Shortest path	977233374(96%)
Path length	4.054
Average number of neighbors	11.644
Number of nodes	10078
Network density	0.001
Network heterogeneity	2.052
Isolated node	0
Number of self-loops	0
Run Time	6167.672(sec)

It was inferred that the isolated nodes are zero in the extracted HIN, which signifies the fact that all the nodes are linked to binary interaction. Also, all the nodes are linked pairwise. The number of connected network represents the connectivity of human interactome network. Only 87 interactions are isolated from the identified 58674 interaction, which indicates stronger connectivity. In the HIN, 96% of shortest path was also identified. The average shortest path between two nodes in the interactome is 4.054. The largest distance between two nodes in the interactome is 15 which is termed as the network diameter. The derived HIN is disconnected. Hence, the diameter can also be described as the maximum node eccentricity which is 15. The average connectivity of a node in the network is 11.644 nodes. The network density value between 0 and 1 implies how densely the network is populated within the edge since self-loop and duplicated edges are ignored in the derived HIN and demonstrates the value as 0.001.

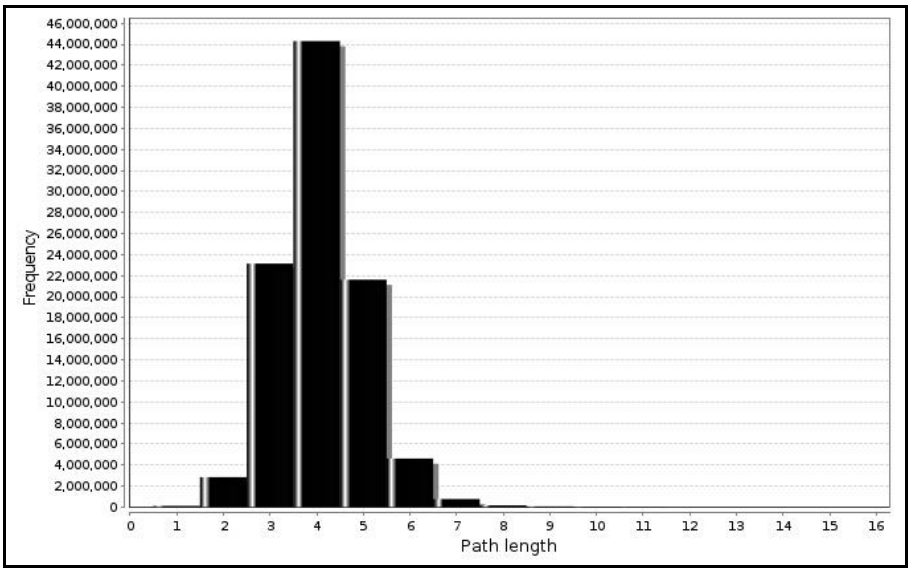

Fig 4.5: Path length analysis

The path length of human interactome is less.

The clustering coefficient of the vertex indicates the intensity of the neighborhood of a vertex. The clustering coefficient of the entire network is the average of the clustering coefficients of the vertices. The clustering coefficient of node is always between 0 and 1. The clustering coefficient obtained was 0.135 as the result of this investigation.

The developed HIN demonstrates an acceptable correlation with the node's degree distribution. Above 55% nodes indicated more than 11 interactions. The nodes degree distribution correlates well with the low power. Fig 4.6 portrays the degree distribution.

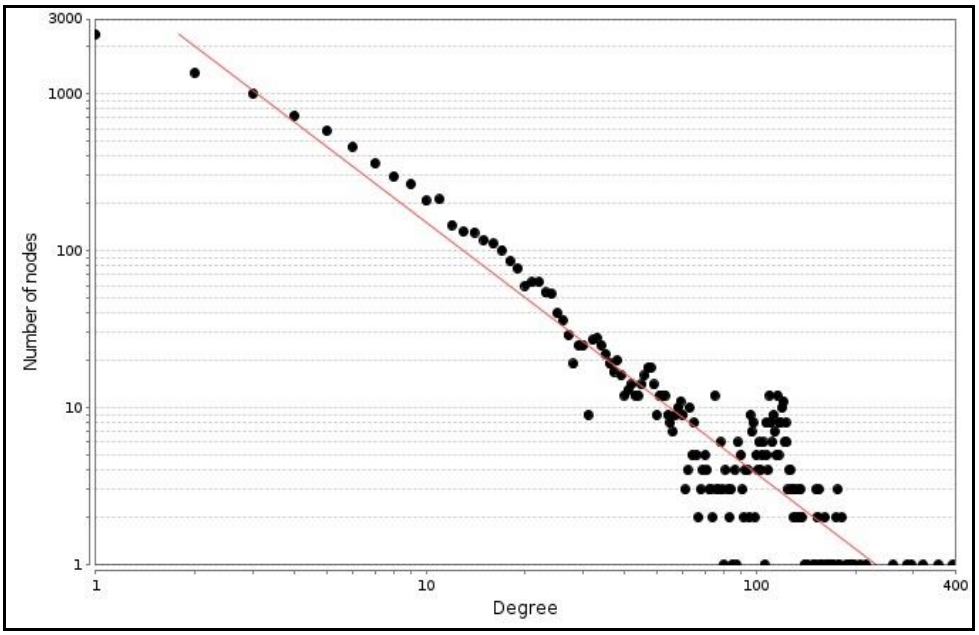

Fig 4.6: Degree Distribution

The nodes degree distribution correlates well with the low power.

4.2 Evaluation of Cancer and Non-Cancer Complexes in HIN

Cancer proteins were collected from valid and curated database specifically CBio and Sanger. 13944 and 3165 cancer proteins were

mapped in HIN. The different dataset possessed different format for the cancer proteins. So, in order to preserve the uniformity in naming convention, the resultant cancer proteins were assigned a unique ID by means of personalized Perl programs. The cancer dataset obtained was labelled as Cancer Proteins (CP).

Table 4.4 demonstrates the analysis inferred from the integration of cancer proteins from CBio and Sanger databases.

Table 4.4: Analysis by integration of cancer proteins

Database	**No of proteins**	**Duplicated**	**Unique**
Sanger	13944	10798	3146
CBio	3165	19	

The CP dataset was mapped against the HIN dataset to reconstruct cancer interactions in human interactome. Fig 4.7 portrays the MCODE clusters.

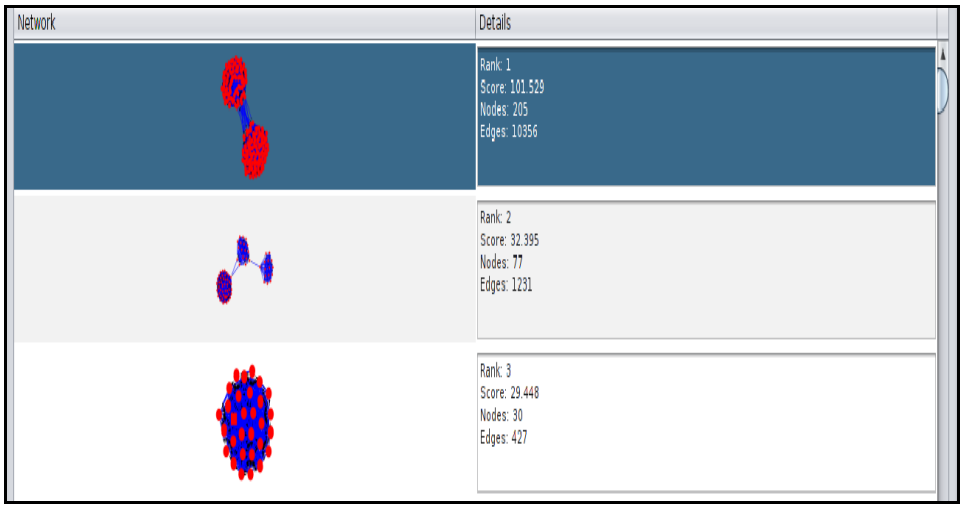

Fig 4.7: MCODE clusters

Each circle indicates proteins clusters. Blue color indicates interaction and red color indicates proteins. The right side of the image displays the number of edges and nodes.

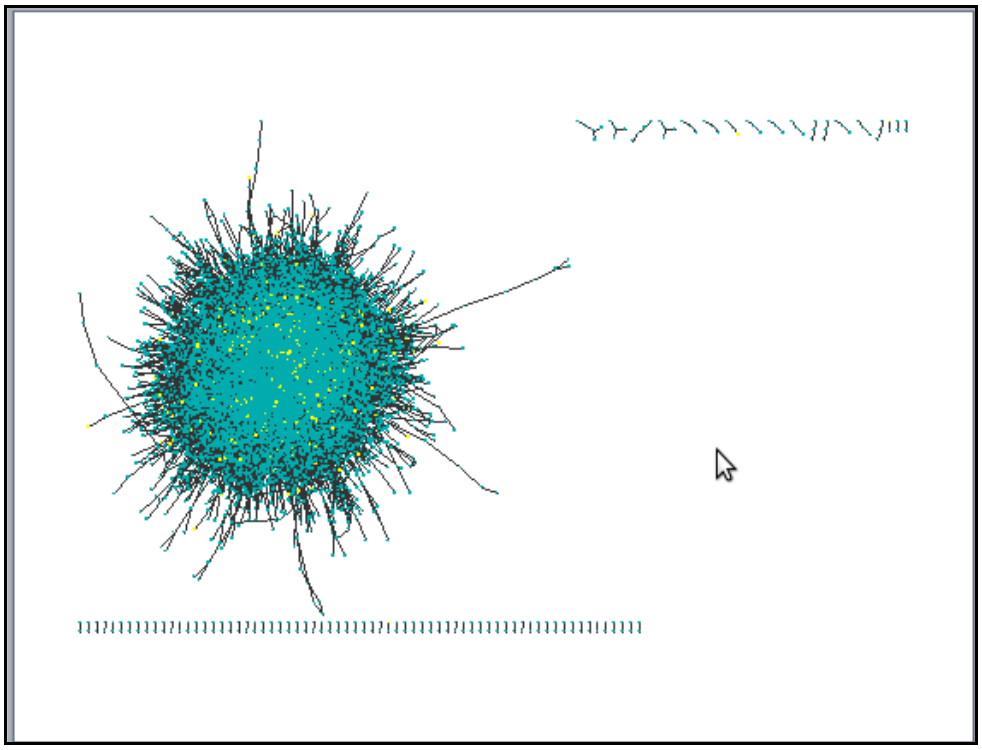

Fig 4.8: Representation of cancer interaction

All the cancer protein covered belong to the largest network in the interactome. The proteins highlighted in yellow color represents the cancer proteins.

4.2.1 Connected Component Algorithm

The largest component from HIN was extracted by employing the Connected Component Algorithm (CCA). This algorithm mainly plays a pivotal role in detaching all the unconnected network from the largest network.

The reconstructed CCA network possess 9886 proteins and its 58568 interactions. Table 4.5 illustrates the implications from a CCA network.

Table 4.5: CCA network

Clustering Coefficient	0.138
Isolated network	1
Network diameter	15
Network radius	8
Network centralisation	0.039
Shortest path	977231104(100%)
Path length	4.054
Average number of neighbours	11.849
Number of nodes	9886
Network density	0.001
Network heterogeneity	2.032
Isolated node	0
Number of self-loops	0
Run Time	6287(sec)

From the implications of CCA network, it was understood that, if the isolated node are zero then all the nodes are linked to binary interaction. All the nodes are linked pairwise. The number of connected network represents the connectivity of the CCA network along with 58568 interaction. The only one network isolated is the CCA network. This result indicates that the stronger connectivity. 100% shortest path is included in the largest network. The average shortest path between two nodes in the CCA network is 4.054 which is same as the HIN. The network diameter is 15 and is same as HIN. The derived HIN is disconnected. So, the diameter can also be described as the maximum node eccentricity which is 15. The average connectivity of a node in the network is 11.849 nodes which is also equivalent to the HIN. The network density value lies in between 0 and 1, which in turn implies the density of a populated network with edge as the self-loop and duplicated edges are ignored in the derived CCA network and demonstrates the value 0.001. Fig 4.9 portrays a CCA network. More

information on result can be retrieved from http://www.bioteccancernet.cusat.ac.in/cca.

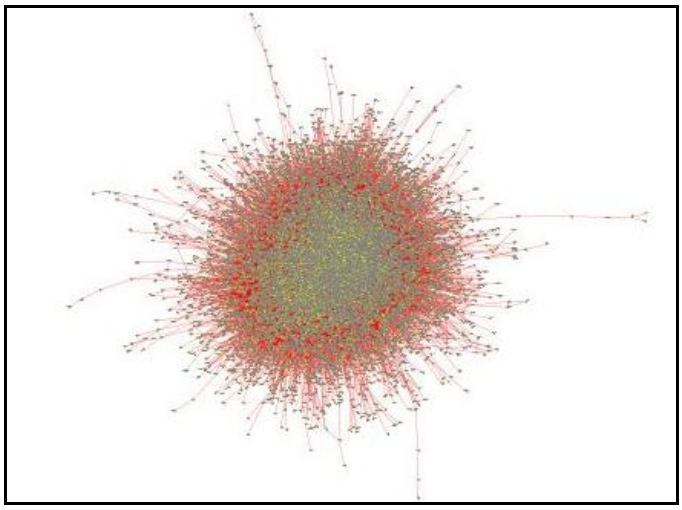

Fig 4.9: CCA network

This the cytoscape view of highest network in a human interactome.

Fig 4.10 portrays the betweenness centrality. The betweenness centrality of a node reflects the amount of control that this node exerts over the interactions in the network. The measure of betweenness centrality favors the node with dense sub network. Fig 4.10 illustrates that the value for all the nodes, betweenness centrality lies between 0 and 1.

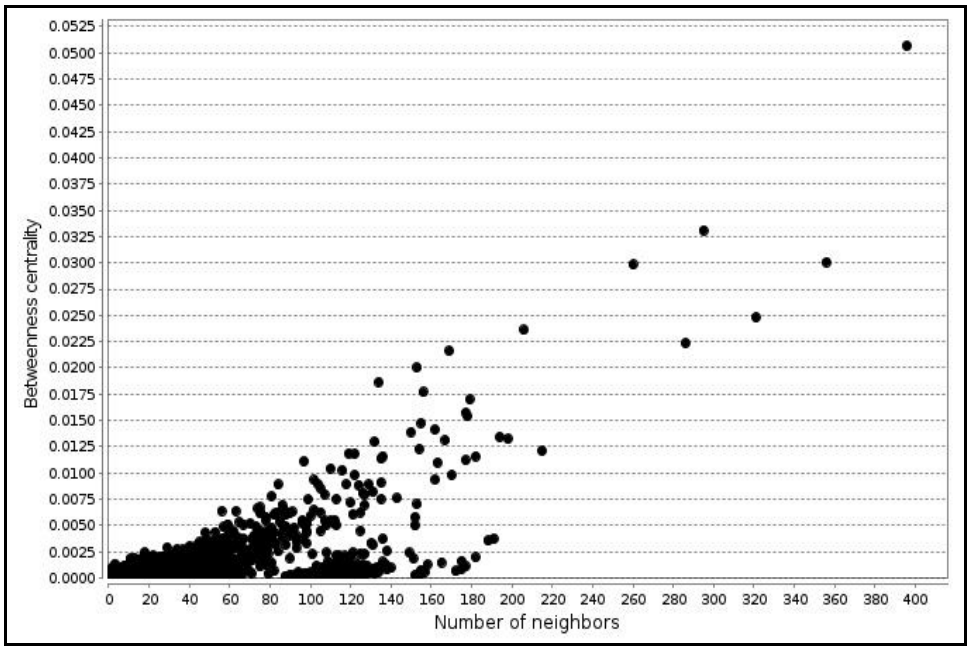

Fig 4.10: Betweenness centrality

In betweenness centrality, the value for all the nodes, betweenness centrality lies between 0 and 1.

Clustering coefficient of vertex indicates the intensity of the neighborhood of a vertex. The clustering coefficient of the entire network is the average of the clustering coefficients of the vertices. The clustering coefficient of node always lies in between 0 and 1. The value obtained is 0.138 for clustering coefficient which is very close to human interactome.

Clustering coefficient, betweenness centrality and node distribution displays the highest connectivity of CCA network, which includes most of the cancer proteins. The study of major domain from the CCA network is imperative since the proteins are tightly connected in CCA. Fig 4.11 portrays the clustering coefficient.

