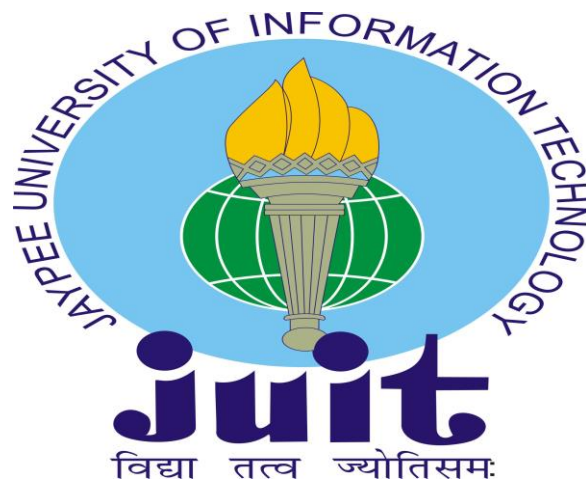


# **Annotation of Biological Systems using Top down approach of Systems Biology**

Enrollment No - 122503  
Name - Rajinder Gupta  
Supervisor - Dr. Tiratha Raj Singh



May - 2014

Submitted in partial fulfillment of the Degree of

**Master of Technology**

in

**Computational Biology**

DEPARTMENT OF BIOTECHNOLOGY AND BIOINFORMATICS

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY,

WAKNAGHAT, SOLAN - 173234, H.P., INDIA

## Index

Chapter Number	Topics	Page Number
	Certificate from the Supervisor	II
	Acknowledgement	III
	Summary	IV
	List of Figures	V
	List of Tables	VI
	List of Symbols and Acronyms	VII
Chapter-1	Introduction	1
Chapter-2	Review of Literature	6
Chapter-3	Methodology	10
Chapter-4	Results	19
Chapter-5	Annotations	26
Chapter-6	Inferences	32
Appendix A	Scripts for data generation	34
Appendix B	Database creation commands on MySQL	48
Appendix C	Data transformation commands on Pentaho Kettle	52
Appendix D	MySQL commands for search option at GUI	66
	Bibliographic references	68
	List of Posters	71
	Resume	72

## CERTIFICATE

This is to certify that the work titled “**Annotation of Biological Networks using Top down approach of Systems Biology**” submitted by “**Mr. Rajinder Gupta**” in partial fulfillment for the award of degree of M.Tech. Computational Biology of Jaypee University of Information Technology, Waknaghat has been carried out under my supervision. This work has not been submitted partially or wholly to any other University or Institute for the award of this or any other degree or diploma.

Signature of Supervisor	.....
Name of Supervisor	Dr. Tiratha Raj Singh
Designation	Assistant Professor
Date	23 /05/2014

## Acknowledgement

I would like to express my profound gratitude to my guide Dr.Tiratha Raj Singh for his guidance, support and constant encouragement throughout the course of this project work. He has been more than just my project guide; at times a mentor to rescue me out of my doubts, at times a philosopher to hear my miseries and still at times a friend with whom I can share myself. He has always pushed me to work hard and still gave me enough space to cultivate and implement my own ways of dealing with the problems. Also he taught me life values which I will never forget. Thank you very 'many' much 'Sir'.

I would like to acknowledge Vice Chancellor Sir and Dean & HOD BT & BI Prof. R. S. Chauhan for providing me with an opportunity to be a part of the institute and to complete my Master's degree.

Here, I want to give a special mention to Dr. DipankarSengupta, who has been a source of immense motivation and inspiration both for my academic and personal life.He was never, and I know will never be, more than a phone call away. With his acquaintance of Omni-domain knowledge and skills he has helped me in almost every aspect I have asked him for. Thanks a lot 'Sir'.

In addition I would like to thank all the faculty members of BT & BI Department of JUIT, who have mended a novice in the field of Computational Biology into a well-equipped Computational Biologist.Adding on to the list, I would like to thank Somlata Mam for providing us with work space.

Also, I would like to thank Ms. ManikaSehgal, Mr. Imran Shah and Mr. Ashwani Kumarwho have always been available to help and to provide me with helpful insights on my work. Also, all other PhD scholars have been a great help in my work.

I would also like to appreciate the part that my classmates have played in shaping up this project work. I would like to acknowledge Ms. Anuja Mishra and Ms. JaaiVadke for their constant boosting and cheering me up at hard times which has helped me a lot. Thank a lot folks!!

A big thanks to all open source tool providers because of which I was able to complete my project.

I would like to thank the Almighty God for his grace throughout my life. And above all my Mom and Dad who have always supported me through the thick& thin and have been a constant source of encouragement and support,and lastly my younger brother who is a friend and support to me.

Rajinder Gupta

## Summary

The current study deals with the whole network rather than an individual biological entity and thus brings forward the emergent properties of the system. The concepts of systems biology are applied to the disease pathway to predict the probable targets out of it. FANMOD is used to generate the network motifs from the pathways. From these images the frequency of the nodes involved in the generated network motifs are calculate and further annotated using three new terms viz.  $TO_G^P$ CNM,  $T_G^P$ CNM &  $TF_G^P$ allNM that are introduced in this study for helping in revealing the dependency of a system on a set of genes or proteins or both. The final list of probable targets for different diseases; as in this study non-small cell lung cancer, small cell lung cancer, pancreatic cancer, prostate cancer and renal cancer; bring forward new information which implies that different nodes for related or unrelated cancers are predicted by the model to be the probable targets. Such a thing will bring to limelight a new wave in drug discovery process which states that a single drug can be effective for more than 1 disease.

## List of Figures

- Figure 1.1 Image representation of a graph (A): A real network or graph; (B): A randomized network or graph
- Figure 3.1 ProTaS work flow
- Figure 3.2 Overview of the pipelined methodology
- Figure 3.3 Data model of the database developed on CA Erwin DM
- Figure 3.4 Home Screen of ProTaS
- Figure 3.5 Search Panel of ProTaS
- Figure 4.1 Network motif images for Non-small cell Lung Cancer.
- Figure 4.2 Network motif images for small cell Lung Cancer.
- Figure 4.3 Network motif images for Renal Cancer.
- Figure 4.4 Search page results; disease name, description and link
- Figure 4.5 Search page results; list of probable nodes, node number in pathway, threshold value and external description link
- Figure 4.6 Search page results; Adjacency matrix and statistical parameters viz. Z-score, P-Value and Significance Profile
- Figure 5.1 List of Significant nodes for Non-small Cell Lung Cancer at threshold > 5%
- Figure 5.2 List of Significant nodes for Renal Cancer at threshold > 7%

## List of Tables

- Table 5.1 Kids & Pencil example to illustrate three terms:  $TO_G^P$ CNM,  $T_G^P$ CNM &  $TF_G^P$ allNM
- Table 5.2 List of significant nodes for 5 diseases taken into study and their respective thresholds

## List of Symbols and Acronyms

1. NM: Network Motifs
2. P-value: It is the number of random networks in which a motif occurred more often than in the original network, divided by the total number of random networks.
3. Significance Profile (SP): It is a vector of Z-scores of a particular set (of same size) of motifs which is normalized to 1.
4.  $TF_G^P$ allNM: Total Frequency of a particular Gene/Protein in all Network Motifs.
5.  $T_G^P$ CNM: Total number of all Genes/Proteins participating in the conserved Network Motif with respect to the corresponding Network Motif.
6.  $TO_G^P$ CNM: Total occurrences of all Genes/Proteins participating in the conserved Network Motif with respect to the corresponding Network Motif.
7. Z-score: It is the original frequency of the network minus the random frequency of the network divided by the standard deviation of the random network.