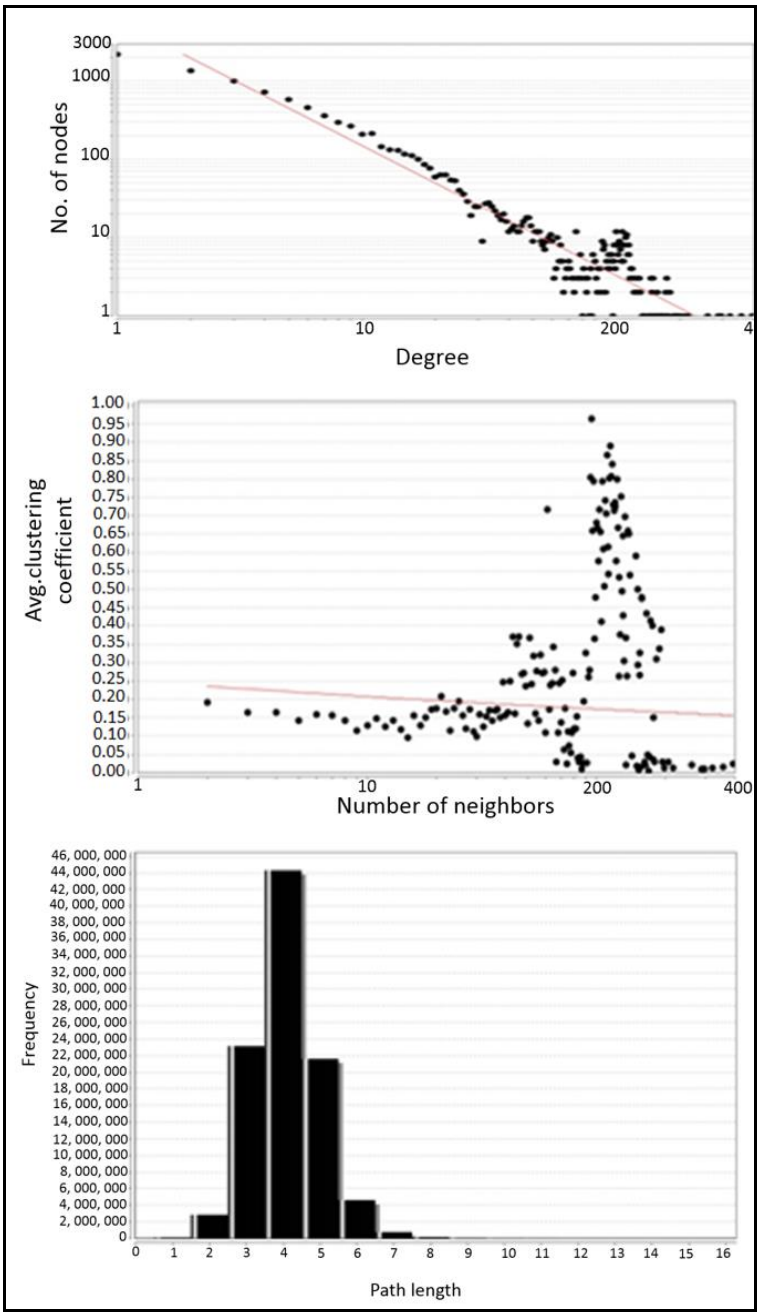

Fig 4.11: Clustering coefficient

Clustering coefficient, betweenness centrality and node distribution displays the highest connectivity of CCA network, which includes most of the cancer proteins.

4.2.2 Molecular complex detection method

Molecular complex detection method (MCODE) identified clusters in the CCA network. MCODE algorithm discovered highly connected biological modules from huge protein network. High-scoring clusters have a high density value. MCODE derived totally 121 highly connected modules from the CCA network. The further analysis considered only 99 MCODE clusters as the density score value is >2. More information on result can be retrieved from http://www.bioteccancernet.cusat.ac.in/mcode.

In the derived dataset, the rank 1 modules own the highest number of protein as well as interactions such as 205 and 10356 respectively. This data is present in only one module. In contrast, the least number of protein 3 and interaction 3 contains 38 number of modules. 121 protein domains were unique. 99 complexes having total 1141 unique proteins were selected and the total interactions was measured to be 14340.

Table 4.6: Cluster and density score

Cluster	Density Score	No: proteins	No: Interactions
1	101.529	205	10356
2	32.395	77	1231
3	29.448	30	427
4	22	22	231
5	19	19	171
6	19	19	171
7	14.1	21	141
8	13	13	78
9	12.923	14	84
10	9.846	14	64
11	8.5	9	34
12	8.25	25	99
13	8.1	41	162
14	8	8	28
15	7	7	21
16	6.933	16	52
17	6	6	15
18	5.6	6	14
19	5	5	10
20	5	5	10
21	5	5	10
22	5	5	10
23	5	5	10
24	4.809	48	113
25	4.5	5	9
26	4.5	5	9
27	4.296	28	58
28	4.25	9	17
29	4	4	6
30	4	5	8
31	4	4	6
32	4	4	6
33	4	4	6
34	4	4	6
35	4	4	6
36	4	4	6

37	4	4	6
38	4	4	6
39	4	4	6
40	3.933	91	177
41	3.917	25	47
42	3.714	8	13
43	3.58	82	145
44	3.375	17	27
45	3.333	4	5
46	3.333	4	5
47	3.333	4	5
48	3.333	4	5
49	3.333	4	5
50	3.333	4	5
51	3.333	4	5
52	3.333	4	5
53	3.333	4	5
54	3.333	4	5
55	3.333	4	5
56	3.333	4	5
57	3.333	4	5
58	3.333	7	10
59	3.176	18	27
60	3	5	6
61	3	3	3
62	3	3	3
63	3	3	3
64	3	3	3
65	3	3	3
66	3	3	3
67	3	3	3
68	3	3	3
69	3	3	3
70	3	3	3
71	3	3	3
72	3	3	3
73	3	3	3
74	3	3	3
75	3	3	3

76	3	3	3
77	3	3	3
78	3	3	3
79	3	3	3
80	3	3	3
81	3	3	3
82	3	3	3
83	3	3	3
84	3	3	3
85	3	3	3
86	3	3	3
87	3	3	3
88	3	3	3
89	3	3	3
90	3	3	3
91	3	3	3
92	3	3	3
93	3	3	3
94	3	3	3
95	3	3	3
96	3	3	3
97	3	3	3
98	3	3	3
99	3	5	6
100	2.958	72	105
101	2.952	43	62
102	2.857	8	10
103	2.8	6	7
104	2.762	22	29
105	2.75	9	11
106	2.739	93	126
107	2.733	31	41
108	2.692	27	35
109	2.667	7	8
110	2.667	4	4
111	2.667	4	4
112	2.667	4	4
113	2.667	4	4
114	2.667	4	4

115	2.667	4	4
116	2.667	4	4
117	2.667	4	4
118	2.667	4	4
119	2.5	5	5
120	2.424	34	40
121	2.4	16	18

Table 4.7 illustrates the total number of modules, the number of protein and the number of interactions present in that module.

Table 4.7: Proteins and interactions

	No. of modules	**No. of proteins**	**No. of interactions**
Modules	121	1550	14873
Selected modules	99	1141	14340

4.3 Validation of complexes

In order to validate the complexes, the preliminary gene ontology analysis was executed and the resultant data was further assessed using the statistical analysis. 1141 proteins was used for GO analysis and also used for the classification of proteins. For the validation, 99 clusters were divided into three categories such as clusters having only normal proteins and its normal interaction, clusters having only cancer proteins and its cancer interaction, and finally clusters having both cancer and non-cancer proteins and its cancer and non-cancer interaction.

Table 4.8 and Fig 4.12 portrays the protein distribution. The data from the table confirms that the exceptionally tightly connected protein modules displays the over expression of cancer proteins.

Table 4.8: Proteins distributions

	No. of modules	No. of proteins	No. of cancer interactions	No. of normal proteins
Modules	121	1550	873	677
Selected modules	99	1141	785	356

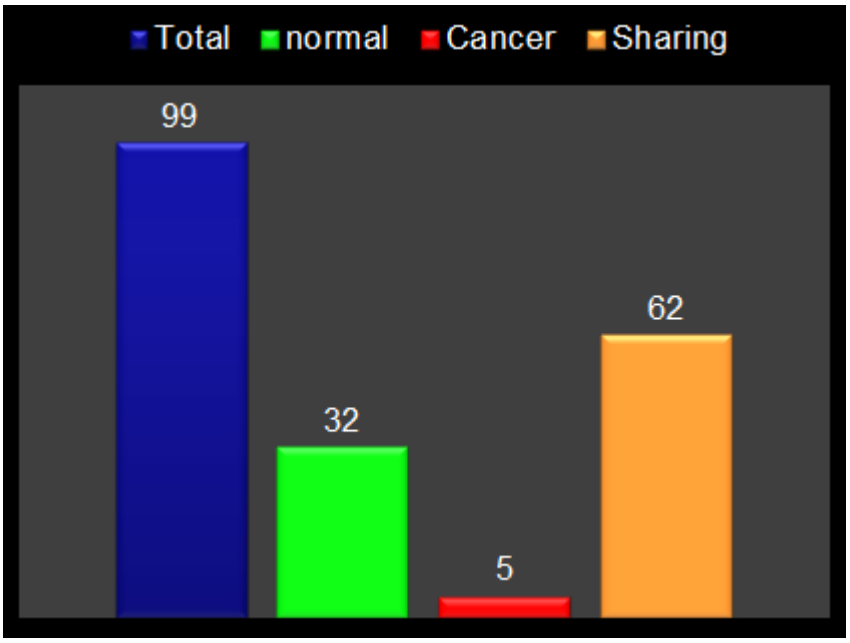

Fig 4.12: Distribution chart for MCODE complexes

(62 represents the cancer and normal protein clusters. 32 represents the normal protein clusters and 5 represents the cancerous protein clusters.)

The GO analysis of the protein networks were executed by utilizing the GO database. Hyper geometric exact test was employed to extract the information pertaining to the biological function enriched genes. The entire set of genes were subjected to the simultaneous test for multiple GO categories such as biological process (BP), molecular function (MF),

cellular component (CC), and p-values were calculated for each proteins present in the cluster. The outcome was GO annotations of 1550 proteins including the MCODE proteins and the corresponding p-values. A total of 121 modules were subjected to GO analysis and one of the module is explained in table 4.9a and b. Fig 4.13 portrays the validation of modules. More information on result can be retrieved from http://www.bioteccancernet.cusat.ac.in/go.

These results indicate that the strategy of permitting proteins to fit into different clusters appears to be effective for grouping multi-functional proteins into manifold functional groups. This also helps to align the biologically relevant modules with the corresponding different protein functions. In the cluster, p-value of all the proteins are further exploited to recognize the p-value of the cluster.

Table 4.9a: GO interpretation of one module

List of proteins from one module
RPL18 HNRPK RPL17 RPL19 RPL14 HNRPF RPL13 U2AF2
RPL15 HNRPD HNRPR RPLP2 HNRPM HNRPL
SFRS6 SFRS7 SFRS4 SFRS5 DHX38 RBM8A
EIF1AX RPLP0 HNRPU SFRS9 RPLP1 RPL26L1
U2AF1 FAU RPL10 RPL11 LSM2 RPL12 PTBP1 EIF2S3
SFRS1 SFRS3 SFRS2 PCF11 RPS18 RPS19 HEAB RPS16 RPS17
EIF2S1 RPS14 EIF2S2 RPS15 RPS12 RPS13 RPS10
RPS11 FUS NHP2L1 RPS25 RPS26 RPS27 RPS28 RPS29
CDC40 RPS20 RPS21 HNRPH2 HNRPH1 SFRS11
RPS23 RPS24 DHX9 RPSA RPS9 RPS6 RPS5 WBSCR1
RPS8 RPS7 RPL18A EIF3S10 EIF3S12 YBX1
RPL41 EIF4A2 EIF4A1 RPL3L RPS4Y1 EEF1G
THOC4 EEF1D POLR2H POLR2G
POLR2F POLR2E POLR2L POLR2K
POLR2J POLR2I RPL27A RPL35 RPL36 RPL37
RPL38 POLR2D SF3B5 RPL39 SF3B4 POLR2C
POLR2B SF3B3 SF3B2 POLR2A SF3B1 RPL30 RPL32
RPL31 CD2BP2 RPL34 SNRP70 RPL26 RPL27
RPL24 ETF1 SF3A2 SF3A1 RPL28 RPL29 SF3A3 PAPOLA
RPL23 RPL22 RPL21 PHF5A HNRPA2B1 TXNL4A
NCBP2 NCBP1 RPL36A EIF3S1
SNRPD3 SMC1L1 EIF5 ASCC3L1 EIF5B
SNRPD1 RBM5 SNRPD2 DDX23 DNAJC8
EIF3S7 EIF3S6 EIF3S9 EIF3S8 EIF3S3
EIF3S2 RPS27A EIF3S5 EIF3S4
PABPN1 SNRPA1 RPL35A EFTUD2
MAGOH EEF2 HNRPA3 HNRPA1 HNRPA0
SNRPB SNRPA SNRPF SNRPE UBA52
SNRPG EEF1B2 SNRPB2 RPL7 RPL6 RPL9
RPL8 RPL3 RPL5 RPL7A RPL4 RPL10A
EEF1A1 CSTF2 RPL23A RBMX EIF4B EIF4E
GTF2F1 GTF2F2 RPL37A CSTF1 RPS2 RPS3
RPS3A PCBP1 PCBP2 RPS4X PRPF4 PRPF6 EIF4G1
CPSF3 CPSF2 CPSF1 RPS15A HNRPC RPL13A

Table 4.9b: GO interpretation of one module

GO-ID	p-value	Description
6412	1.08E-24	Translation
10467	9.44E-19	gene expression
34645	3.70E-17	cellular macromolecule biosynthetic process
9059	4.34E-17	macromolecule biosynthetic process
44267	5.30E-16	cellular protein metabolic process
19538	3.44E-15	protein metabolic process
22613	4.08E-15	ribonucleoprotein complex biogenesis
42254	5.49E-15	ribosome biogenesis
44249	1.60E-13	cellular biosynthetic process
9058	2.67E-13	biosynthetic process
44260	2.14E-11	cellular macromolecule metabolic process
43170	5.42E-11	macromolecule metabolic process
44085	1.32E-09	cellular component biogenesis
70925	2.79E-09	organelle assembly
44237	5.97E-09	cellular metabolic process
42255	7.18E-09	ribosome assembly
6417	1.46E-08	regulation of translation
10608	3.05E-08	posttranscriptional regulation of gene expression
8152	3.68E-08	metabolic process
44238	4.32E-08	primary metabolic process
22618	4.50E-08	ribonucleoprotein complex assembly
32268	4.75E-08	regulation of cellular protein metabolic process
51029	2.44E-07	rRNA transport
6407	2.44E-07	rRNA export from

		nucleus
51246	6.75E-07	regulation of protein metabolic process
6364	1.88E-05	rRNA processing
6405	2.52E-05	RNA export from nucleus
16072	2.57E-05	rRNA metabolic process
50658	7.02E-05	RNA transport
51236	7.02E-05	establishment of RNA localization
50657	7.73E-05	nucleic acid transport
51168	1.22E-04	nuclear export
34622	1.35E-04	cellular macromolecular complex assembly
6450	1.47E-04	regulation of translational fidelity
6403	1.57E-04	RNA localization
9987	1.83E-04	cellular process
15931	2.09E-04	nucleobase, nucleoside, nucleotide and nucleic acid transport
34470	2.43E-04	ncRNA processing
51169	2.93E-04	nuclear transport
6913	2.93E-04	nucleocytoplasmic transport
6396	3.00E-04	RNA processing
42257	3.60E-04	ribosomal subunit assembly
34621	3.98E-04	cellular macromolecular complex subunit organization
6448	4.23E-04	regulation of translational elongation
65003	7.55E-04	macromolecular complex assembly
34660	7.68E-04	ncRNA metabolic process
42273	1.29E-03	ribosomal large subunit biogenesis
43933	1.39E-03	macromolecular complex subunit organization

27	1.79E-03	ribosomal large subunit assembly
22607	2.39E-03	cellular component assembly
462	2.60E-03	maturation of SSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA)
30490	3.26E-03	maturation of SSU-rRNA
10468	5.68E-03	regulation of gene expression
50686	6.28E-03	negative regulation of mRNA processing
33119	6.28E-03	negative regulation of RNA splicing
48025	6.28E-03	negative regulation of nuclear mRNA splicing, via spliceosome
465	6.28E-03	exonucleolytic trimming to generate mature 5'-end of 5.8S rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA)
10556	6.46E-03	regulation of macromolecule biosynthetic process
470	6.80E-03	maturation of LSU-rRNA
463	6.80E-03	maturation of LSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA)
31326	7.53E-03	regulation of cellular biosynthetic process
9889	7.87E-03	regulation of biosynthetic process
16070	1.16E-02	RNA metabolic process

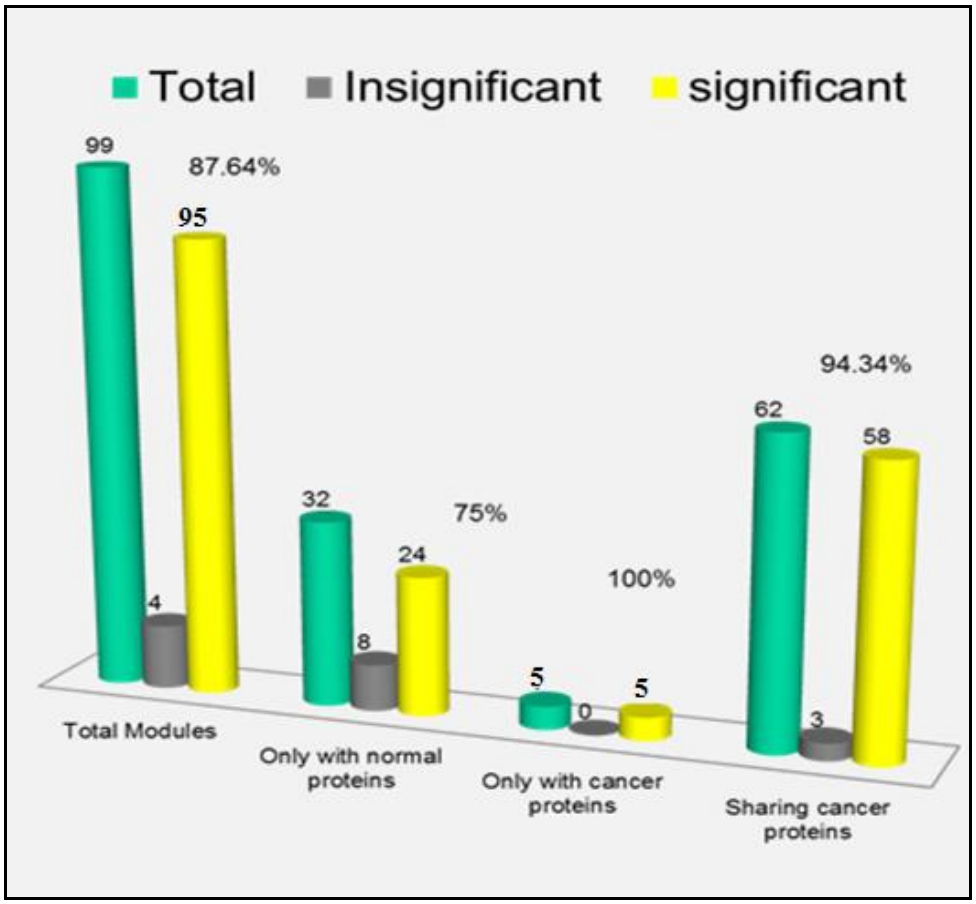

Fig 4.13: Validation of modules

In 99 clusters, 95 clusters are valid. In 95 clusters, 24 normal protein clusters are valid, 5 are with cancer cluster and 58 has both cancer and normal clusters.

Table 4.10 displays p-value of 99 clusters. A cut-off of 0.08 was used. If the p-value of a cluster is below the cut-off, it is considered to be insignificant. 94.34% accuracy was attained in this validation. All these four clusters share cancerous and non-cancerous interaction groups. The statistical analysis also estimated 95 modules to be biologically relevant.

Table 4.10: Cluster p-value

Cluster	p-value
1	0.005
2	0.046
3	0.072
4	0.012
5	0.028
6	0.059
7	0.002
8	0.0025
9	0.0095
10	0.0071
11	0.053
12	0.024
13	0.068
14	0.073
15	0.005
16	0.0049
17	0.0082
18	0.0083
19	0.0094
20	0.0045
21	0.049
22	0.031
23	0.072
24	0.019

25	0.031
26	0.047
27	0.033
28	0.067
29	0.079
30	0.0084
31	0.0049
32	0.0073
33	0.0016
34	0.0067
35	0.0078
36	0.002
37	0.0085
38	0.077
39	0.019
40	0.036
41	0.041
42	0.053
43	0.044
44	0.0085
45	0.0069
46	0.0018
47	0.025
48	0.029
49	0.0083
50	0.0094

51	0.0083
52	0.0018
53	0.0073
54	0.0093
55	0.038
56	0.043
57	0.038
58	0.0019
59	0.0064
60	0.0078
61	0.053
62	0.075
63	0.063
64	0.0016
65	0.086
66	0.940
67	0.0081
68	0.051
69	0.0021
70	0.075
71	0.0033
72	0.032
73	0.0391
74	0.0037
75	0.00931
76	0.913

77	0.0183
78	0.0037
79	0.0910
80	0.0291
81	0.0837
82	0.8910
83	0.0061
84	0.0031
85	0.0090
86	0.0029
87	0.0083
88	0.0071
89	0.061
90	0.039
91	0.0097
92	0.039
93	0.870
94	0.0086
95	0.0019
96	0.0042
97	0.016
98	0.0014
99	0.0082

4.3.1 Grouping of proteins based on the GO annotations

The complex proteins are clustered based on GO annotations like molecular function, cellular component, and biological process. According to the inference from cellular component, 686 out of 1141 proteins were nuclear proteins. Many of the proteins were part of the chromatin and located at the centromere region of a chromosome. Others were part of the histone acetyltransferase or deacetylase complexes or replication fork, cytosol, etc.

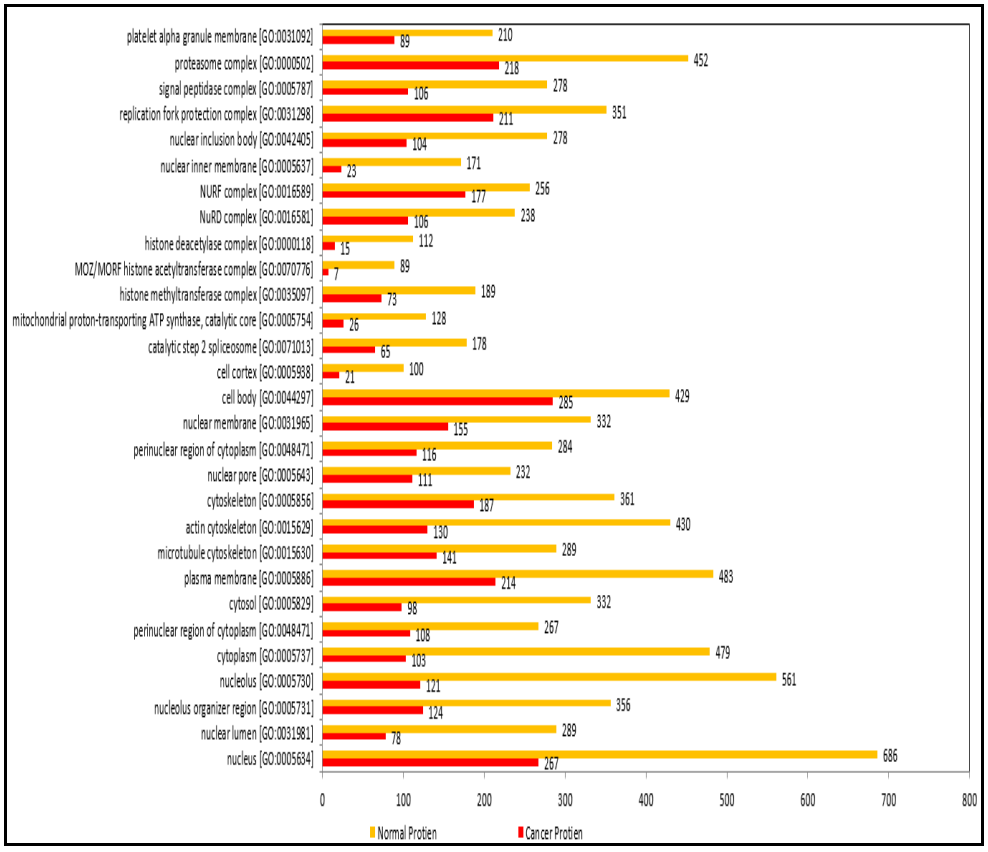

Fig 4.14 Cellular component

In the graph, orange color indicates the normal proteins and red color indicates the cancer proteins. The left hand side depicts the cellular component and right hand side indicates the number of proteins.

In a total of 686 proteins, 234 proteins represent proteins involved in cancer. In cellular component, nuclear lamina contains 8 proteins, which is the lowest. Fig 4.14 portrays the distribution of cellular component.

In the biological process, 324 proteins from 1141 proteins belong to nucleic acid metabolic processes and 143 proteins represented proteins involved in cancer. Many other proteins were related to mitosis (nuclear envelope disassembly, sister chromatid cohesion, spindle localization or cytokinesis), cell cycle regulation (cell cycle progression, cell cycle phase transition, cell cycle checkpoint, regulation of mitotic cell cycle, cell cycle DNA replication) and mRNA transport, histone acetylation, etc. Fig 4.15 portrays the distribution of BP.

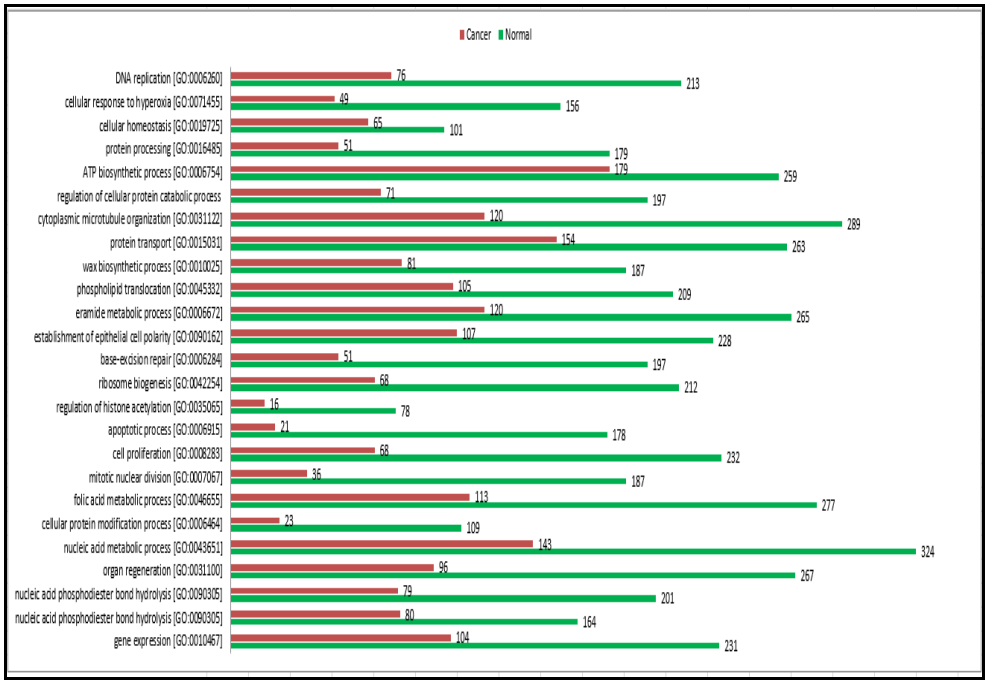

Fig 4.15 Biological processes

In the graph, red color indicates the cancer proteins and green color indicates the normal proteins. The left hand side depicts the biological processes and right hand side indicates the number of proteins.

Finally, the molecular functions of majority complex protein were investigated and concluded that 101 proteins out of the 267 DNA binding proteins were cancer proteins. 89 proteins out of the 175 RNA binding proteins were cancer proteins. Fig 4.16 portrays the molecular function.

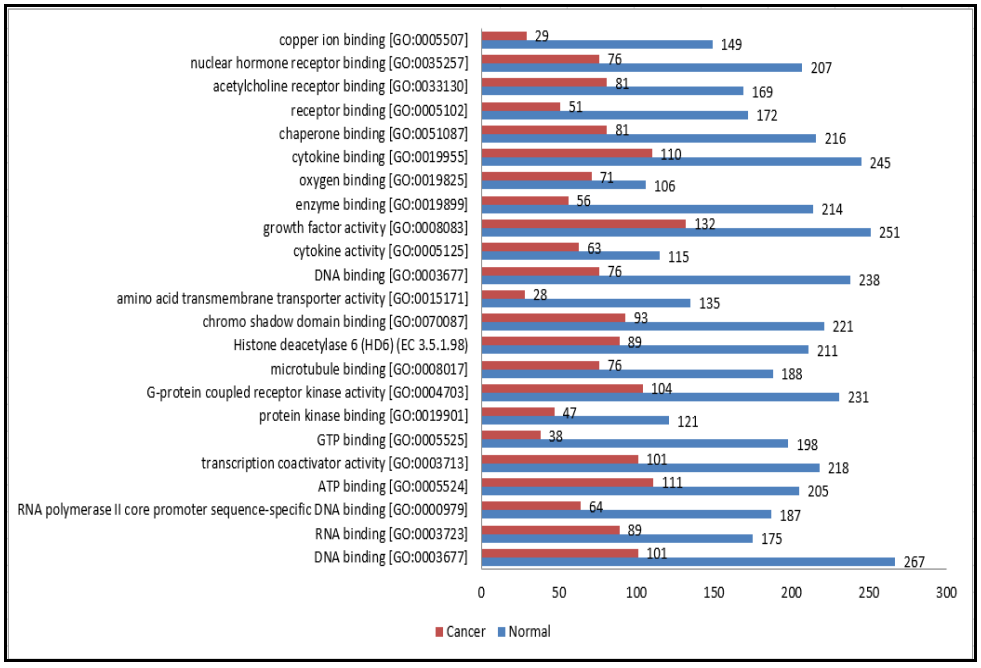

Fig 4.16 Molecular function

In the graph, blue color indicates the normal proteins and red color indicates the cancer proteins. The left hand side depicts the molecular function and right hand side indicates the number of proteins.

4.4 Characterisation

4.4.1 Degree distribution

Degree distribution of the networks has the prospect to explain the counter behavior like overlapping and clustering. All the cancerous Protein interaction networks (PINs) possess some of the higher degree selective nodes conflicting to the non-cancerous PINs. Surveillance of this observation leads to a logical proposal that few of the giant nodes originate

in cancer PINs possessing a massive degree distribution and results in arbitrarily discrete nodes.

In the cluster, many links are found in majority of the nodes. In them, certain nodes display more number of links. This kind of observations are found in numerous investigations in the field of biological networks. The current investigation demonstrates that an assimilated protein cluster contains assorted links with scale-free network. The average degree of the 785 cancer proteins was 38.72. The average degree of 356 non-cancerous proteins was 23.13. This suggests that the average degree of cancerous proteins are ominously higher compared to the non-cancerous proteins. Therefore, cancer encoded proteins has the tendency to interact intensely with other proteins and have comparatively better connectivity in the entire network. Fig 4.17 portrays the more detailed view of the degree characteristics.

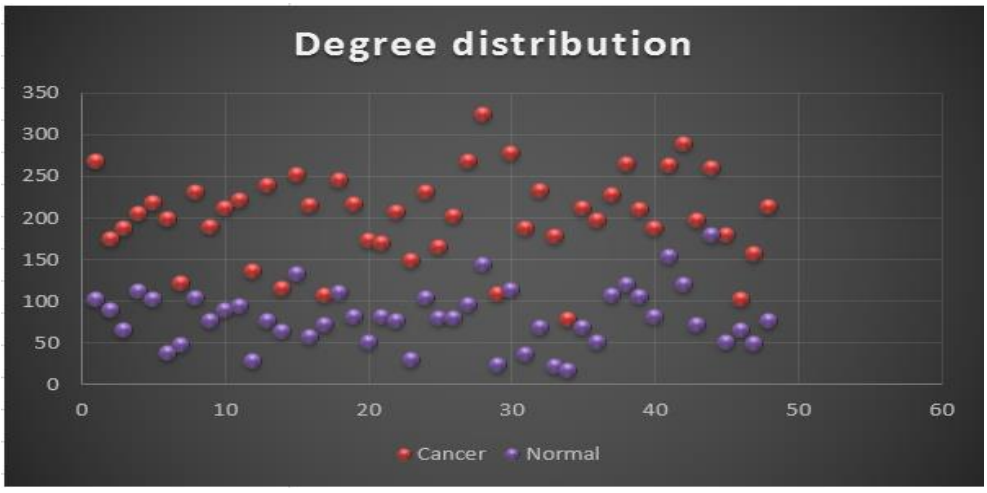

Fig 4.17 Degree distribution

In the graph, purple color indicates the normal proteins and red color indicates the cancer proteins. The left hand side indicates the number of proteins and right hand side indicates the degree.

The cancer proteins inclined to slant towards the higher degree compared to the non-cancerous proteins. The value was calculated as 38.72 for degree centrality of cancer protein by employing test statistics. The value for normal protein was calculated as 23.13. Hence, it can be concluded that the degree of cancerous protein is higher than the non-cancerous protein.

4.4.2 Hub proteins

Remarkably, the cancerous proteins in a human interactome exhibited greater connectivity and were inclined to be hub proteins compared to the normal proteins. This observation reflects that the cancerous genes play imperative role in a human interactome.

The highly connected nodes are generally demarcated as hubs. Presently, description of hubs is an uncertain concern in the study of biological network. A cutoff, that is, degree >10 was employed to delineate hubs in this study. According to this cutoff, 250 (81.1%) of the cancer proteins were categorized as hubs. This number is considerably greater than the normal proteins 6.1%. These interpretations specified that the cancerous proteins are more probable to be network hubs compared to the essential or control proteins. Table 4.11 lists few of the major hubs. More information on result can be retrieved from http://www.bioteccancernet.cusat.ac.in/hub.

Table 4.11: Major hubs

List of major hubs
AR
ATM
BRCA1
BRCA2

CDH1
GARS
HEXB
KRAS
LMNA
MSH2
PTK3CA
TP53
MADTL1
RAD54L
VAPB
CHEK2
BSCL2
BRIP1

4.4.3 Betweenness centrality

Betweenness centrality was the next factor calculated for this study. The number of shortest paths transient through the node is the betweenness centrality of a node. Or in other words it embodies the significance of a node in the network. In a biological network, this calculation reveals the degree of signals that possess paths across the node.

The average value for betweenness was generated as 761 and 141.3 for cancer and non-cancerous proteins respectively. Fig 4.18 portrays the ratio of cancer and non-cancerous proteins for numerous values of betweenness. The figure supports the fact that the value of betweenness for cancer proteins is greater compared to the non-cancerous proteins.

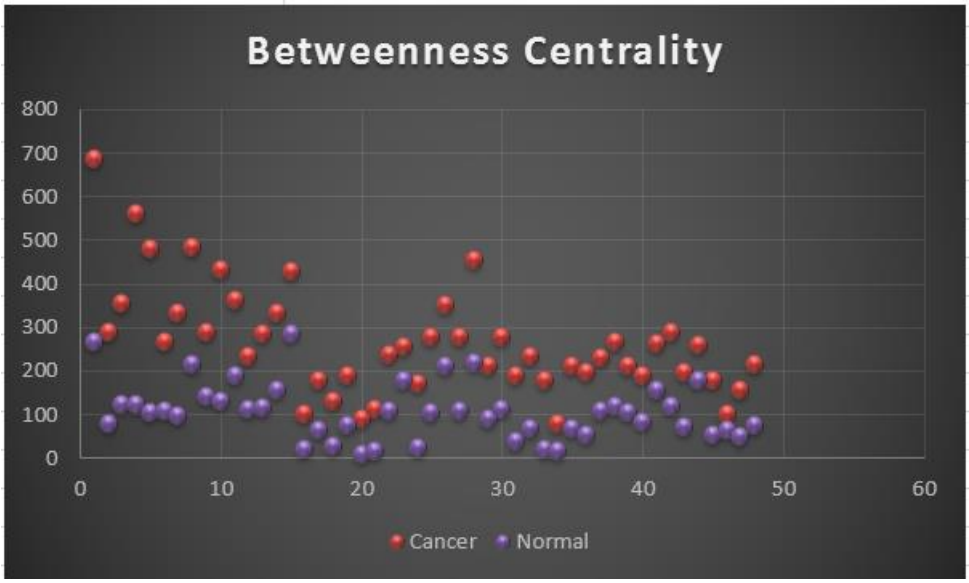

Fig 4.18 Betweenness centrality

In the graph, purple color indicates the normal proteins and red color indicates the cancer proteins. The left hand side displays the number of proteins and right hand side indicates the betweenness centrality.