# Chapter 1

## Introduction

With the evolution of human race so did evolve the diseases we fell hostile to. More complex systems as of ours', translates to more complicated and intricately regulated diseases that attack it. The complexity of our system is not defined by the simple elements that interact in different ways to give a complex system but by the interaction of simple elements (such as ions) and complex elements (such as genes, proteins, regulatory factors etc.) at varied conditions, at varied concentrations and that too in different systems differently. Also a biological system is comprised of large numbers of functionally diverse and frequently multifunctional sets of elements (internally or externally of the system) which interact selectively and nonlinearly to produce coherent and complex behaviors. In spite of such complexities and variations across systems certain patterns are well conserved and regulated.

The complexity of our system calls for diseases with still more complex regulation pathways and to add onto this, more complex are the ways to crack them down. Today, we are surrounded by enormous number of diseases that pose a threat to our society and race; alongside new forms of these diseases keep getting apparent and adding on to it are still the new diseases that are making their mark. Some of the major diseases that have scared the populations across the globe include cancer, diabetes, Alzheimer's, AIDS etc. Some of the new diseases that petrified the people, not too way back, are H1N1 (swine flu), bird flu etc., and the bad thing was how quickly they got evolved to new forms as in case of H1N1 within a couple of months we had the mutated version of the disease namely, H5N1.

There are certain diseases for which till date we do not have the ultimate cure, and numerous people die of such diseases every year. And Cancer is one of it and the most vicious of them all. It accounts for 1 in 8 deaths across the globe, more than AIDS, malaria, and tuberculosis combined [1]. In our body, under normal conditions, only that much number of new cells is produced as much are required by the body to maintain its proper functioning. To achieve such an immaculate precision, certain regulatory, signaling proteins, some external factors etc. are all required. But when there is some abnormality in the cell's genome or under some external factors or both; it gets cancerous and

loses all of its ability to regulate the division of cells and they start proliferating at much higher rates, such an uncontrolled division of cells result into formation of tumors and the malignant tumors are called cancer. Cancers are usually named after the part of the body they infect as if they have infected the pancreas they are named as pancreatic cancer but if the malignant cancer cells from their organ of origin travel to some other body part and infect it, and then they are named by adding a suitable suffix to the name of the cancer, e.g.: Malignant tumors of the blood-forming tissue are designated by the suffix -emia. Thus, leukemia refers to a cancerous proliferation of white blood cells (leukocytes) to other body parts [2].

Today there are a number of cancers that have become an indispensable part of our lifestyles not because of some necessity but because of the living ways that we have adopted. The prime reason of cancer is tobacco smoking either actively or passively, exposure to UV, obesity, unhealthy eating etc. Some of the cancers are common in all human beings; while others have their likings viz. lung & bronchus cancer is more prominent in males whereas breast cancer is in women; Asians are more prone to lung & bronchus cancer whereas Caribbean are more susceptible to prostate cancer; cancer is also age discriminating as the aged people are more vulnerable to cancers [3]. Of all the cancers the deadliest one is the lung cancer which accounts for nearly 27% of all the cancer deaths worldwide; then stands breast cancer, prostate cancer, colon cancer and many more in this order of lethality[4].

Extensive research has been undertaken by different organizations, laboratories, research groups and individuals trying to crack the big nut of finding the perfect cure for the diseases in lesser amount of time and at lower costs. But so far we have just managed to take only baby steps towards our ultimate goal of finding the cure. And when we talk about cancer a lot has been unraveled but none of it has, till date, provided with solid grounds for eventual cure of the disease. One of the things about cancer that has dazzled the research communities for years now is its immaculate control over the cells' life cycle. It not just controls it but also fastens the cell cycle to higher rates resulting in formation of unwanted mass of cells.

The traditional approaches for the study and analysis of the disease pathways were (and still are) focused on a single biological component of the system; which is selected from the literature

survey or personal acquaintance of the researchers. Very less or no care is taken for the components of the system that seem less important in the pathways or are thought of unrelated entities. Many a times we overlook such things if we go by the traditional methods of target search. The need of the hour is to use new and novel approaches to search the probable targets using the holistic approaches rather than the reductionist ones, which we have been using since ages, with less frequently getting the desired outcomes. We really need to come out of the box and try out new and different things.

With the advancements in technology and increase in our domain knowledge newer fields of study have emerged; one of the most dissertated of them is ‘Systems Biology’ [7], of which the domain of network motifs has gained immense admiration. The concepts of systems biology to counteract the complexities of biological systems have not been so far employed to good use. The field of target search has been an area of immense time, labor and intellect consuming but we still find ourselves, on occasions, with ambiguous targets without a factual description for the same. At such a place the knowledge of systems biology can come handy. systems biology is a domain of science where we are more concerned and thus, study the emergent properties of a system and its constituent elements viz. genes, proteins, metabolites, chemical compounds etc. as one whole system rather than the effect of a particular gene or protein or else on the whole system. It can also be put forward as the study of the mechanisms underlying complex biological processes as integrated systems of many interacting components.

In a biological system, the functions and its emergent properties rely on a combination of the networks and the specific elements involved in the network. The system may comprise of interactions at intra or inter cellular, tissues, individuals or species; depending on the requirements of the study. These interactions are nothing but one or the other sort of networks or graphs; some examples of a network from our surroundings include food chains, social interactions, evolutionary relationships, electric circuits, internet, social networking and so on. There is no end to this list; we currently are envired by networks. Networks are so important for turning the wheel of life that they are omnipresent. They make up the basis of the smallest possible dependencies in the biological world; as say, a molecule (one in a billions that help regulating the proper functioning of our system) needs to interact with another molecule, it just cannot go to it and perform its chore but there are defined set of interactions and modifications which one has to follow to fulfill the eventual purpose of its existence;

and all this is achieved by introducing ‘networks’. What networks do here is that they delineate the actual path to be followed for a process to take place correctly and efficiently.