All the 785 cancer proteins displayed the value of betweenness to be higher than zero. The average value of betweenness for the cancer proteins is 8.73 × 104, which is considerably higher compared to the essential proteins, that is, 4.96 × 104.

4.4.4 Clustering coefficient

The ratio between the number of neighboring edges present and the number of possible neighbors is the clustering coefficient. This measurement aids in understanding how good a node is connected to its direct interactors. The higher density of a network connection is indicated by a node's higher clustering coefficient. It was observed that the value of clustering coefficients for around 785 cancerous proteins was within the range of 0-0.3. Remarkably, the normal proteins displayed higher

proportion when the value for the clustering coefficient was 0 or >0.9 and cancerous proteins displayed the lowest proportion. Generally, the neighbors of cancerous proteins connected with each other in a lower rate compared to the non-cancerous proteins. It was evident from many other investigations that the average clustering coefficient of cancerous proteins are lower compared to the normal proteins. Fig 4.19 portrays the clustering coefficient of cancer and normal proteins.

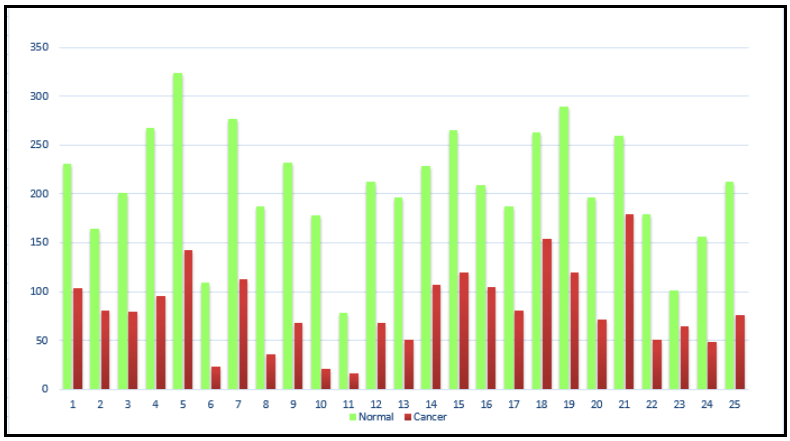

Fig 4.19: Clustering coefficient

In the graph, green color indicates the normal proteins and red color indicates the cancer proteins. The left hand side indicates the number of proteins and right hand side indicates the clustering coefficient.

4.4.5 Shortest path length

The shortest path length from one node to other nodes in a network was also considered for this investigation. The average value derived for cancer protein was 4.11 and normal proteins 5.67 respectively. This makes it evident that the path from cancer proteins to other proteins is shorter compared to normal proteins to other proteins. This comparison specifies that the competence of cancer proteins interacting with each other could be greater than the normal proteins in the human protein interaction network.

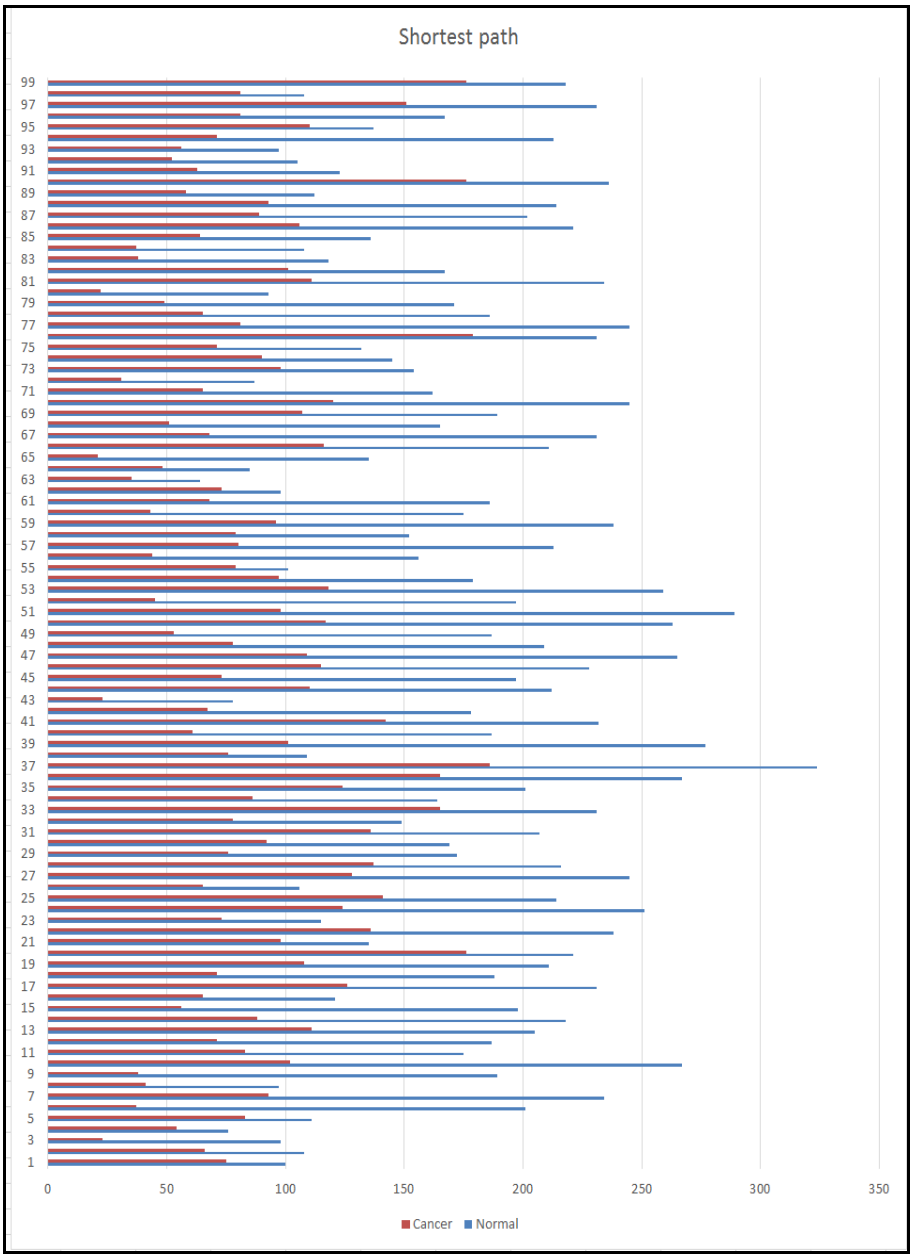

Fig 4.20: Shortest path length of cancer and normal proteins

In the graph, the red color indicates the cancer proteins and blue color indicates the normal proteins. The right hand side indicates the number of proteins and left hand side indicates the shortest path length.

Fig 4.20 displays the ratio of cancer and normal proteins for shortest path distances. The figure substantiates that the shortest path distance of cancer proteins is much shorter than that of normal proteins.

4.4.6 Pathway analysis

All the proteins present in each of the 99 clusters were identified. All the identified protein was further added to the KEGG database in order to detect its pathway. This pathway offers effective and valid information to find out the phenotype classification as well as the linking pathway. 1141 cluster proteins are unique and there is no physical interaction among the clusters. But, the clusters are linked with cancer pathway as portrayed in fig 4.21.

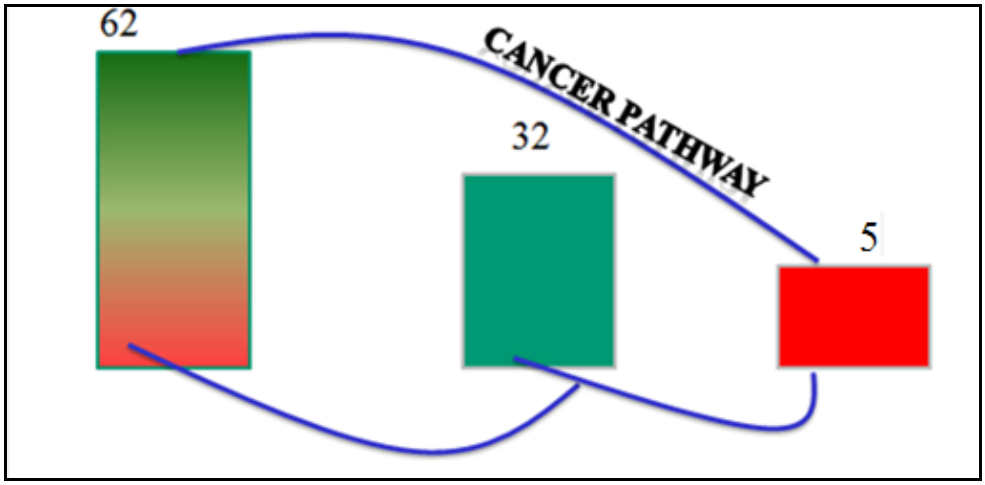

Fig 4.21: Cancer pathway

(62 represents the cancer and normal protein clusters. 32 represents the normal protein clusters and 5 represents the cancerous protein clusters.)

It is observed that, among all the identified pathway cancer pathway owns more protein from complexes. Fig 4.22 portrays the major l1 linking pathway. The highest number of cluster protein is observed in prostate

linking pathway. The cluster proteins are not connected with the cancer proteins physically but they are extremely connected with the cancer linking pathway.

Linking Pathway	Participents from CANC complexes	Cancer	Total No: disease associated
Colorectal	62	Colorectal	5
Pancreatic	66	Pancreatic	1
Glioma	65	Glioma	1
Thyroid	29	Thyroid	1
Acute myeloid leukemia	57	Acute myeloid leukemia	1
Chronic myeloid leukemia	73	Chronic myeloid leukemia	1
Basal cell carcinoma	55	Basal cell carcinoma,Basal cell nevus syndrome	2
Melanoma	71		1
Renal cell carcinoma	66	Renal cell carcinoma,von Hippel-Lindau syndrome	1,Renal cell carcinoma 2,von Hippel-Lindau syndrome
Bladder	38	Bladder	1
Prostate	89	Prostate	1

Fig 4.22: Linking pathway

The phenotype cataloguing was done for the cluster proteins by employing KEGG (BRITE). The key category of cancer was recognized using the data from web as well as literature survey (McLendon *et al.*, 2008). A total of 12 chief cancer causing proteins were recognized from the cluster against the cancer. Table 4.12 displays the CANC complexes listing 12 major cancer causing proteins.

Table 4.12: CANC complexes

Cancer	Participants from CANC complexes
Brest Cancer	58
Ovarian	52
Prostate	49
Leukemia	47
Pancreatic	41
Lymphoma	35
Glioma	34
Skin	28
Lung	28
Stomach	25
Thyroid	23
Brain	19

4.5 CancerNet tool

CancerNet tool is web based and works with sql data base of size 2 GB. The front end uses Perl and PHP, the backend uses mysql and the operating system used is Ubuntu. The data captured from the tightly packed human protein complex (HPC) is integrated to the tool. CancerNet retrieves all relevant information from the protein symbol to interacting partner. The

query passes through the perl-cgi. The query is processed using the SQL language and the required information is fetched from the tightly packed HPC database.

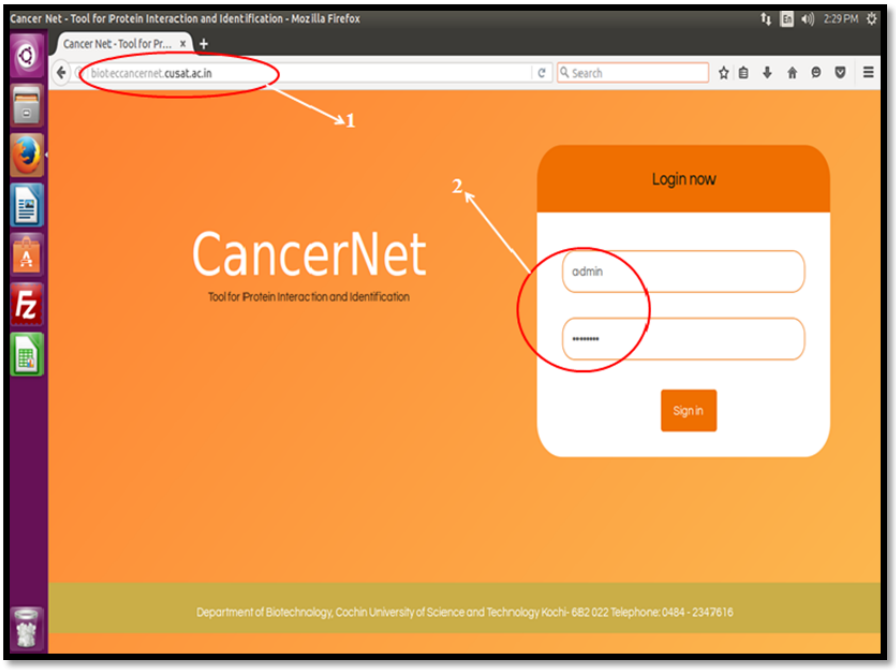

Fig 4.23: Login page of CancerNet

1 highlighted in the image indicates the server name to access the tool and 2 highlighted in the image indicates the fields to enter the login credentials.

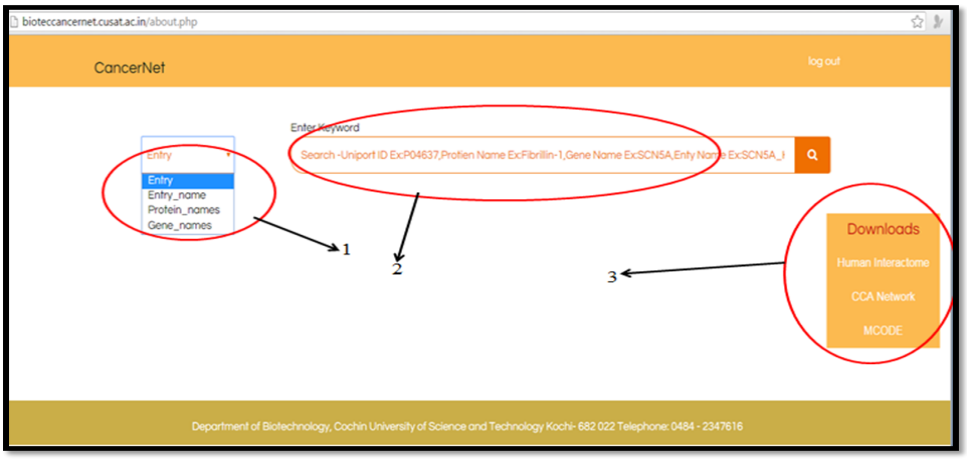

Fig 4.24: Homepage of the CancerNet tool

1 indicates the list of entry, 2 indicates the query field and 3 indicates the download option

The user can select any one of the following four different types of entries:

1. Entry
2. Entry_name
3. Protein_names
4. Gene_names

The entry is actually the uniprot ID which is unique and is used as the primary key in the data base. Entry_name is the organism combined with the uniport ID. After selecting the type of entry, the user can search for the required protein by entering the details in the query field. User is directed to the output page after submitting the query. Download is a beneficial feature, which is added to the tool to download the analyzed human interactome, CCA Network, and MCODE.

The default web pages of the CancerNet tool is mentioned in Appendix II section. Fig 4.25 displays the category and entry fields in the tool.

4.5.1 Query using the CancerNet tool

Fig 4.25: Query field

After selecting the "Protein_names" in the "Select Column" field, the user has to enter the name of the required protein. For example, in the fig 4.25, the user has entered Receptor tyrosine-protein kinase erbB-2protein.

Fig 4.26 displays the output page after entering the protein name. This includes information about the gene ontology and pathway of the queried protein. The gene ontology covers information such as biological process, molecular function, cellular component and Kegg pathway of that protein. In addition, the preliminary information such as the number of interacting partner, gene origin, and related diseases of the protein is also displayed.

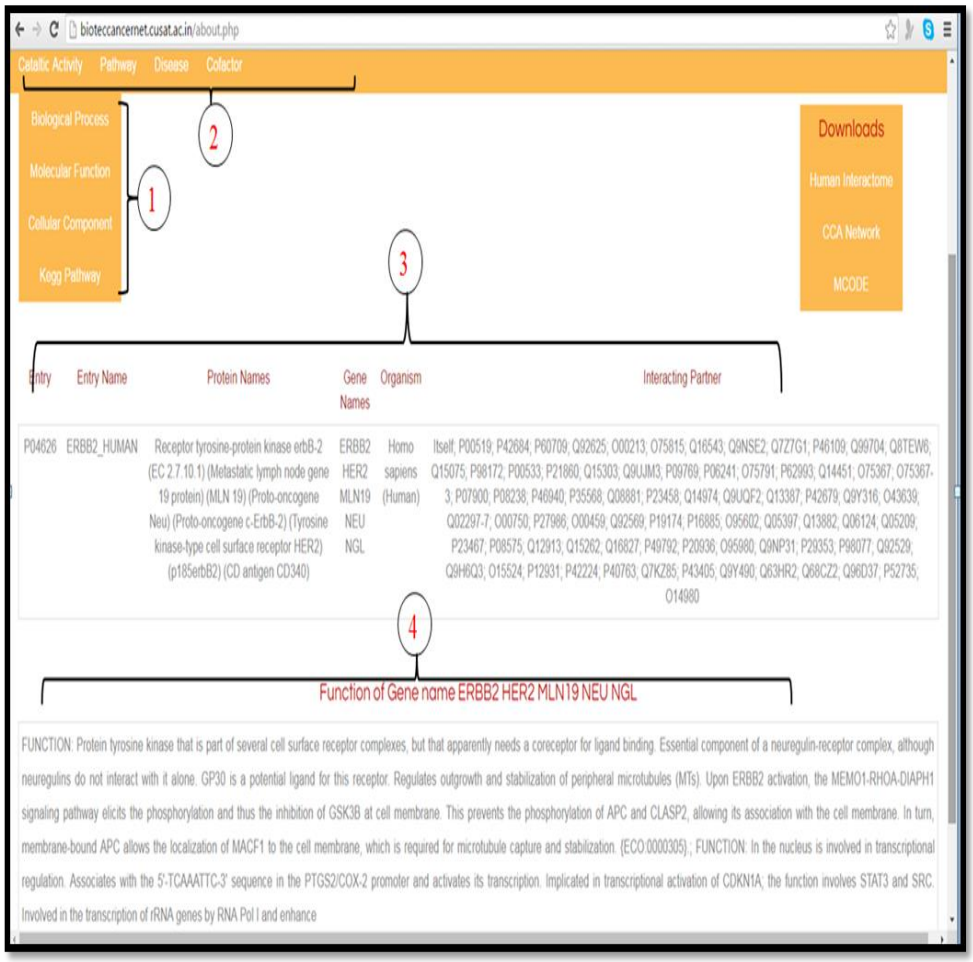

Fig 4.26: Output page

1 and 2 indicates the information fileds which works after clicking the appropriate field. For example, information related to Biological process is revealed after clicking the Biological process tab. 3 and 4 indicates the preliminary information which is displayed by default, that is, the number of genes and its functions as well as the proteins and its interacting partner.

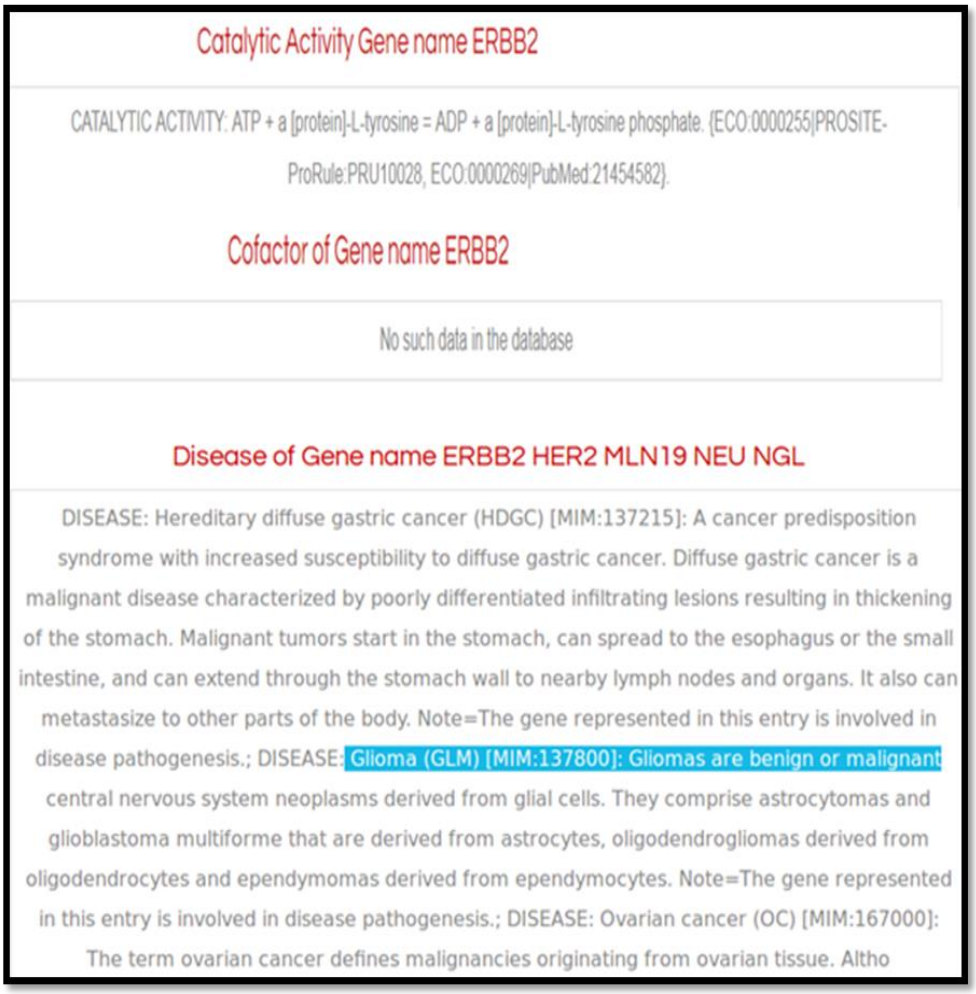

Fig 4.27: On click information: Catalytic activity, Cofactor, and involvement of disease.

The figure displays the information about the gene such as catalytic activity, cofactor and the related disease. Catalytic activity and cofactor of a gene helps in understanding the interaction of proteins. Also, the information on the associated disease is retrieved.

Fig 4.28: Gene ontology annotation of Gene ERBB2

Biological process, cellular component and molecular function are the three annotations covered in the CancerNet tool. The blue highlights indicates the gene annotation ID.

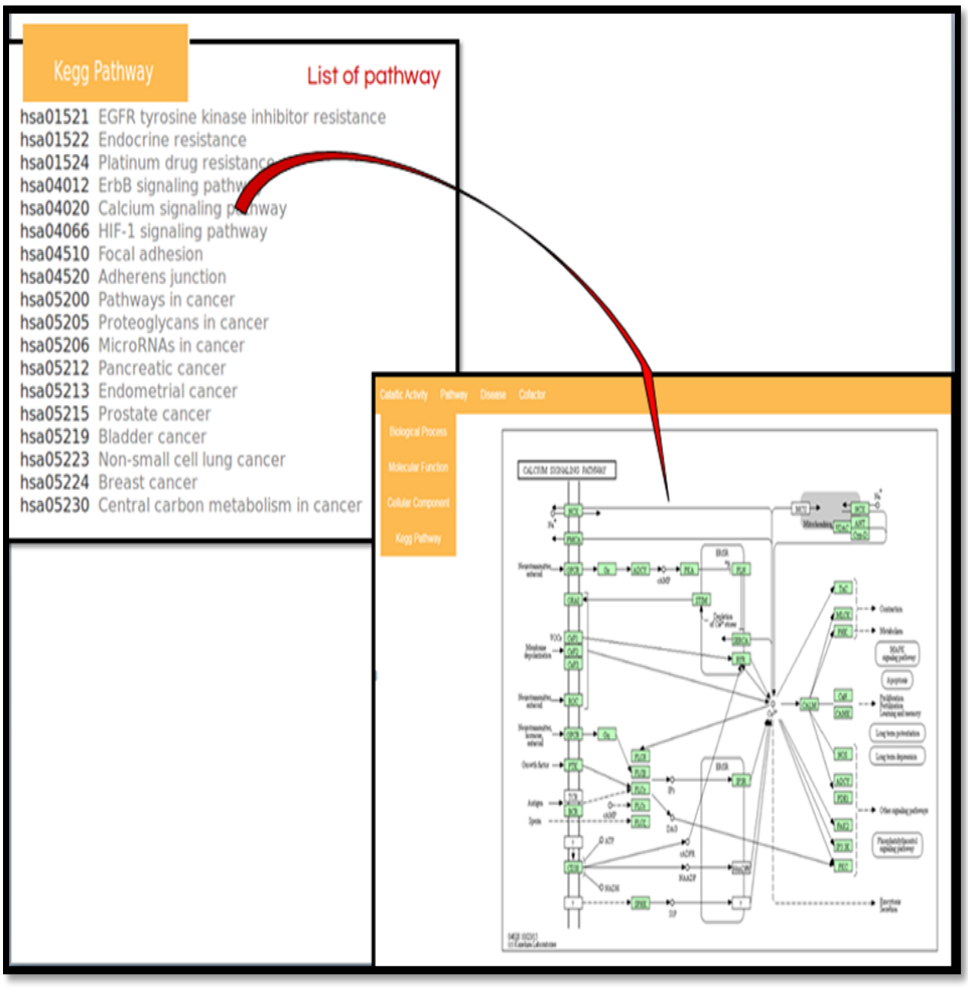

Fig 4.29: List of pathways associated with KEEG

For example, list of ERBB involved pathways. When the user clicks a particular pathways, it displays the respective graphical representation.

The figure lists the associated pathways of the queried protein. One protein can participate in multiple pathways. The researcher can select the desired pathway from the list of pathways displayed. After selecting the interested pathway, the detailed schematic representation of the protein pathway is retrieved along with other proteins which participate in that pathway.

Chapter 5
DISCUSSION

In this study, a novel methodology is proposed to identify cancer and non-cancerous protein complexes in human interactome. This main focus is to display an exceptional platform for investigating the currently existing human interactome and its respective modules. Also, to assimilate the data of protein interactions into various prominent databases and study the tightly packed protein complexes. The further analysis demonstrates the discrepancies in topological properties of cancer and non-cancer proteins. The results of this study demonstrated that investigation of the interactions leads to significantly better predictions about the cancer proteins. Finally, a one-stop solution was introduced to retrieve major cancer and non-cancerous complexes.

5.1 Human interactome created from 79,950 human protein interaction using 4 databases: NCBI reference sequence, Human atlas, Plasma proteome database and Uniprot

In this study, a total of 4 databases were utilized to study the human interactome. This was accomplished by employing a strategical approach.

In the investigations allied to protein interactions, Bader *et al.*, (2004), Jansen *et al.*, (2003) and Lin *et al.*, (2003) advocated that the non-predicted data can be integrated to a prevailing database in an administered learning context. Jaimovich *et al.*, (2006) recommended the alternatives of Markov networks to scrutinize the information on protein interaction. Deng *et al.*, 2004; Letovsky and Kasif, 2003, and Leone and Pagnani, 2005 predicted that the interaction network can be extracted and used it for other

tasks to predict the functions of protein, topological properties, malignancy due to proteins, in grouping the interacting co-expressed proteins (Segal *et al*., 2003), etc. It was remarkable that the overlap inferred from the diverse methodologies was very minute necessitating that the data extracted from the different high-throughput investigation must be manipulated with extreme care (Bader and Hogue 2002; Schachter 2002; von Mering *et al*., 2002).

All these methodologies employs a cataloguing algorithm to assimilate assorted biological datasets. A classifier or the person who distinguishes non-interacting pairs is educated to differentiate among positive or constructive instances of accurately interacting pairs of proteins from the negative instances. Each pair of protein is encrypted as a feature vector where the features embody a specific data source concerned with either PPIs, domain configurations, associated mRNA expressions, or proof extracted from investigational techniques. In the current investigation, 10078 unique proteins were mined from the comparative study of protein data from four prominent databases.

The most challenging and perplexing task is to attain computational replicas for the PPI mechanism. Gray *et al*., (2003) revealed from their investigation that other computational approaches for determining PPI spots are categorized into two groups. The first is the docking method that attempts to map two protein structures to discover the paramount locations on both the structures. These methodologies are pragmatic only to the unraveled structures of protein and are currently prevailing merely for a smaller number of proteins. Li *et al.,* (2003), Kim *et al*., (2006), and Marti-Renom *et al.,* (2007) advocated the use of other techniques if the proteins lacked any homology with solved structures of protein. The other method

was the local sequence interaction prediction method. This method typically employs an algorithm to identify interacting sites. In the above researches, a single method was assigned to study a particular trait of the protein. In the current investigation, data from multiple methodologies were employed to study the specific trait of a protein.

The human proteome is more complex because of the larger number of proteins, post-translational amendments, their splice isoforms, and mechanisms of kinetic protocols when compared to yeast. While almost all the preceding investigations were concentrated on the relative assessment of PPI network in yeast. A broad assessments of human PPI maps using multiple data integration was deficient. A straightforward hypothesis of the outcomes of the yeast maps to human maps could be deceptive concerning the diverse primary biology and methodologies towards mapping. Consequently, a methodical assessment of presently accessible human PPI maps is merited to achieve an enhanced insight into the functional configuration and topological assembly. Based on this, a relative examination of existing human PPI maps was accomplished in this effort. Rhodes *et al.*, (2005) applied a stratagem by employing the sum of probability ratio scores policy to envisage human protein interaction. This probability ratio scores are resultant of the homologous PPI.

The initial endeavor was implemented around 2005 to scrutinize large-scale assimilative mapping of human interactome (Stetzl *et al*., 2005). A protein matrix containing 4500 baits and 5600 preys in a yeast two hybrid system was used to portion together the interactome. Rual *et al*., (2005) executed an analogous yeast-two hybrid investigation and substantiated with association of other biological traits and co-affinity

purification, confirming more than 300 connections to 100 ailment-related proteins.

In computational biology, the worth of human interactome depends on the data collection. Kim *et al*., 2014 and Wilhelm *et al*., 2014 used a few databanks to extract and interpret human PPIs with agreeable publication records. The outcome was 14000 discrete interrelating pairs of protein. The data was consolidated directly from binary databank reserve. In the framework of a network, Walhout and Vidal (2001) understood that the proteins and their respective binding partners provide a stronger perception about the cell functions. More than 10000 direct and exclusive PPIs were annotated in HPRD. This data was derivative of innumerable discrete small-scale experiments printed in literature.

In the current investigation, 10078 proteins and their corresponding 58674 protein interactions were identified by employing four conspicuous databases. In the above researches, the study on human interactome was initiated by integrating the interaction database directly to yield the interactome. But, interactions from the protein-level and protein to PPI mapping was employed in this study. Consequently, it was possible to identify 127 missing interactions. Some of the scholars affirmed the fact that the interactome can be plotted soon by taking benefit of the introduction of reference proteome maps.

According to National Physical Laboratory (NPLI), Technical bulletin, (1987) PPIs are considered as the ultimate to vitally all cellular procedures. The PPIs own enormous substantial functions like devastating a protein, ensuring the creation of a novel binding location, transforming the precision of a protein for its substrate, or fluctuating the dynamic features of proteins. The preceding eras manifested numerous major

milestones to recognize PPIs and by this means exploring more details about these complicated biological systems.

Feldman *et al*., (2008), Bauer-Mehren *et al.*, (2011) and Taylor *et al*., (2009) revealed from their studies that PPI networks have been exploited to achieve acumen into the ailment-related mechanisms, to ascertain new network-based biomarkers, and to regulate the targets of drug. Chatr-aryanmontri *et al*., (2008) interpreted that the vigilant elucidation of PPI information from multiple databanks delivers biologically relevant implications.

5.2 3146 cancer protein interactions identified from human interactome

Walhout and Vidal (2001) confirmed from their study that proteins do not perform as remote entities often but are usually part of a greater protein complexes inside a cell. Any proteomic analysis owns an essential trait of interpretation of interacting proteins, the interactome, and in plotting the analogous binding locations. Based on this result, in the current analysis the human interactome was subjected to the identification of protein complexes.

Bader and Hogue, (2003) and Spirin and Mirny, (2003) also ascertained from their investigation that, the modules in PPI networks discloses both the necessity for removal of indiscreet procedures in addition to the compact interaction among proteins to execute explicit functions. Rives and Galitski, (2003) and Spirin and Mirny, (2003) assimilated a set of interaction network to identify the compactly linked modules of interacting proteins. This culminated in identifying informative modules from interactome.

Jain *et al.,* (2005) and Jain *et al.,* (2012) confirmed from their investigation that the protein complexes which perform explicit biological functions are often found to encompass extremely allied protein modules. The investigations on these protein modules owns a fundamental role in understanding the pathophysiological characteristic properties of intricate ailments such as cancer.

Spirin and Mirny, (2003) also projected that the affiliated sub-network possess 783 proteins and 644 respective interactions with the largest connected component (LCC) in the alignment encompassing 318 proteins and 318 interactions.

In one LCC, remarkably 44% of the nodes, demonstrating 284 proteins, were linked ($p<1\times10-16$, permutation test), signifying that though the proteins are engaged in incongruent roles, most of them are linked by PPIs. Several biological types are epitomized in the LCC of the interactome comprising of signaling of stress, repairing of DNA, transport of vesicle, modification of chromatin, in addition the metabolism of proteins, RNA and lipids. Almost all the functional groups embodied in yeast are also epitomized in the human network and the network is extremely modular.

Nikolaus *et al*., (2014) experimented on the murine and human proteolytic networks and projected that most of the proteins are connected and very few are in not connected components. Therefore, when both the networks are compared, the LCC, that is the major cluster of nodes directly or indirectly linked comprises the enormous mainstream of these proteins. The numbers are derived as 1377 of 1393 (99%) in murine and 1183 of 1230 (96%) in human. In the current investigation, 9886 nodes and 58568 interactions were found in the connected component network. Compared to the study conducted by Nikolaus, in the current investigation, 9886 nodes

and the respective 58568 interactions were identified in a connected component network. The average connectivity of a node in the network is 11.849 as the number of proteins and their interactions are high in the current preliminary set under study.