For a biological system, a network can be defined as the set of nodes and edges where a node represents a gene or protein or any other entity from the system (i.e. involved in the system) and the set of edges define the relationships between the corresponding edges. A network motif is a sub-graph of the whole network and can be defined as a reoccurring pattern in the system, which occur at a significantly higher number of times in the real network as compared to its number of occurrences in the randomized network. The fine line of differentiation between a real and a randomized network is just that the edges between the nodes are assigned randomly, keeping the number of edges and the set of nodes constant for the randomized network. This can be explained by the following example:

Consider a network or a graph which is defined by  $G(V, E)$  where  $G$  is the graph,  $V$  is the set of vertices, denoted by  $V\{v_1, v_2, \dots, v_n\}$  and  $E$  is the set of edges between the vertices denoted by  $E\{(v_1, v_3), (v_4, v_{n-3}), (v_7, v_n), (v_n, v_2), \dots\}$ . Now, if we have a real network or graph (Figure 1(A)), where for  $G$  the defining values are given as  $V\{1, 2, 3, 4, 5\}$  and  $E\{(1,3), (2,3), (2,5), (3,4)\}$ . And for such a graph one of the random network or graph (Figure 1(B)) can be  $V\{1, 2, 3, 4, 5\}$  and  $E\{(1,2), (2,5), (3,4), (4,5)\}$  (Note that the set of vertices,  $V$  remains the same and the number of edges in  $E$  remains the same though the values have changed).



Figure 1: Image representation of a graph (A): A real network or graph; (B): A randomized network or graph

In our current study 5 cancers are taken viz. small cell lung cancer, non-small cell lung cancer, pancreatic cancer, prostate cancer and renal cancer. In US, lung cancer (small cell lung cancer & non-small cell lung cancer, combined) stands at top of the list for highest number of new cases and mortality rate; of which the small cell lung cancer is responsible for 20% & non-small cell lung cancer is for, as high as, 80% new cases. Prostate cancer is the third highest on the list for mortality rate after lung and breast cancer in the stats released by [3]. The numbers of cases for pancreatic cancer are also on a higher side. The cancers from totally different sections of our body have been taken to search out for some similarities, if any between the pathways and the set of biological entities taking part in the disease pathway.

Even in India, the scenario is no different. Cancer mortality in India was around 555,000 in 2010 [5]. And furthermore, the absolute number of cancer deaths in India is expected to increase because of population growth, changed lifestyles and increase in life expectancy. What really is required is the better cure and that too at affordable prices.

## Chapter 2

### Review of Literature

Since the days of Norbert Wiener [6], it started getting apparent the usefulness of systems level understanding in biological systems. But it took more than half a century to really implement the concept of systems level understanding in biological networks because of the less developed branches of science that we today witness like molecular biology, genomics, proteomics etc. and the developments in the field of technology, though in the domain of biotechnology, bioinformatics, computer sciences etc., have enabled us to collect comprehensive datasets on system performance and gain information on the underlying molecules.

The term 'systems biology' was coined by Kitano H. in 2002 [7], though the concept was there a couple years afore and since then it has been a revelation. A system-level understanding of a biological system can be derived from insight into four key properties [7]; 1) System structures: These include the network of gene interactions and biochemical pathways, as well as the mechanisms by which such interactions modulate the physical properties of intracellular and multicellular structures. 2) System dynamics: How a system behaves over time under various conditions can be understood through metabolic analysis, sensitivity analysis, dynamic analysis methods such as phase portrait and bifurcation analysis, and by identifying essential mechanisms underlying specific behaviors. Bifurcation analysis traces time-varying change(s) in the state of the system in a multidimensional space where each dimension represents a particular concentration of the biochemical factor involved. 3) The control method: Mechanisms that systematically control the state of the cell can be modulated to minimize malfunctions and provide potential therapeutic targets for treatment of disease. 4) The design method: Strategies to modify and construct biological systems having desired properties can be devised based on definite design principles and simulations, instead of blind trial-and-error.

As soon as people got to know about the foundations and concepts of systems biology, no later came into existence the talk of network motifs.

Today, there are lot many tools available to work with network;these can be classified based on the type of algorithms [8] that lay down their working principle:

### **Network Centric Algorithms:**

These algorithms start with the network and enumerate all sub-graphs of size k that are present in the target network. They have the benefit that sub-graphs that are not present in the target network are never encountered. These algorithms need to make a count of a sub-graph without which they cannot compute the frequency of the sub graph. Such types of algorithms are:

1. **NeMoFINDER (Network Motif Finder):** It was developed in 2006 by Chen *et al.*[9]for undirected & unlabeled PPI networks. It employs into use frequency and uniqueness thresholds, and the F1 frequency concept. The input network is partitioned into smaller sub-networks for counting of lower-level sub-graphs in the pattern growth tree because of which it is less-sensitive to large networks. The concept of ‘Canonical Adjacency Matrix’ is used to resolve isomorphism through a method called ‘Graph Cousins’ [10].
2. **Kavosh:** It was developed by Kashani et al. [11] in 2009. It uses the pattern growth tree (with node-extension) to count all size-k sub-graphs in the input network. It employs the ‘Revolving Door algorithm’ [12] for traversing the pattern growth tree to ensure that every motif is encountered exactly once. The frequency concept F1 is used, along with a frequency threshold. It also uses the NAUTY algorithm [13] for isomorphism testing, which is efficient as the NAUTY does not generate any redundant candidate motif.
3. **MAVisto (Motif Analysis and Visualization Tool):** It was developed in 2005 by Schreiber and Schwöbbermeyer [14]. It uses a pattern growth method called FPF or Flexible Pattern Finder [15] and uses the downward closure property which is used in data mining, as well as a frequency threshold to snip through the branches of the pattern growth trees.
4. **MFinder:**It was developed by Kashtan et al.[16] in 2005 which is basedon an edge-sampling algorithm [17]. The pattern growth tree is used to count all sub-graphs in the network through edge

extension. Frequency concept F1 is used, but motifs are required to be induced sub-graphs. It also uses concentration as the significance metric.

5. **FANMOD**(FAstNetwork **MO**tif**D**etection): It was developed by Wernicke *et al.* [18 - 23] in 2006. It employs a node-sampling strategy and a pattern growth tree using node-extension. It is fast, and can search for node size 3 to 8 in directed and undirected networks. It uses F1 frequency concept, and motifs are required to be induced. It uses concentration as the significance metric and NAUTY [13] for isomorphism checking. FANMOD's randomized enumeration algorithm is known as Rand-ESU.

### Motif Centric Algorithms

These are able to compute frequency of any given sub-graph in the target network and thus enabling them to directly verify whether the query sub-graph is a motif. In order to compute frequencies of all size-k graphs, a motif-centric algorithm first enumerates all possible sub graphs of size-k.

1. **Grochow**: It was developed by Grochow and Kellis [24] in 2007. It is based on the concept of symmetry breaking with mapping to count the frequency of a sub-graph. It utilizes the 'geng' and 'directg' packages given by McKay [25 - 27] to generate all non-isomorphic sub-graphs of size-k; it then checks if any of them is a motif, exhaustively. Though, mapping is an exact strategy, but it is possible to first use sampling to pick a random (large) sub-graph from the network, and then use symmetry breaking and mapping to determine its frequency.
2. **MODA**: It was developed by Omid *et al.* [10] in 2009. It uses Expansion tree, which is a pattern growth tree that implies symmetry breaking to generate unique extensions. It then uses mapping for the first level of the expansion tree. Information from the previous level is then used to efficiently enumerate the extensions for each subsequent level. Because of the combination of efficient strategies, it is fast and can handle relatively large networks and motifs. It also allows arbitrary overlap between sub-graphs, using frequency concept F1.

Though the work on network motifs has been greatly accepted and appreciated but its implementation in the field of medicine is still in its infancy and need to be further explored and



worked on. In case of cancer the combinatorial work (cancer + systems biology; cancer + network motifs) is still lesser and needs to be quickly taken into account. The efficiency of such implementations is going to affect the overall cost and time requirements of the R&D departments.

## Chapter 3

### Methodology

A novel approach has been devised to predict the most significant or center of interest targets viz. genes or proteins or a mix of both, from the diseased pathways which can be further used for in detail analysis of the disease and hence for drug discovery. The soul of our model lies at the fact that Nature never invests its effort and resources on something that is insignificant, and if by any chance any such thing does happen it gets eliminated from the system. The pathways and regulatory networks in the biological systems are well regulated and soundly structured to provide immaculate precision without a miss. The efficiency and adeptness of these networks can be concluded from the fact that, under normal conditions, each part and function is well regulated and efficiently clinched to its fate.

The model comprises of 4 steps, namely:

- Getting the pathway
- Generating Network Motifs
- Annotation of Network Motifs
- Development of ProTaS

The steps are shown in the figure 3.1; starting from getting the pathways from KEGG to the final list of probable targets. Step 1 is getting the pathways from KEGG (discussed in detail in this Chapter), Step 2 is generating the network motifs (discussed in detail in this Chapter), Step 3 is calculation of certain parameters (discussed in detail in Chapter 5) and the final step of creating ProTaS (discussed in detail in this Chapter).