The complexes encompassing manifold PPI partners performs many cellular functions. One of the essential units in PPI networks are the molecular complexes and predicting them is one of the utmost tasks in the investigation of PPI networks. Ho *et al.,* (2002), Mering *et al.,* (2002), and Gavin *et al*., (2006) found that the high-output investigational methodologies employed to regulate the complexes of protein sets complexes on a proteome-wide scale usually undergo problems with false negative rates and high false positive. Therefore, innumerous computational efforts were made hitherto to categorize associated functional modules or complexes. Amau *et al.,* (2005), Adams *et al.,* (2006), and Chu W *et al.,* (2006) proposed that the non-administered graph clustering approach was employed to identify the automatic complex or to detect the related functional module in majority of the hitherto methodologies and Aittokallio *et al.,* (2006) also attempted to ascertain likewise or compactly connected subgraphs of clusters or nodes.

Numerous studies tried to fragment the PPI graph into separate extremely connected complexes or clusters. King *et al.,* (2004) segregated the nodes of a specified graph into discrete clusters, grounded on their contiguous interactions, by employing a local search algorithm based on cost. Dunn *et al*., (2005) divided the network into complexes by eliminating the edges with the maximum centralities. The process of edge-removal repeatedly revalidated betweenness till a stable number of edges were eliminated.

Spirin and Mirny (2003) found in a preceding study conducted on yeast that in an interaction network, the extremely interrelated augmented groups did not develop fortuitously.

The complexes or the compactly connected areas in a network are derived similarly to finding clusters. Md Altaf-Ul-Amin *et al.,* (2006) displayed and scrutinized few complexes of proteins generated by the projected algorithm from an archetypal PPI networks of yeast and E.coli.

The approach of King *et al.,* (2004) was initiated by combining a maiden haphazard clustering and repeatedly passing one node from a cluster to other cluster in an arbitrary manner to increase the cost of clustering. The clusters are filtered grounded on the cluster size, compactness and functional homogeneousness after generating the clusters bearing in mind the standards of the recognized biological groups.

Becker *et al.,* (2012) found that the algorithms like OCG having the capability to group proteins into manifold graph modules permits the recognition of multifunctional proteins. Hartwell *et al.,* (1999) confirmed that these modules are similar to the functional components of the network and comprises clusters of extremely connected proteins engaged in the equivalent cellular function.

The fact is noteworthy that such preventive delineation of modules directs to an augmented strength of the identified components concerning the false positive interactions. Spirin and Mirny (2003) proposed that when an extensive proportion of PPIs are eliminated, the recognized components still create extremely connected groups.

Palla *et al.,* (2005) and Jonsson *et al.,* (2006) found that the identification of PPIs which occurs inside similar cellular processes were executed hitherto by clustering methods. According to Jeong *et al.,* (2001)

and Gunsalus *et al.*, (2005), this can be validated from the fact that the protein subnetworks engaged in a demarcated cellular process is profoundly interrelated by direct PPIs compared to the ones anticipated fortuitously.

Wuchty and Almaas (2005) arrived at the conclusion that though connectivity provides a hint of one protein's prominence, the cataloguing of topological role of highly connected proteins grounded on their locality is also probable. In a real network, it is not necessary that all the subgraphs are equally substantial or significant. Motifs are considered to be the subgraphs that befall more frequently in a specified network compared to only the anticipated fortuitously. In a lot of previous investigations executed by Yeger-Lotem *et al.*, (2004)], the presence of simple motifs of building network in PPI graphs and transcription regulation networks were postulated. Shen-Orr *et al.*, (2002) and Milo *et al.*, (2002) recently confirmed that there are quite a lot of effectual tools designed to enable the recognition of motifs. Many complex networks exhibit firm dogmas of structural design. Aittokallio *et al.*, (2006) network motifs is very constructive to the researchers as they aid them to recognize the rudimentary structural features of an explicit network. In the yeast protein interaction network, there is an elevated amount of evolutionary preservation of network motifs. In a network of miscellaneous species, the confluence evolution to the equivalent motif categories was also witnessed in the transcription-regulatory network. Barabasi *et al.*, (2004) confirmed that all these interpretations has an incredible relevance to the biological studies.

5.3 Connected component and molecular detection method finds 99 major protein complexes in human interactome

In the current investigation, the result shows stronger connectivity. 100% shortest path is included in the connected component network. The average shortest path between two nodes in the CCA network is 4.054. The network diameter is 15 and the average connectivity of a node in the network is 11.849 nodes which displays stronger connectivity of each node in the CCA network. Involvement of rich complexes in CCA network was recognized based on the current investigation, when compared to similar investigations.

In the investigation by Gavin *et al*., (2002), fresh data extracted from 588 biochemical purifications were epitomized by applying the spoke model to retrieve 3,225 conjectural PPIs amid 1,363 proteins as input to MCODE. Gavin *et al*., (2002) reported a list of 232 manually interpreted complexes of protein based on the unique refinement data. As part of a larger complex, this data was filtered to eliminate five identified complexes. Each of these complex encompassed single proteins. Also, six complexes encompassing two or three proteins which are existing in the data set. In a set of complexes, a filtered set of 221 complexes was generated to evaluate MCODE though few of these complexes displayed substantial overlap with other complexes.

The study of Gavin *et al.,* (2002) postulated the competence of MCODE to recognize vital complexes from the connected interaction network. In the current study also, MCODE was employed to identify clusters in the CCA network. 121 highly linked biological modules from massive CCA protein network was identified by employing the MCODE algorithm. The high-scoring clusters exhibit high density value. 121

complexes encompassing 1550 unique proteins were selected and 14873 interactions were derived.

There were a lot of methodologies permitting cluster overlap as few proteins are part of functional components or manifold complexes. Bader *et al.*, (2003) attempted to identify compactly connected areas in a large PPI network by employing vertex weights to embody indigenous neighborhood density with the MCODE method.

Brohee and van Helden (2006) did a relative validation for PPI networks of four clustering algorithms, that is, Super Paramagnetic Clustering (SPC), Markov Clustering (MCL), Restricted Neighborhood Search Clustering (RNSC), and Molecular Complex Detection (MCODE). They concluded that RNSC, MCL, and MCODE were stronger for the graph amendments compared to the other two algorithms.

In the current analysis, 99 MCODE clusters were considered having a density score value >2. A total of 99 complexes having total 1141 unique proteins were selected and the total interactions was derived as 14340.

Franke *et al.*, (2006), Masseroli *et al.*, (2005) and Van Driel *et al.*, (2005) lately witnessed the advent of assimilative procedures to recognize plausible genes of maladies in the interludes of linkage concomitant to the sickness based on the assimilation of information like the information on expression and Gene Ontology categories.

A hint on the disparity in evolutionary traits of cancer and non-cancer proteins can be retrieved from the detection of a variance in the number of collaboration associates among the two groups. Fraser *et al.*, (2002), Eisenberg and Levanon, (2003), and Wuchty, (2004) undeniably confirmed that the age of proteins or genes and rate of evolution was the topic of numerous scientific publications and also showcased growing

proof for an association among the number of collaborations and age of proteins. Jordan *et al*., (2003) was doubtful with this association and published about that in a latest publication. According to Saeed and Deane (2006), a correlation exists, even though it is reliant on the comprehensiveness and eminence of the database under investigation. All these results conclude that cancerous proteins, which causes detrimental transformation of functions could be older compared to the non-cancerous proteins.

Jeong *et al*., (2001) indicated that, in yeast the extremely connected proteins are the ones which have phenotypical significance, and grave on the survival aspect of an organism. Saidet *et al*., (2004) described from their investigation that the toxicity-controlling proteins demonstrate a vast amount of interactions. In cancer proteins, the augmented connectivity recommends that they execute a significant role in a protein network.

In recent times, more biological milieu for the PPI networks were incorporated by a lot of imperative researches by expanding or mining PPI graphs into explicit supersets or subsets. Goh *et al*., (2007) illustrated the 'human disease network' which encompassed the maladies and ailment genes concomitant by identified disease-gene binding. This network discovered a common genetic origin of many ailments by exploring all known phenotype and disease-gene bindings in just one graph hypothetical background.

In a PPI network, Radivojac *et al*., (2008) attempted to detect disease–gene linkages by encrypting all the genes grounded on the dispersal of shortest path lengths to each genes supplementary with ailments or owning acknowledged functional interpretations.

In the current analysis, 4 clusters encompassing only cancer proteins, 32 clusters encompassing only normal proteins and 62 clusters encompassing both cancer and non-cancer proteins were grouped. We identified 1141 proteins in 99 clusters. In them, 785 proteins are cancerous and 356 were normal proteins. This study indicated the involvement of diseases in selected clusters which is in alignment with the study done by Jeong.

Jonsson and Bates (2006) proposed that an open debate is ongoing about the differentiation of genes existing in precarious aliments like Cancer based on their location in a PPI network and characteristic properties. The main example is the centralities and better connectivity of the cancerous genes when compared to the normal genes. It is observed that the number of proteins the cancer proteins interact with is explicitly higher and also contribute in central hubs compared to the peripheral ones, reflecting their participation and better centrality in networks which contributes to the backbone of a proteome.

5.4 Topological property differences exist between cancer and normal proteins

Jeong *et al*., (2001) confirmed from their analysis that the identification of network characteristics like degree distribution, connectedness, etc common across multiple PPI networks was the novel comparative investigations of PPI networks. It was hypothesized from this study that a scale-free topology is followed in most of the PPI networks. A Power-law distribution, f(d) ~ cd_, is followed by the degree distribution f(d) of the nodes in these networks. f(d) is represented as the nodes frequency with d as degree. Similar investigations postulated the significant

roles conveyed by the high-degree hub proteins in a PPI network. Almost all the paths are directed through such hub-proteins in a PPI network.

Aittokallio *et al*., (2006) postulated from his investigation that if the degree distributions of a complete network and arbitrarily experimented subnets disclose the identical family with respect to probability distributions, then the chances to generalize from subnets to the properties of a whole network is possible. Barabasi *et al*., (1999) postulated that the real-world networks possess short average path lengths and degree distributions.

Jeong (2001) studied the degree distribution of proteins at network level and derived a high value for cancer proteins. In the current investigation, the study was at the protein complex level and also derived a high value of degree distribution for cancer proteins compared to the non-cancerous proteins. Also, the properties like shortest path length and clustering coefficient was analyzed. It was concluded that the cancer proteins displayed lesser shortest-path distance and clustering coefficient.

Samanta and Liang, (2003) in their earlier study in the PPI network of yeast illustrated that the function of proteins partaking an abnormally vast neighboring proteins as it is possible to forecast functional links among proteins. The algorithm used to extract the high-throughput PPI data are not sensitive to false positives or noises, which serves as the advantage.

Kim *et al*., (2006) differentiated PPIs by utilizing the atomic-resolution information from the 3D structure of proteins. Lage *et al*., (2007) also recommended a structural measure to investigate the segmented PPI hubs and offered acumen about the evolutionary rate of these hub proteins. Lage also produced a phenome-interactome network by assimilating interactions of human proteins which are quality-monitored with an

authenticated and phenotype similarity score derived computationally. Consequently, it was possible to identify the formerly anonymous complexes probable to be connected with ailment. Linding *et al*., (2007) developed a methodology on similar lines known as NetworKIN.

During the initial investigation of Jeong *et al*., (2001) and Han *et al*., (2004), they discovered that the hub proteins were probable to be programmed for indispensable genes in a PPI network. Instead of studying the protein's pairwise interaction, the study at complex level should be focused for a better comprehension of a protein network and its essentiality. Hart *et al*., (2007) of late accomplished a similar investigation and substantiated that indispensable proteins are intense in some of the complexes, which is also ubiquitous in anticipated complexes.

Guerrero *et al*., (2008), Milenković and Pržulj (2008) demonstrated from few illustrations that proteins owning analogous topological vicinities display similar biological features. Jonsson and Bates (2006) concluded from their investigation that cancerous genes display inordinate centralities and connectivity equaled to the non-cancerous genes. However, Goh *et al*., (2007) proposed that the affiliation among ailment genes and the respective network degrees require a better and vigilant contemplation because most of the sickness induced genes lack the inclination to program for hub proteins.

In the current investigation, cancer proteins displayed higher degree distribution. Kim, Lage and Jeong identified the high degree nodes as hub proteins which has the tendency to cause cancer. In the current investigation, high degree was noticed in cancer proteins and the further analysis led to the conclusion that hub proteins are cancerous.

Hua Li (2010) enhanced the forecasting system by creating a novel algorithm and methodologies. This algorithm was employed on a human PPI network to create a genome-wide well-designed interpretation. This algorithm sided to calculate and lower the impact of hub proteins on perceiving functionally linked proteins. In the human PPI network, the interpretations of GO and KEGG was used as a self-governing and impartial standards to assess the algorithm introduced by Hua Li. Hua Li highlighted that his study enriched the complete excellence of functional implications for human proteins.

In the current investigation, 1141 proteins from the complexes were identified and were added to the KEGG database in order to detect its pathway. 521 KEGG pathways were derived. Hua Li did not extend his pathway analysis in the field of cancer network. But, few linking pathways were observed on the current investigation. Majority of the linking pathways were associated with cancer pathway. In the cancer pathway, the participation of proteins from the complexes were considerably higher compared to the normal pathway. This study was further extended to carry out the phenotype classification of cancer using KEGG (Brite) and we concluded that Breast cancer was more dominant compared to other cancer types.

Hua Li further assigned 466 KEGG pathway interpretations to 274 proteins and 123 GO interpretations to 114 proteins. Finally, Hua Li was able to group 1729 proteins according to their functional relations and executed pathway analysis to further classify numerous subclusters extremely augmented in definite signaling pathways.

In the current study, 1141 proteins were added to the GO analysis. It was found that most of the proteins were located at the nucleus component.

234 proteins were linked to cancer. In the cellular component, 234 proteins from 686 represented cancer. In the biological process, 143 proteins exhibited cancer from 324 proteins. In molecular function, 101 proteins from the 267 DNA binding proteins were cancer proteins. 89 proteins from the 175 RNA binding proteins were cancer proteins. The number of proteins used for GO and KEGG analysis was higher compared to Hua Li because the datasets retrieved from the proposed method is greater.

5.5 CancerNet tool developed for protein and its related information

The number of human PPI databases are increasing day by day but their application to medical and biological field are restricted because of the dissimilar information (Yildirim *et al.*, 2007; Goh *et al*., 2007; Ideker and Sharan, 2008 and Braun *et al*., 2008).

A common platform with direct access is necessary to access the PPI data. Hence, CancerNet was developed to provide a one stop solution to retrieve information on human interactome using one assimilated platform. In CancerNet, not only the interacting partner but also 15 other information like catalytic activity, type of disease, cofactor, biological process, molecular function, cellular component, KEGG pathway, name of the related gene, name of the protein, function of the protein, uniprotID, name of the organism, MCODE data, CCA data, and human interactome is displayed. In CancerNet, the KEGG pathway information is displayed in the same window and not in the KEGG website. In HIN, noise removal is supported which served as an advantage to the scholars to retrieve accurate data.

Chapter 6
SUMMARY

The wide-ranging replica of molecular processes are reinforced by large-scale maps of protein interactions. Likewise, an organized demonstrating methodology of cellular processes are based on entirely sequenced genomes supporting currently as a requirement for complete maps of PPIs. The advancement in revealing interactome was slower in comparison to the extremely efficiency of mapping genome projects, particularly for the human interactome. In recent times, there is a mounting number of both computational and experimental exertions to improve the systematic maps of human interactome. In experimental and computational approaches, both have their own pros and cons. Hence, vigilant validation of these maps are obligatory to circumvent prejudices in experimental approaches and elevated rate of false positive interactions in discrete maps.

This study displays an exceptional scaffold for investigating and assimilating the currently existing human proteins and its integration databases. Also, other methods result in the prediction of tightly packed protein complexes. The further analysis demonstrates the disparities in topological properties of cancer and non-cancer proteins.

The diverse information of human protein from various databases such as Uniprot, NCBI Refsq, Plasma proteome and human protein atlas are integrated into a single database in a single format since different databases follow different access number. A total of 9432 unique proteins were recognized. The key intention of this study was to recognize all the existing binary interaction among the 9432 unique proteins. The

assimilated human protein database such as HPRD, IntAct, MINT, and DIP was scrutinized against the integrated human PPI database to procure a human interactome prototype. These interactions were derivative of either Y2H-assays, literature reviews or deduced on the foundation of homologous interactions in other living organisms. The investigation displayed that the existing maps have just a trivial, but a substantial overlay. However, the majority of proteins interaction can be found in multiple databases. The omitted protein interactions in cohesive databases were recognized by using orthologous based quest. The databanks or the datasets used in the research possessed diverse criterions and formats for the nomenclature of PPIs owned. So, to retain the uniformity in nomenclature standards, the subsequent PPIs were apportioned a unique ID by employing personalized Perl programs. Finally, 76072 interaction data was identified and entitled as Human Protein-Protein Interactions (HPPIs).

Cytoscape was used to investigate after plotting the human interactions as a network. The statistical study displayed that the engendered interactome was robust as 95% proteins are associated and lacked self-interaction. Therefore, the proteins were epitomized as nodes and interactions as edges.

The integration of the contemporary human PPI networks are favorable as they share complementary information. Nevertheless, the assimilation of information from assorted reserves is a tedious task because the data was fundamentally engendered by employing innumerable investigational circumstances by applying diverse identifiers. Also, this information or data is stored normally in dissimilar formats. Consequently, it was obligatory to carry out a vigilant investigation of existing challenges present in the human PPI data and also the steps employed for the

efficacious integration. To overcome these defies, a database was designed and implemented for integrating human PPI networks from various resources. This assimilated framework was designated as Human Protein-Protein Interactions (HPPIs). HPPIs in 76072 interactions and more than 10078 unique proteins were collected from twelve major PPI sources. HPPIs offers numerous dimensions such as GO, partner, annotation, and pathway for the quality valuation of all the interacting pairs.