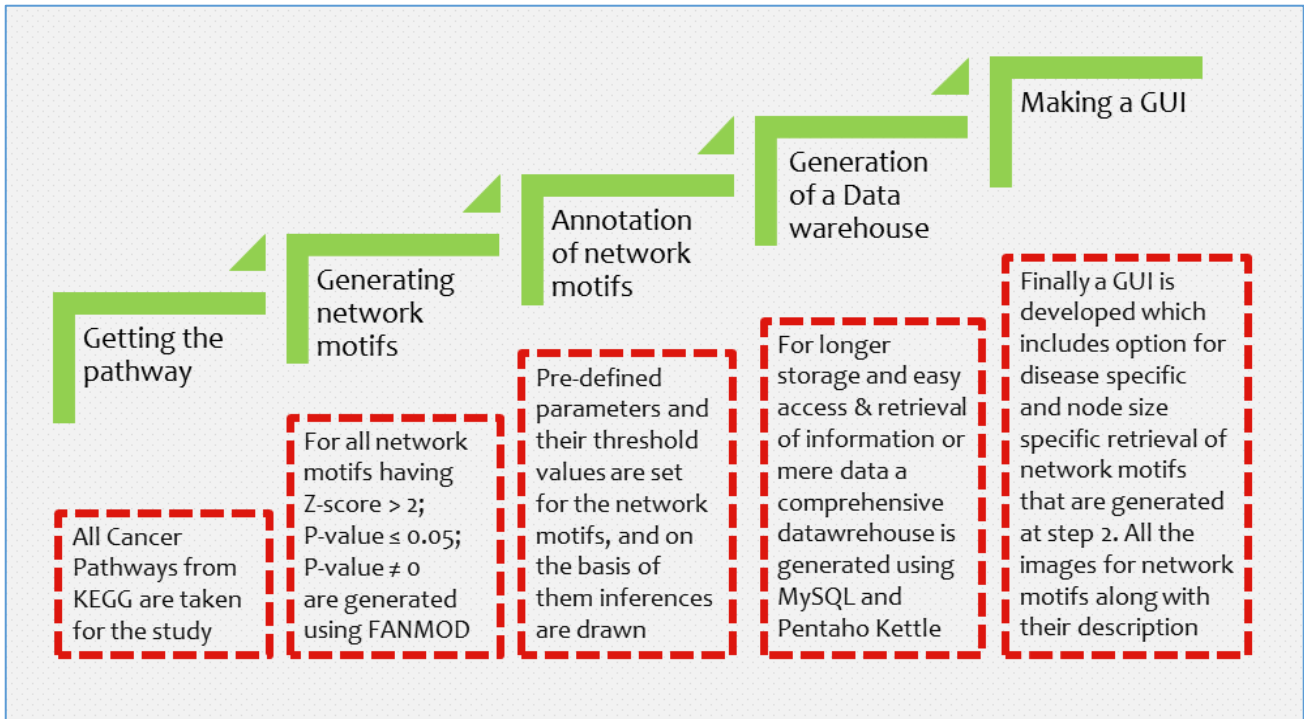


Figure 3.1: ProTaS work flow

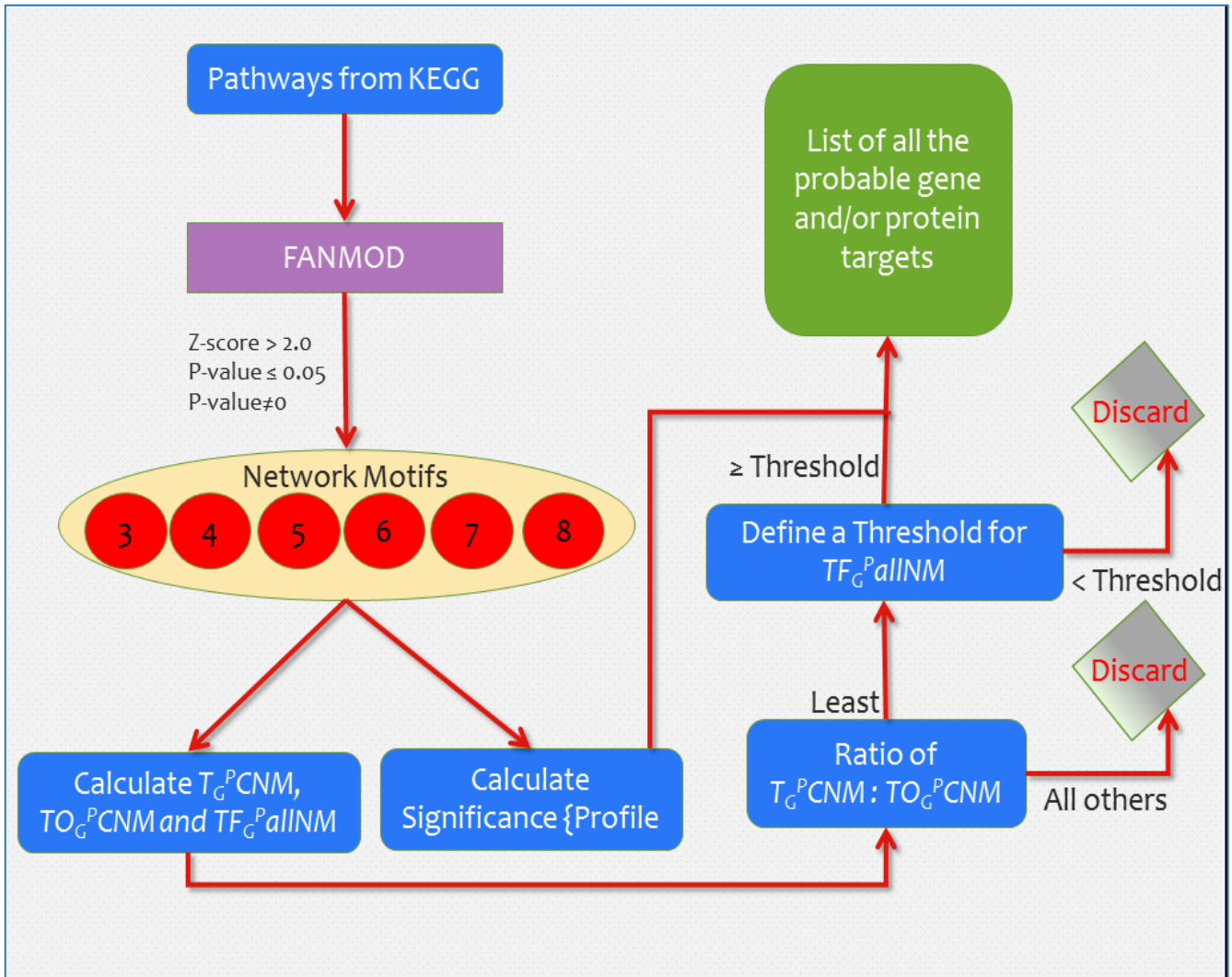


Figure 3.2: Overview of the pipelined methodology

### Getting the pathways

All the pathways for current study are taken from KEGG (Kyoto Encyclopedia of Genes and Genome) [28]; the pathways are taken for *Homo sapiens*. KEGG is a database resource for understanding higher-order functions and utilities of the biological system, such as the cell or the organism, from genomic and molecular information. It is a computer representation of the biological system, consisting of building blocks and wiring diagrams, which can be used for modeling and simulation as well as for browsing and retrieval. The wiring diagrams involve endogenous molecules, that are directly encoded

in the genome (proteins and RNAs) and those that are indirectly encoded through biosynthetic/biodegradation pathways (metabolites, glycans etc.) [29].

### Generating Network Motifs

FANMOD (FAst Network MOtif Detection tool) developed by Wernicke et al. in 2006 [23]. It is used to generate network motifs from both directed and undirected graphs. It can generate network motifs of size 3 to 8. Three values that are taken for study are Z-score, P-value and Significance Profile. The Z-score is the original frequency minus the random frequency divided by the standard deviation of the random network [30]. It is given by the equation:

$$Z - score = \frac{F_1(m) - F_{1,r}(m)}{\sigma_r(m)}$$

where

$F_1(m)$  is Frequency of motif ‘m’ in target network.

$F_{1,r}(m)$  is Frequency of a motif ‘m’ in random network.

$\sigma_r(m)$  is standard deviation of a motif ‘m’ in random network.

The P-value of a motif is the number of random networks in which it occurred more often than in the original network, divided by the total number of random networks. [30] It is given by the equation:

$$P - value = \frac{E}{N^2}$$

where

E is number of edges

$N^2$  is total number of nodes which includes self-edges

More specifically, as in case of FANMOD, it is calculated by the equation given:

$$P - value = \frac{N_{real} - N_{rand}}{N_{totrand}}$$

where

$N_{real}$  is number of times a motif has occurred in the target network

$N_{rand}$  is number of times a motif has occurred in the random network

$N_{totrand}$  is total number of the random networks

The Significance Profile (SP) is a vector of Z-scores of a particular set of motifs which is normalized to 1 [22]. SP is given by the equation:

$$Significance\ Profile = \frac{Z(m_i)}{\sqrt{\sum_{i=1}^n Z(m_i)^2}}$$

where

$Z(m_i)$  is Z-score of network motif number 'i' and size 'm'

n is the number of network motifs

Motifs are considered to be statistically significant and overrepresented if they have a Z-score greater than 2.0 (Kashtan *et al.*, 2002) [22] and the accuracy to be greater than 95% i.e. P-value is  $\leq 0.05$  for a minimum of 1000 random networks [22]; same are taken for this study.

### **Annotation of Network Motifs**

Today we have quite a few approaches and tools [9 – 11, 14, 16, 18] to detect the network motifs from the biological pathways but the real art is not finding the network motifs but extracting some knowledge out of these reoccurring patterns. Here we have introduced three new terminologies to annotate the network motifs; two of which are network motif specific and third one is entity specific, these are given below (For details refer to Chapter 5):

Network motif specific:

- $TO_G^P$ CNM
- $T_G^P$ CNM

Entity specific:

- $TF_G^P$ allNM

Using these terminologies and putting a threshold function on them we have extracted a list of probable targets for the 5 cancer pathways taken into study. The ratio value for  $TO_G^P$ CNM:  $T_G^P$ CNM is taken to be the least and for the value of  $TF_G^P$ allNM a minimum threshold is taken after thorough analysis to establish the authenticity of the final list of targets.

### **Development of ProTaS**

ProTaS stands for Probable Target Search It is a web interface developed to search for the probable targets for a disease pathway. The current release consists of information for only 5 cancer pathways viz. small cell lung cancer, non-small cell lung cancer, pancreatic cancer, prostate cancer and renal cancer. The core of ProTaS comprises of two parts, namely:

- A Database
- A GUI

Development of the database

Data model (Figure 3.1) is generated using CA ERwin Data Modeler, it is created using MySQLQuery Browser version 1.1.20 and all the data transformations are done using Pentaho Kettle (Spoon version 3.1.0). The model has 6 tables: diseases consists of diseaseId, diseaseName, description and KEGG link [28]; nodes consists of nodeId, diseaseId, nodeName and description; freq consists of freqId, nodeId, AdjMatrix, diseaseId and freq; cal\_values consists of valueId, adjMatrix, diseaseId, zscore, pvalue and sp; (For a detailed description on creation of database refer Appendix B and for data transformation commands on Kettle refer Appendix C).

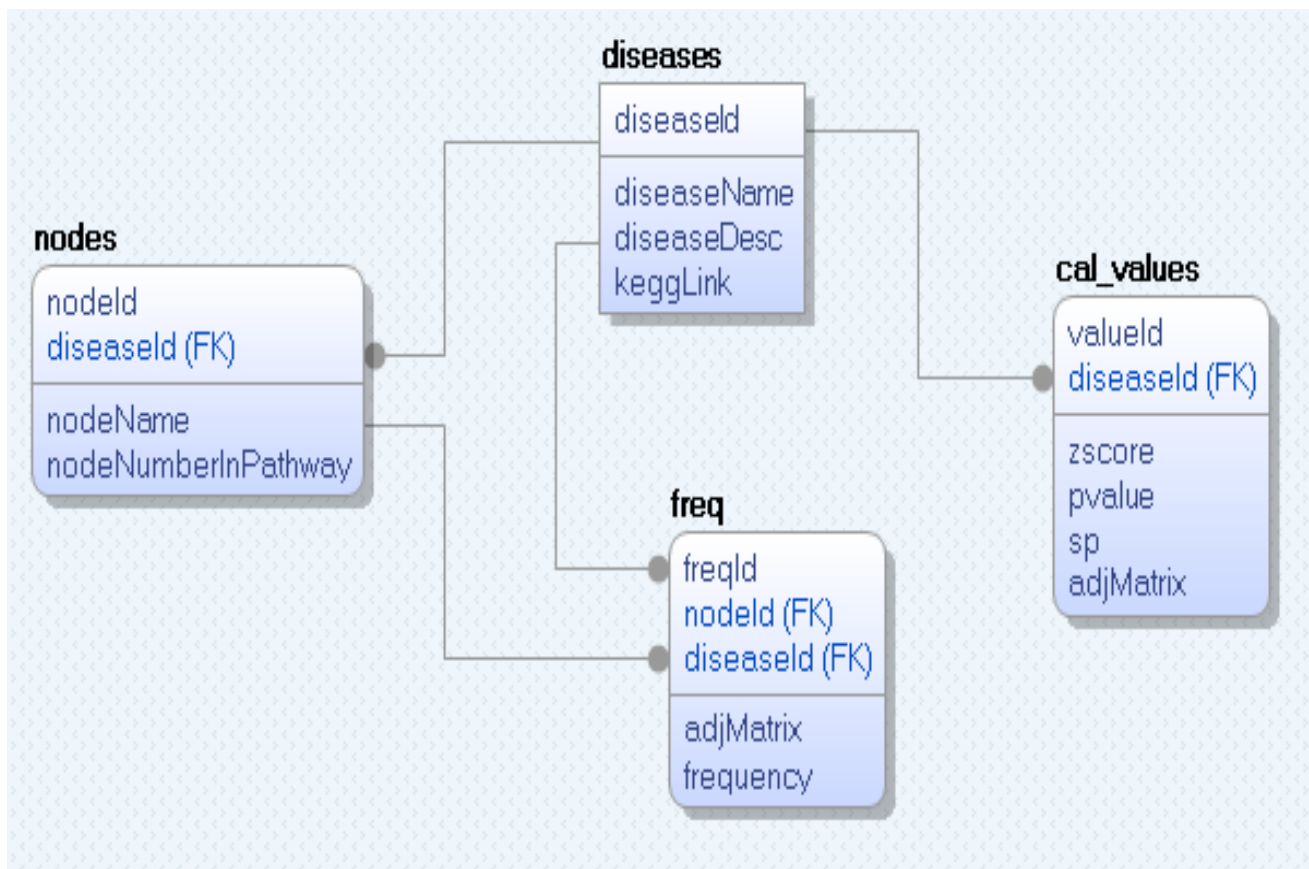


Figure3.3: Data model of the database developed on CA Erwin DM

### Development of the GUI

The GUI can be reached at <http://www.bioinfoindia.org/protas>. The GUI has been developed in HTML5, validation is done using Javascript and database connectivity is established using PHP.

On the GUI, the simple search can be done for the disease by first selecting the desired disease (small cell lung cancer, non-small cell lung cancer, pancreatic cancer, renal cancer and prostate cancer) and the threshold value which ranges from >1 to >10. The complete list of commands working at the backend for retrieving the results can be found at Appendix D. Image (no) gives a look at the GUI how it looks and how the results are displayed.