Connected component method was employed to investigate the human interactome network for removing all the distracted interactions. The outcome was an extremely connected human protein network in the human interactome. The highly connected human protein network strongly adhered to all the network property as human interactome. The reconstructed CCA network owned 9886 proteins and its 58568 interactions. The integrated human cancer protein database was then investigated in the highly connected human protein network to retrieve cancer and non-cancer interactions.

Cancer proteins were collected from the exceedingly legitimate and curated database explicitly CBio and Sanger to map cancer proteins in HPPIs. As demarcated earlier, the unique 3146 cancer protein dataset obtained was entitled as Cancer Proteins (CP). The CP dataset was mapped against the HPPIs dataset to reconstruct cancer interactions in human interactome. This result infers that stronger connectivity. 100% shortest path is included in the largest network. The average shortest path between two nodes in the CCA network is 4.054 which is same as the HPPIs.

Molecular complex detection method (MCODE) identified highly connected modules in the CCA network. MCODE algorithm discovered highly connected biological modules from huge protein network. High-

scoring clusters have a high density value. MCODE derived totally 121 highly connected modules from the CCA network. The further analysis considered only 99 MCODE clusters as the density score value is >2. The selected complexes based on the MCODE density value was then subjected to the validation process.

The statistical investigation revealed 94.34% exactitude. Among them, only four clusters were plotted which was lower than the cut off value. All these four clusters share cancerous and non-cancerous interaction groups. The statistical analysis also estimated 94 modules to be biologically relevant. The clusters were categorized as cancerous protein cluster and non-cancerous protein cluster from a total of 6232462 clusters respectively.

The studies conducted with the help of Gene ontology revealed that most of the cancerous and non-cancerous proteins from the cluster proteins are positioned within the nucleus component. One of the important biological process of the cluster protein is the metabolism of nucleic acid. The molecular function is a major aspect of the GO analysis and it demonstrated that this major function of complex protein is pertaining to the areas like DNA binding and RNA binding. It was found after the molecular function analysis that 101 proteins from the 267 DNA binding proteins were cancer proteins. 89 proteins from the 175 RNA binding proteins were cancer proteins. This result emphasizes on the significance of GO analysis information in drug discovery (Müller and Brown, 2012).

When the four topological measurements are considered, it is evident that the cancer proteins possess unfavorably diverse topological properties compared to the non-cancerous proteins. Cancer proteins incline to possess shorter shortest-path distance, greater connectivity and betweenness, and frailer clustering coefficient compared to the non-

cancerous proteins. It was also observed that the recessive cancerous proteins possessed robust network features compared to the dominant cancerous proteins. These analysis also supported the fact that the global topological properties of cancer and non-cancerous proteins in a human interactome were similar to the highly compactly packed protein complex.

The nodes with considerably higher number of acquaintances in the network are entitled as hub nodes. In a whole network, these hub nodes are vital in the information flow exchange. 250 hub cancer proteins and 138 normal proteins were identified respectively by employing degree distribution, which in turn confirmed the amelioration of hub proteins in a cluster.

The pathway level information was also integrated in this study. Since, proteins are unique in cluster, the information on linking cluster is crucial. One of the major linking property is the pathway analysis. From the comparative analysis, it was reported that the highest number of cluster proteins are present in prostate cancer. The cluster proteins are not connected with the cancer proteins physically but they are extremely connected with the cancer linking pathway.

The phenotype classification using the KEGG (BRITE) revealed that perilous diseases like breast cancer, ovarian cancer and prostate cancer displayed the highest number of proteins accumulated from cluster.

CancerNet is a web-based tool, which encompasses the entire data about the human interactome. CancerNet is comparatively a user friendly web-interface for the scholars because it provides a single platform to extract data about the human interactome. In uniprot database, the user switch between windows to access the data but CancerNet furnishes all the information in one window.

Chapter 7

CONCLUSION

Many researchers were able to initially identify the potential of network biology in many research areas. The advancement and innovation in experimental modus operandi and computational approaches will endure to augment the exposure and sensitivity of PPI networks. A focus on interactomics exclusively in its application to research on cancer will be on the assortment of dissimilar categories of networks. This includes transcriptional monitoring, PPI and metabolic networks to empower the conception of detailed molecular replicas of dangerous diseases such as neurodegenerative ailments and oncogenesis. Additionally, the assimilation of interactions networks with the magnificent sets of data engendered by persistent disease-related sequencing, microarray, or imaging tasks are expected to deliver us with molecular records of exceptional facts for the human related to health and sickness. Consequently, network biology assures considerable contribution to an enhanced insight of the intricacy of disease and in due course to its antidote.

The following are the major findings in this study:

- Human interactome created from 79,950 human protein interaction using 4 databases: NCBI reference sequence, Human atlas, Plasma proteome database and Uniprot.
- 3146 cancer protein interactions identified from human interactome.
- Connected component and molecular detection method finds 99 major protein complexes in human interactome.

- Topological property differences exist between cancer and normal proteins.
- CancerNet tool developed for protein and its related information.

In this investigation, the approach is based on the multi-step interaction network sifting phases, in which PPI data was assimilated with expression information from healthy and diseased person by employing numerous phases. This methodology efficiently predicted several well-known and unique gene disease transformers. This technique is very scalable, and also has the advantage of extending the investigation to other human maladies, as long as the required information is present. On the other hand, this integration necessitates manual gathering of information and stacks of programming efforts for the execution of such type of investigation. Consequently, the imminent target of CancerNet would be to afford such a facility within the CancerNet framework, to systematize the process of discovering the disease modifiers and scrutinizing them in combination with disease-relevant biological information. Specifically, a user can upload their own expression data to screen the examined network for a specific disorder or disease, and this screened network can then be assimilated further with other biological pathways or functional aspects.

CancerNet is an extremely user-friendly tool. Lot of validations were done and it was confirmed that this tool is a one stop solution to investigate normal and cancer protein interactions. GO analysis displays most of the proteins from CANC complex existent in the nucleus. Drugs designed against nucleus protein are the very vital. For example, Dengue and HIV virus are the main proteins, which attack the nucleus and results in lethal consequences for human body. CancerNet is an accolade to the researchers working on cancer investigation to scrutinize definite cancer stimulated

protein networks, networking properties supplementing drug discovery, biological function of diverse proteins, interacting cohorts of specific proteins and its networks, etc. The characteristic features like assortativity, clustering co-efficient, degree distribution, betweenness of centrality, pathway analysis, and shortest path length resulted in the ratification of presence of hub CANC complex in human bodies. This recommended system can also be applied for other entities of biological concern.

The study of protein interaction is a major milestone and can be extended to modern technology such as Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR). The result of CRISPR completely depends on the interaction of Cas9 and Guide RNA or gRNA. This study can help us to identify similar interacting complexes like Cas9.

REFERENCES

Adams, J. J., Pal, G., Jia, Z., & Smith, S. P. (2006). Mechanism of bacterial cell-surface attachment revealed by the structure of cellulosomal type II cohesin-dockerin complex. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(2), 305-310.

Adamson, E. D. (1987). Oncogenes in development. *Development*, *99*(4), 449-471.

Aittokallio, T., & Schwikowski, B. (2006). Graph-based methods for analysing networks in cell biology. *Briefings in bioinformatics*, *7*(3), 243-255.

Albert, R., Jeong, H., & Barabási, A. L. (1999). Internet: Diameter of the world-wide web. *Nature*, *401*(6749), 130-131.

Altaf-Ul-Amin, M., Shinbo, Y., Mihara, K., Kurokawa, K., & Kanaya, S. (2006). Development and implementation of an algorithm for detection of protein complexes in large interaction networks. *BMC bioinformatics*, *7*(1), 207.

Arnau, V., Mars, S., & Marín, I. (2005). Iterative cluster analysis of protein interaction data. *Bioinformatics*, *21*(3), 364-378.

Attri, A. K., & Minton, A. P. (2005). Composition gradient static light scattering: a new technique for rapid detection and quantitative characterization of reversible macromolecular hetero-associations in solution. *Analytical biochemistry*, *346*(1), 132-138.

Axelrod, D. (2001). Total internal reflection fluorescence microscopy in cell biology. Traffic, 2(11), 764-774.

Axelrod, D. (2003). [1] Total internal reflection fluorescence microscopy in cell biology. *Methods in enzymology*, *361*, 1-33.

Bader, G. D., & Hogue, C. W. (2002). Analyzing yeast protein–protein interaction data obtained from different sources. *Nature biotechnology*, *20*(10), 991-997.

Bader, G. D., & Hogue, C. W. (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC bioinformatics*, *4*(1), 1.

Bader, G. D., Donaldson, I., Wolting, C., Ouellette, B. F., Pawson, T., & Hogue, C. W. (2001). BIND—the biomolecular interaction network database. *Nucleic acids research*, *29*(1), 242-245.

Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*(5439), 509-512.

Barabasi, A. L., & Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nature reviews genetics*, *5*(2), 101-113.

Barbe, L., Lundberg, E., Oksvold, P., Stenius, A., Lewin, E., Björling, E., & Andersson-Svahn, H. (2008). Toward a confocal subcellular atlas of the human proteome. *Molecular & Cellular Proteomics*, 7(3), 499-508.

Batada, N. N., Hurst, L. D., & Tyers, M. (2006). Evolutionary and physiological importance of hub proteins. *PLoS Comput Biol*, 2(7), e88.

Becker, E., Robisson, B., Chapple, C. E., Guénoche, A., & Brun, C. (2012). Multifunctional proteins revealed by overlapping clustering in protein interaction network. *Bioinformatics*, *28*(1), 84-90.

Berman, H., Henrick, K., Nakamura, H., & Markley, J. L. (2007). The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic acids research*, *35*(suppl 1), D301-D303.

Braun, P., Tasan, M., Dreze, M., Barrios-Rodiles, M., Lemmens, I., Yu, H., & Venkatesan, K. (2009). An experimentally derived confidence score for binary protein-protein interactions. *Nature methods*, *6*(1), 91-97.

Brohee, S., & Van Helden, J. (2006). Evaluation of clustering algorithms for protein-protein interaction networks. *BMC bioinformatics*, *7*(1), 1.

Bry, F., & Kröger, P. (2003). A computational biology database digest.*Distributed & Parallel Databases*, *13*(1), 7-42.

Bry, F., & Kröger, P. (2003). A computational biology database digest.*Distributed & Parallel Databases*, *13*(1), 7-42.

Charbonnier, S., Gallego, O., & Gavin, A. C. (2008). The social network of a cell: recent advances in interactome mapping. *Biotechnology annual review*, *14*, 1-28.

Charbonnier, S., Gallego, O., & Gavin, A. C. (2008). The social network of a cell: recent advances in interactome mapping. *Biotechnology annual review,14*, 1-28.

Chen, S. C., Zhao, T., Gordon, G. J., & Murphy, R. F. (2007). Automated image analysis of protein localization in budding yeast. *Bioinformatics*,23(13), i66-i71.

Cho, Y. R., Hwang, W., & Zhang, A. (2007, October). Optimizing flow-based modularization by iterative centroid search in protein interaction networks.

In*2007 IEEE 7th International Symposium on BioInformatics and BioEngineering* (pp. 342-349). IEEE.

Chu, B. (1974). Laser light scattering. Academic Press., New York, N.Y

Creixell, P., Schoof, E. M., Erler, J. T., & Linding, R. (2012). Navigating cancer network attractors for tumor-specific therapy. *Nature biotechnology*,*30*(9), 842-848.

Daigle, S. R., Olhava, E. J., Therkelsen, C. A., Majer, C. R., Sneeringer, C. J., Song, J., ... & Jin, L. (2011). *Selective killing of mixed lineage leukemia cells by a potent small-molecule DOT1L inhibitor.* Cancer cell, 20(1), 53-65.

Davies, H., Hunter, C., Smith, R., Stephens, P., Greenman, C., Bignell, G., ... & Parker, A. (2005). Somatic mutations of the protein kinase gene family in human lung cancer. *Cancer research*, *65*(17), 7591-7595.

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

Dunn, R., Dudbridge, F., & Sanderson, C. M. (2005). The use of edge-betweenness clustering to investigate biological function in protein interaction networks. *BMC bioinformatics*, *6*(1), 1.

Eisenberg, D., Marcotte, E. M., Xenarios, I., & Yeates, T. O. (2000). *Protein function in the post-genomic era*. Nature, 405(6788), 823-826.

Eisenberg, E., & Levanon, E. Y. (2003). Human housekeeping genes are compact. *Trends in Genetics*, *19*(7), 362-365.

Ewing, R. M., Chu, P., Elisma, F., Li, H., Taylor, P., Climie, S., ... & Taylor, R. (2007). Large-scale mapping of human protein–protein interactions by mass spectrometry. *Molecular systems biology*, *3*(1), 89.

Fields, S., & Song, O. K. (1989). A novel genetic system to detect protein protein interactions.

Franke, L., Van Bakel, H., Fokkens, L., De Jong, E. D., Egmont-Petersen, M., & Wijmenga, C. (2006). Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *The American Journal of Human Genetics*, *78*(6), 1011-1025.

Garraway, L. A., & Lander, E. S. (2013). Lessons from the cancer genome. *Cell*, 153(1), 17-37.

Gavin, A. C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., ... & Edelmann, A. (2006). Proteome survey reveals modularity of the yeast cell machinery. *Nature*, *440*(7084), 631-636.

Gavin, A. C., Bösche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., ... & Remor, M. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, *415*(6868), 141-147.

Gingras, A. A., White, P. J., Chouinard, P. Y., Julien, P., Davis, T. A., Dombrowski, L., & Marette, A. (2007). Long-chain omega-3 fatty acids regulate bovine whole-body protein metabolism by promoting muscle insulin signalling to the Akt–mTOR–S6K1 pathway and insulin sensitivity. *The Journal of physiology*, *579*(1), 269-284.

Giot, L., Bader, J. S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., & Vijayadamodar, G. (2003). A protein interaction map of Drosophila melanogaster. *Science*, 302(5651), 1727-1736.

Glory, E., & Murphy, R. F. (2007). Automated subcellular location determination and high-throughput microscopy. *Developmental cell*, 12(1), 7-16.

Goh, K. I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., & Barabási, A. L. (2007). The human disease network. *Proceedings of the National Academy of Sciences*, *104*(21), 8685-8690.

Guerrero, A. V., Quang, P., Dekker, N., Jordan, R. C., & Schmidt, B. L. (2008). Peripheral cannabinoids attenuate carcinoma-induced nociception in mice. *Neuroscience letters*, *433*(2), 77-81.

Gunsalus, K. C., Ge, H., Schetter, A. J., Goldberg, D. S., Han, J. D. J., Hao, T., ... & Li, N. (2005). Predictive models of molecular machines involved in Caenorhabditis elegans early embryogenesis. *Nature*, *436*(7052), 861-865.

Han, J. D. J., Bertin, N., Hao, T., Goldberg, D. S., Berriz, G. F., Zhang, L. V., ... & Vidal, M. (2004). Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature*, *430*(6995), 88-93.

Hanahan, D., & Weinberg, R. A. Hallmarks of cancer: the next generation Cell 144, 646–674 (2011). *CAS ISI PubMed Article*.

Hart, R. G., Pearce, L. A., & Aguilar, M. I. (2007). Meta-analysis: antithrombotic therapy to prevent stroke in patients who have nonvalvular atrial fibrillation. *Annals of internal medicine*, *146*(12), 857-867.

Hartwell, L. H., Hopfield, J. J., Leibler, S., & Murray, A. W. (1999). From molecular to modular cell biology. *Nature* 402. *C47–C52*.

Hennessy, B. T., Smith, D. L., Ram, P. T., Lu, Y., & Mills, G. B. (2005). Exploiting the PI3K/AKT pathway for cancer drug discovery. *Nature* reviews Drug discovery, 4(12), 988-1004.

Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., ... & Margalit, H. (2004). IntAct: an open source molecular interaction database. *Nucleic acids research*, 32(suppl 1), D452-D455.

Hishigaki, H., Nakai, K., Ono, T., Tanigami, A., & Takagi, T. (2001). Assessment of prediction accuracy of protein function from protein–protein interaction data. *Yeast*, 18(6), 523-531.

Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., Moore, L., Adams, S. L., ... & Yang, L. (2002). Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature*, *415*(6868), 180-183.

Hughes, J. P., Rees, S., Kalindjian, S. B., & Philpott, K. L. (2011). Principles of early drug discovery. *British journal of pharmacology*, *162*(6), 1239-1249.

Hughes, J. P., Rees, S., Kalindjian, S. B., & Philpott, K. L. (2011). Principles of early drug discovery. *British journal of pharmacology*, 162(6), 1239-1249.

Ideker, T., & Sharan, R. (2008). Protein networks in disease. *Genome research*, *18*(4), 644-652.

Imielinski, M., Berger, A. H., Hammerman, P. S., Hernandez, B., Pugh, T. J., Hodis, E., ... & Sougnez, C. (2012). Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell*, 150(6), 1107-1120.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., & Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome.*Proceedings of the National Academy of Sciences*, *98*(8), 4569-4574.

Jameson, D. M., Croney, J. C., & Moens, P. D. (2003). [1] Fluorescence: Basic concepts, practical aspects, and some anecdotes. Methods in enzymology, 360, 1-43.