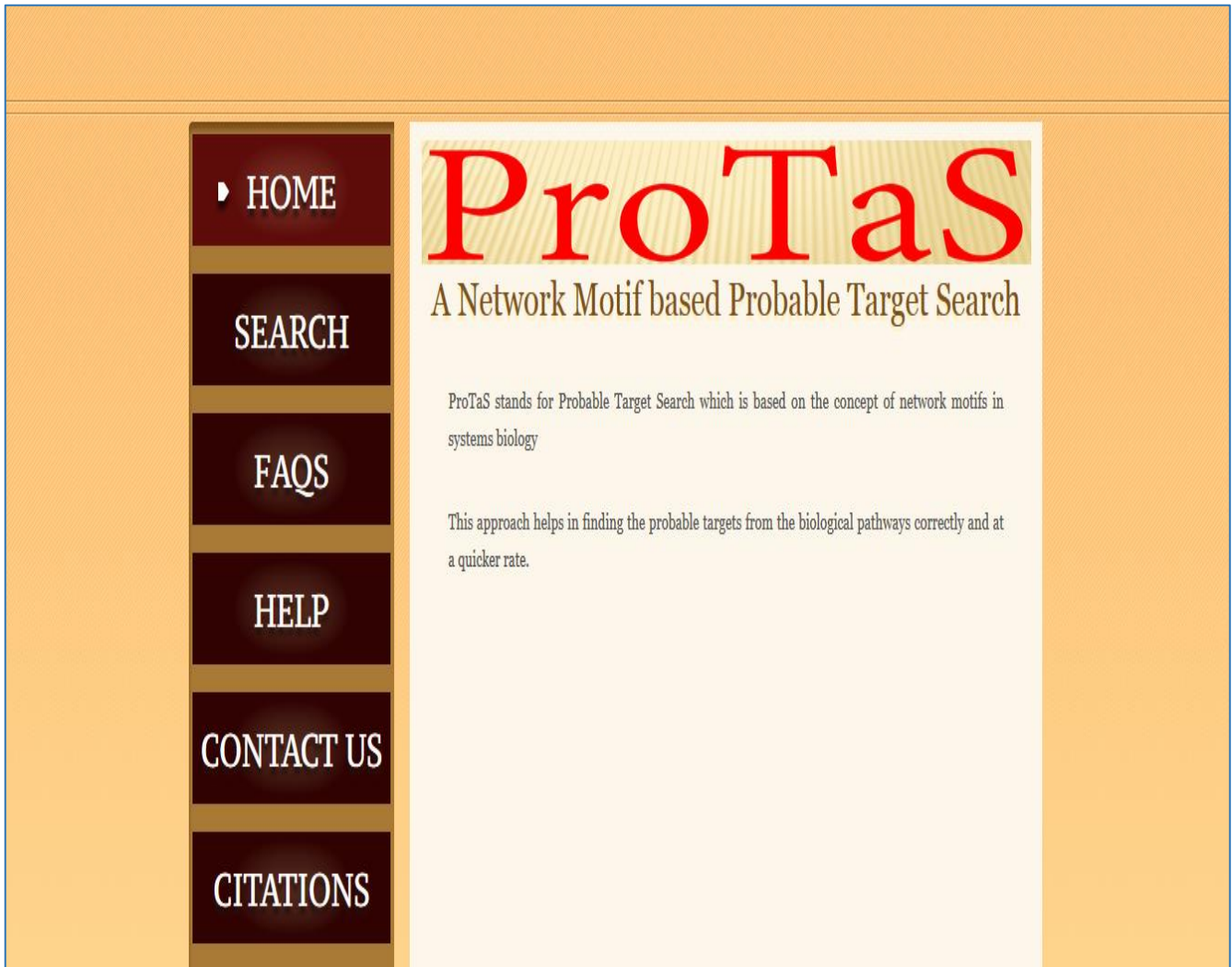


Figure 3.4: Home Screen of ProTaS

HOME

SEARCH

FAQS

HELP

CONTACT US

CITATIONS

DISCLAIMER

# ProTaS

A Network Motif based Probable Target Search

## Search Panel

### Disease Specific Probable Target Search

Select Disease

Select Cancer ▾

Select minimum % frequency of Gene/Protein

> 1% ▾

Submit Reset

Figure 3.5: Search Panel of ProTaS

## Chapter 4

### Results

In our study we have considered 5 cancers viz. non-small cell lung cancer, small cell lung cancer, pancreatic cancer, prostate cancer and renal cancer. The pathways for these are extracted from KEGG in .xml file format. Perl script is used to extract the relationships from the .xml file for the genes and proteins from the pathway (refer to Appendix A (I)), for e.g.: Node 1 of the pathway is interacting (activating, repressing, state change etc.) with some other node, Node 2; and the resultant text file (.txt) is used for the generation of network motifs of node size 3 to 8 using FANMOD. The number of random networks is set to 1000, Z-score  $> 2.0$  and P-value  $\leq 0.05$  [22]. Also the network motifs that are occurring at least 5 times are taken into account; as for the others it gives an undefined value for z-score or p-value or both, making them less significant.

Using FANMOD different files are generated such as .OUT file which has the network motif ID, Adjacency matrix, Frequency (Original), mean-frequency (Random), standard-deviation (Random), Z-Score and p-Value; .dump file has the adjacency matrix and all the corresponding nodes present in that network motif; and .html files having all the images for the network motifs in generalized form i.e. only a representative image for all the network motifs of a particular adjacency matrix. Then aPHPscript is used to generate all the instances of network motifs (adjacency matrices), a summary table (.xls) for frequency of all the nodes for all the network motifs and another .xls file for node numbers in the pathway and their corresponding names for all the pathways in study (refer Appendix A(II)). The summary table is in 2D data distribution to reduce it to a simple 1 to 1 relationship file a script written in Perl is used (refer appendix A (III)).

Some of the generalized network motifs and their set of instances are given in figure 4.1, 4.2 & 4.3 The data files generated viz. summary table, 1 to 1 relationship file, list of nodes etc. are transferred into the database; which is developed using MySQL (refer Appendix B); using Pentaho Kettle (refer Appendix C).

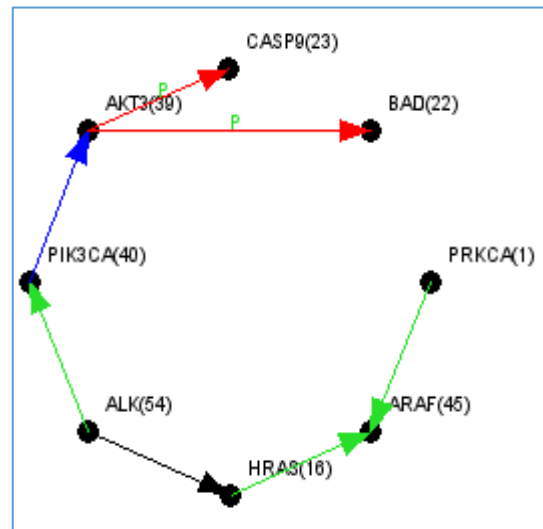
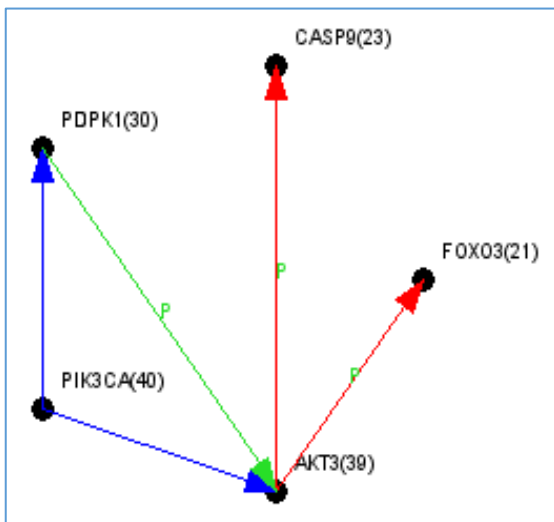
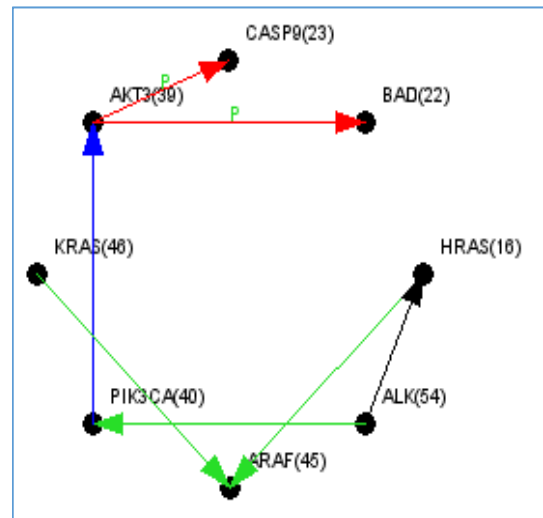
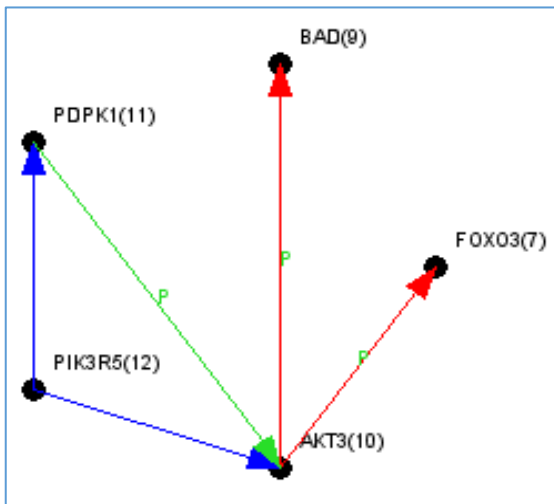
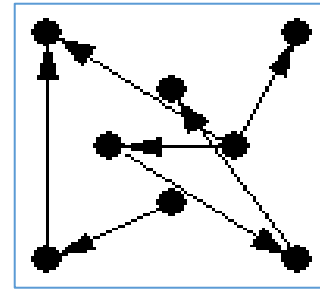
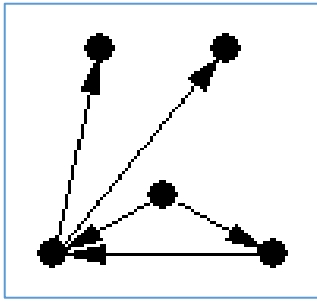