Jeong, H., Mason, S. P., Barabási, A. L., & Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature*, 411(6833), 41-42.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., & Barabási, A. L. (2000). The large-scale organization of metabolic networks. *Nature*, *407*(6804), 651-654.

Jonsson, P. F., & Bates, P. A. (2006). Global topological features of cancer proteins in the human interactome. *Bioinformatics*, *22*(18), 2291-2297.

Kameyama, K., & Minton, A. P. (2006). Rapid quantitative characterization of protein interactions by composition gradient static light scattering. *Biophysical journal*, *90*(6), 2164-2169.

Kar, G., Gursoy, A., & Keskin, O. (2009). Human cancer protein-protein interaction network: a structural perspective. *PLoS Comput Biol*, 5(12), e1000601.

Kelly, T. K., De Carvalho, D. D., & Jones, P. A. (2010). Epigenetic modifications as therapeutic targets. *Nature biotechnology*, 28(10), 1069-1078.

Kim, D. H., Noh, J. D., & Jeong, H. (2004). Scale-free trees: The skeletons of complex networks. *Physical Review E*, *70*(4), 046126.

Kim, P. M., Lu, L. J., Xia, Y., & Gerstein, M. B. (2006). Relating three-dimensional structures to protein networks provides evolutionary insights. *Science*, *314*(5807), 1938-1941.

King, A. D., Pržulj, N., & Jurisica, I. (2012). Protein complex prediction with RNSC. *Bacterial Molecular Networks: Methods and Protocols*, 297-312.

Krause, R., & Wild, D. L. (2006). Identifying protein complexes in high-throughput protein interaction screens using an infinite latent feature model. In *Pacific Symposium on Biocomputing* (Vol. 11, pp. 231-242). World Scientific.

KRAUSE, R., & WILD, D. L. (2006). Identifying protein complexes in high-throughput protein interaction screens using an infinite latent feature model. In *Pacific Symposium on Biocomputing* (Vol. 11, pp. 231-242). World Scientific.

Lage, K., Karlberg, E. O., Størling, Z. M., Olason, P. I., Pedersen, A. G., Rigina, O., ... & Moreau, Y. (2007). A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nature biotechnology*,*25*(3), 309-316.

Lambert, J. P., Ivosev, G., Couzens, A. L., Larsen, B., Taipale, M., Lin, Z. Y., & Pawson, T. (2013). Mapping differential interactomes by affinity

purification coupled with data-independent mass spectrometry acquisition.*Nature methods*, *10*(12), 1239-1245.

Lee, S. A., Chan, C. H., Tsai, C. H., Lai, J. M., Wang, F. S., Kao, C. Y., & Huang, C. Y. F. (2008). Ortholog-based protein-protein interaction prediction and its application to inter-species interactions. *BMC bioinformatics*, 9(12), 1.

Lehner, B., & Fraser, A. G. (2004). A first-draft human protein-interaction map. *Genome biology*, *5*(9), 1.

Lewis, E. N., Qi, W., Kidder, L. H., Amin, S., Kenyon, S. M., & Blake, S. (2014). Combined dynamic light scattering and Raman spectroscopy approach for characterizing the aggregation of therapeutic proteins.*Molecules*, *19*(12), 20888-20905.

LI, H. (2010). Network topology in human protein interaction data predicts functional association.

Li, S., Armstrong, C. M., Bertin, N., Ge, H., Milstein, S., Boxem, M., & Goldberg, D. S. (2004). A map of the interactome network of the metazoan C. elegans. *Science*, *303*(5657), 540-543.

Lin, C., Cho, Y. R., Hwang, W. C., Pei, P., & Zhang, A. (2007). Clustering methods in protein-protein interaction network. *Knowledge Discovery in Bioinformatics: techniques, methods and application*, 1-35.

Linding, R., Jensen, L. J., Ostheimer, G. J., van Vugt, M. A., Jørgensen, C., Miron, I. M., ... & Metalnikov, P. (2007). Systematic discovery of in vivo phosphorylation networks. *Cell*, *129*(7), 1415-1426.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., ... & Cho, J. H. (2009). Finding the missing heritability of complex diseases. *Nature*, *461*(7265), 747-753.

Marcotte, E. M., Pellegrini, M., Ng, H. L., Rice, D. W., Yeates, T. O., & Eisenberg, D. (1999). Detecting protein function and protein-protein interactions from genome sequences. *Science*, *285*(5428), 751-753.

Maslov, S., & Sneppen, K. (2002). Specificity and stability in topology of protein networks. *Science*, *296*(5569), 910-913.

Masseroli, M., Galati, O., Manzotti, M., Gibert, K., & Pinciroli, F. (2005). Inherited disorder phenotypes: controlled annotation and statistical analysis for knowledge mining from gene lists. *BMC bioinformatics*, *6*(Suppl 4), S18.

Mayer, G. (2009). Data management in systems biology I-Overview and bibliography. *arXiv preprint arXiv:0908.0411*.

McLendon, R., Friedman, A., Bigner, D., Van Meir, E. G., Brat, D. J., Mastrogianakis, G. M., ... & Yung, W. A. (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216), 1061-1068.

Memišević, V., & Pržulj, N. (2012). C-GRAAL: Common-neighbors-based global GRA ph AL ignment of biological networks. *Integrative Biology*, *4*(7), 734-743.

Memišević, V., Wallqvist, A., & Reifman, J. (2013). Reconstituting protein interaction networks using parameter-dependent domain-domain interactions. *BMC bioinformatics*, 14(1), 1.

Meyerkord, C. L., & Fu, H. (2015) Protein-Protein Interactions. Methods and Applications, 1st Edn, eds, *New York: Humana Press*.

Milenkovic, T., & Przulj, N. (2008). Uncovering biological network function via graphlet degree signatures. *arXiv preprint arXiv:0802.0556*.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., & Alon, U. (2002). Network motifs: simple building blocks of complex networks.*Science*, *298*(5594), 824-827.

Newberg, J. Y., Li, J., Rao, A., Pontén, F., Uhlén, M., Lundberg, E., & Murphy, R. F. (2009, June). Automated analysis of human protein atlas immunofluorescence images. In *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro* (pp. 1023-1026). IEEE.

Newman, M. E. (2003). The structure and function of complex networks. *SIAM review*, *45*(2), 167-256.

Ofran, Y., & Rost, B. (2003). Analysing six types of protein–protein interfaces. *Journal of molecular biology*, 325(2), 377-387.

Palla, G., Derényi, I., Farkas, I., & Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society.*Nature*, *435*(7043), 814-818.

Peri, S., Navarro, J. D., Amanchy, R., Kristiansen, T. Z., Jonnalagadda, C. K., Surendranath, V., & Ibarrola, N. (2003). Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome research*, 13(10), 2363-2371.

Persico, M., Ceol, A., Gavrila, C., Hoffmann, R., Florio, A., & Cesareni, G. (2005). HomoMINT: an inferred human network based on orthology mapping of protein interactions discovered in model organisms. *BMC bioinformatics*,*6*(4),1.

Phizicky, E. M., & Fields, S. (1995). Protein-protein interactions: methods for detection and analysis. *Microbiological reviews*, 59(1), 94-123.

Prasad, T. K., Kandasamy, K., & Pandey, A. (2009). Human Protein Reference Database and Human Proteinpedia as discovery tools for systems biology. *Reverse Chemical Genetics: Methods and Protocols*, 67-79.

Radivojac, P., Baenziger, P. H., Kann, M. G., Mort, M. E., Hahn, M. W., & Mooney, S. D. (2008). Gain and loss of phosphorylation sites in human cancer. *Bioinformatics*, *24*(16), i241-i247.

Rossin, E. J., Lage, K., Raychaudhuri, S., Xavier, R. J., Tatar, D., Benita, Y., & International Inflammatory Bowel Disease Genetics Constortium. (2011). Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. *PLoS Genet*, *7*(1), e1001273.

Rossin, E. J., Lage, K., Raychaudhuri, S., Xavier, R. J., Tatar, D., Benita, Y., ... & International Inflammatory Bowel Disease Genetics Constortium.

(2011). Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. *PLoS Genet*, *7*(1), e1001273.

Rual, J. F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., ... & Klitgord, N. (2005). Towards a proteome-scale map of the human protein–protein interaction network. *Nature*, *437*(7062), 1173-1178.

Saeed, R., & Deane, C. M. (2006). Protein protein interactions, evolutionary rate, abundance and age. *BMC bioinformatics*, *7*(1), 1.

Said, M. R., Begley, T. J., Oppenheim, A. V., Lauffenburger, D. A., & Samson, L. D. (2004). Global network analysis of phenotypic effects: protein networks and toxicity modulation in Saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences*, *101*(52), 18006-18011.

Sam, L., Liu, Y., Li, J., Friedman, C., & Lussier, Y. A. (2007). Discovery of protein interaction networks shared by diseases. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* (p. 76). NIH Public Access.

Samanta, M. P., & Liang, S. (2003). Predicting protein functions from redundancies in large-scale protein interaction networks. *Proceedings of the National Academy of Sciences*, *100*(22), 12579-12583.

Schwikowski, B., Uetz, P., & Fields, S. (2000). A network of protein–protein interactions in yeast. *Nature biotechnology*, *18*(12), 1257-1261.

Sevimoglu, T., & Arga, K. Y. (2014). The role of protein interaction networks in systems biomedicine. *Computational and structural biotechnology journal*,11(18), 22-27.

Sharan, R., Ulitsky, I., & Shamir, R. (2007). Network-based prediction of protein function. *Molecular systems biology*, *3*(1), 88.

Shen-Orr, S. S., Milo, R., Mangan, S., & Alon, U. (2002). Network motifs in the transcriptional regulation network of Escherichia coli. *Nature genetics*,*31*(1), 64-68.

Some, D., & Kenrick, S. (2012). Characterization of protein-protein interactions via static and dynamic light scattering. INTECH Open Access Publisher.

Spirin, V., & Mirny, L. A. (2003). Protein complexes and functional modules in molecular networks. *Proceedings of the National Academy of Sciences*,*100*(21), 12123-12128.

Stelzl, U., & Wanker, E. E. (2006). The value of high quality protein–protein interaction networks for systems biology. *Current opinion in chemical biology*, *10*(6), 551-558.

Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F. H., Goehler, H., ... & Timm, J. (2005). A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, *122*(6), 957-968

Tarassov, K., Messier, V., Landry, C. R., Radinovic, S., Molina, M. M. S., Shames, I., ... & Michnick, S. W. (2008). An in vivo map of the yeast protein interactome. *Science*, 320(5882), 1465-1470.

Titz, B., Schlesner, M., & Uetz, P. (2004). What do we learn from high-throughput protein interaction data?. *Expert review of proteomics*, 1(1), 111-121.

Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., & Qureshi-Emili, A. (2000). A comprehensive analysis of protein–protein interactions in Saccharomyces cerevisiae. *Nature*, *403*(6770), 623-627.

Uhlén, M., Björling, E., Agaton, C., Szigyarto, C. A. K., Amini, B., Andersen, E., ... & Berglund, L. (2005). A human protein atlas for normal and cancer tissues based on antibody proteomics. Molecular & Cellular Proteomics,4(12), 1920-1932.

Van Driel, M. A., Cuelenaere, K., Kemmeren, P. P. C. W., Leunissen, J. A., Brunner, H. G., & Vriend, G. (2005). GeneSeeker: extraction and integration of human disease-related information from web-based genetic databases.*Nucleic acids research*, *33*(suppl 2), W758-W761.

Vidal, M., Cusick, M. E., & Barabasi, A. L. (2011). Interactome networks and human disease. *Cell*, *144*(6), 986-998.

Virtaneva, K., Wright, F. A., Tanner, S. M., Yuan, B., Lemon, W. J., Caligiuri, M. A., ... & Krahe, R. (2001). Expression profiling reveals fundamental biological differences in acute myeloid leukemia with isolated trisomy 8 and normal cytogenetics. *Proceedings of the National Academy of Sciences*,*98*(3), 1124-1129.

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., & Kinzler, K. W. (2013). Cancer genome landscapes. *science*, 339(6127), 1546-1558.

Von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S., & Bork, P. (2002). Comparative assessment of large-scale data sets of protein–protein interactions. *Nature*, *417*(6887), 399-403.

Wang, F., Wang, X., Yuan, C. G., & Ma, J. (2010). Generating a prion with bacterially expressed recombinant prion protein. *Science, 327*(5969), 1132-1135.

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

Weinstein, I. B., & Joe, A. K. (2006). Mechanisms of disease: oncogene addiction—a rationale for molecular targeting in cancer therapy. *Nature clinical practice Oncology, 3*(8), 448-457.

Wells, J. A., & McClendon, C. L. (2007). Reaching for high-hanging fruit in drug discovery at protein–protein interfaces. *Nature*, 450(7172), 1001-1009.

Wuchty, S. (2004). Evolution and topology in the yeast protein interaction network. *Genome research*, *14*(7), 1310-1314.

Wuchty, S., & Almaas, E. (2005). Peeling the yeast protein network. *Proteomics*, *5*(2), 444-449.

Xing, E. P., Wu, W., Jordan, M. I., & Karp, R. M. (2004). Logos: a modular bayesian model for de novo motif detection. *Journal of Bioinformatics and Computational Biology*, *2*(01), 127-154.

Yeger-Lotem, E., Sattath, S., Kashtan, N., Itzkovitz, S., Milo, R., Pinter, R. Y., ... & Margalit, H. (2004). Network motifs in integrated cellular networks of transcription–regulation and protein–protein interaction. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(16), 5934-5939.

Yıldırım, M. A., Goh, K. I., Cusick, M. E., Barabási, A. L., & Vidal, M. (2007). Drug—target network. *Nature biotechnology*, *25*(10), 1119-1126.

Zhang, A. (2009). Protein interaction networks: computational analysis. *Cambridge University Press*.

Zhang, W., Sun, F., & Jiang, R. (2011). Integrating multiple protein-protein interaction networks to prioritize disease genes: a Bayesian regression approach. *BMC bioinformatics*, *12*(1), 1.

Zhu, H., Bilgin, M., Bangham, R., Hall, D., Casamayor, A., Bertone, P., ... & Mitchell, T. (2001). Global analysis of protein activities using proteome chips. *science*, 293(5537), 2101-2105.

APPENDIX I

Design and Implementation

As the amount of data on protein interactions is growing rapidly, there is ongoing demand for integrated platforms with a high degree of flexibility. Such platforms should not be only easily accessible but also be consistently updated. Data should be accurately integrated from different sources and queries should be processed in minimal time. The structure of the platform should be extensible to new data without changing its data structure. Thus, a careful design and implementation of the system and the selection of computational approaches to assemble heterogeneous data sources are crucial. Traditional computational approaches like object-oriented software and relational databases can be cumbersome and time-consuming. Typically, persisting data objects from SQL tables with a Hypertext Preprocessor or Personal Home Page (PHP) connection and prepared SQL statements may be easy for simple objects.

Integration of data

Data integration from multiple resources results in many concerns. As many databases are involved, the retrieved data had different structures and patterns. The main challenge was to align them to a common pattern and assign a unique identifier. When this is executed manually, it becomes cumbersome task because of the huge data. It is impossible to change the structure of the data from all the database at a time. So, Perl programs were employed to automate this task.

The other concern was the presence of lot of duplicated data. This also was resolved by the use of automated Perl programs. Similarly, the

self-interacting protein information was fixed by using automated Perl programs.

Transfer of database from local to global server

The total size of the database used for CancerNet tool is 2 GB. Usually, phpmyadmin is used for installation and update of database. But, the allowed capacity to upload is 20 MB. So, the inbuilt configurations were customized to resolve this issue.

Before converting SQL files and uploading in the MySQL database, the spreadsheet were first converted to .csv and then to SQL file which was tedious task owing to the huge number of files.

APPENDIX II

Tool related challenges

Generally, list of pathways are not accessible in any other website other than the KEGG website. Only the accession number is displayed and when the accession number is clicked, it directs to the particular website. But, using PHP graph module was used to rectify this problem. It is possible to parse the variable to KEGG web server and it aids in fetching all the pathways in the CancerNet webpage avoiding visit to other websites. The same procedure is used to display the KEGG pathway in the graphical display.

As mentioned in the section 5.5, out of the 16 information the data related to KEGG is retrieved real-time and is not stored on the database.

Figure 1: Homepage
This is the homepage of the CancerNet tool.

Figure 2: Information tabs

This image displays the various information furnished by the CancerNet tool.

Related links

The link to the homepage of CancerNet is http://bioteccancernet.cusat.ac.in/. The username is **admin** and the password is **admin123**.

The following is the list of related links:

Data	**Link**
Human interactome	**http://bioteccancernet.cusat.ac.in/hin**
CCA network	http://bioteccancernet.cusat.ac.in/cca
MCODE	http://bioteccancernet.cusat.ac.in/mcode
Hub proteins	http://bioteccancernet.cusat.ac.in/hub
Cancer proteins	http://bioteccancernet.cusat.ac.in/cancer
Cancer phenotype	http://bioteccancernet.cusat.ac.in/phenotype
Pathways	http://bioteccancernet.cusat.ac.in/pathways

LIST OF PUBLICATIONS

Peer Reviewed Publications

1. **Arinnia Anto** and Padma Nambisan (2014), Using multi level algorithmic methods for identifying highly interacting human protein complexes and various protein pathways, International Journal of Computational Bioinformatics and In Silico Modeling, Vol. 3, No. 4 (2014): 433-439
2. **Nayana Parambayil**, Aiswarya Chenthamarakshan, Arinnia Anto, Sudha Hariharan and Padma Nambisan (2014), Computational Studies on Lip H Isolated From Ganoderma Lucidum Gd88, Directory of open access journals, Belgrade, 67(3), 817-828, 2015