Figure 4.1: Network motif images for Non-small cell Lung Cancer.

**Column 1:** A generalized network motif of size 5 and its corresponding 2 instances.

**Column 2:** Another generalized network motif of size 8 and its corresponding 2 instances.

**Legend:** → Activation → Expression → Repression → Unknown p: Phosphorylation

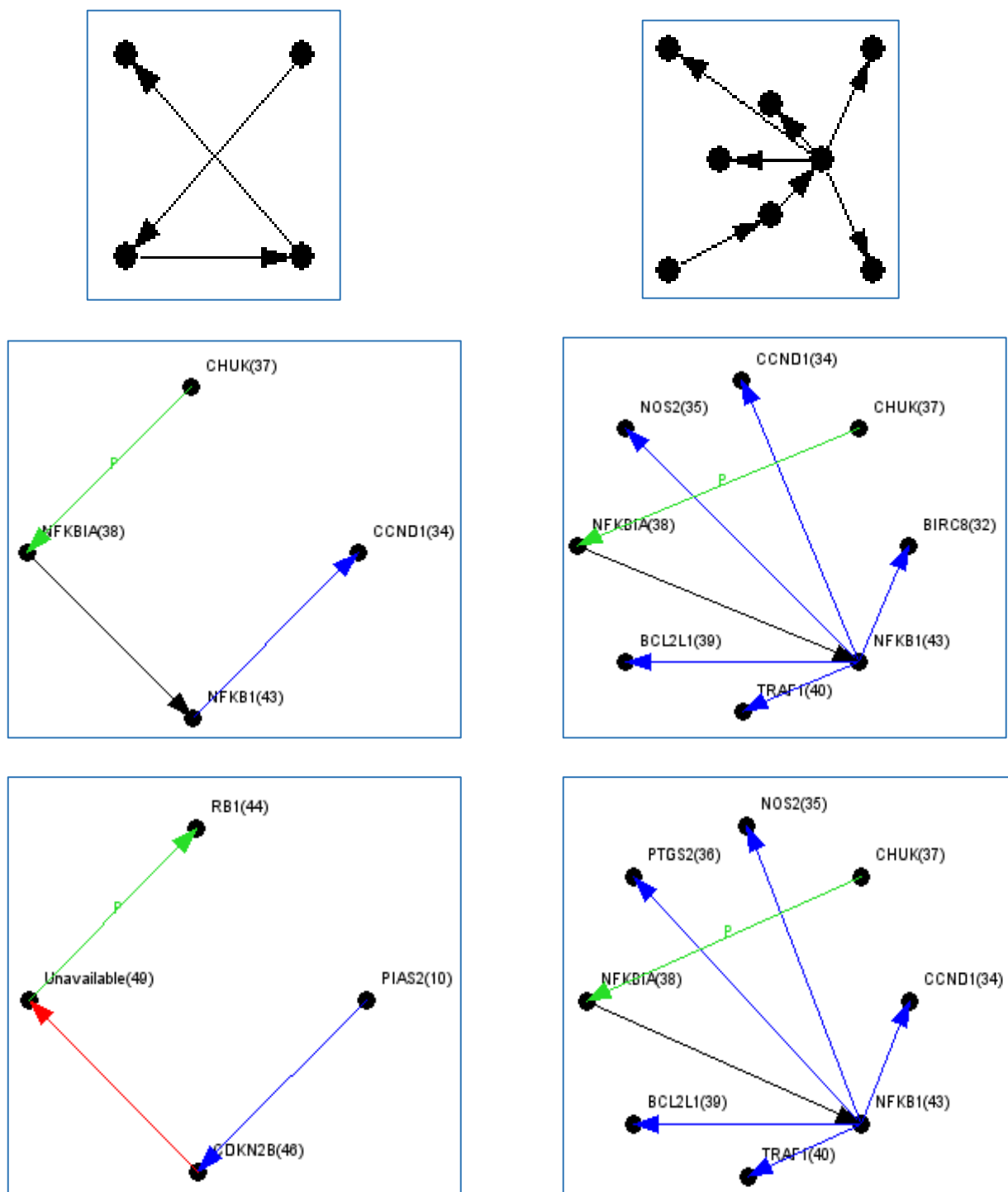


Figure 4.2: Network motif images for small cell Lung Cancer.

**Column 1:** A generalized network motif of size 4 and its corresponding 2 instances.

**Column 2:** Another generalized network motif of size 8 and its corresponding 2 instances.

**Legend:** → Activation → Expression → Repression → Unknown p: Phosphorylation

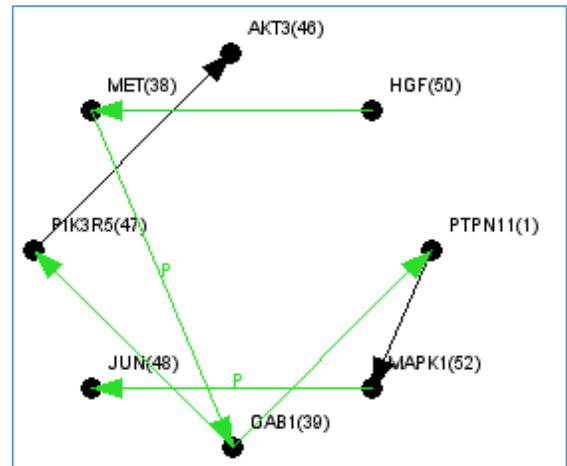
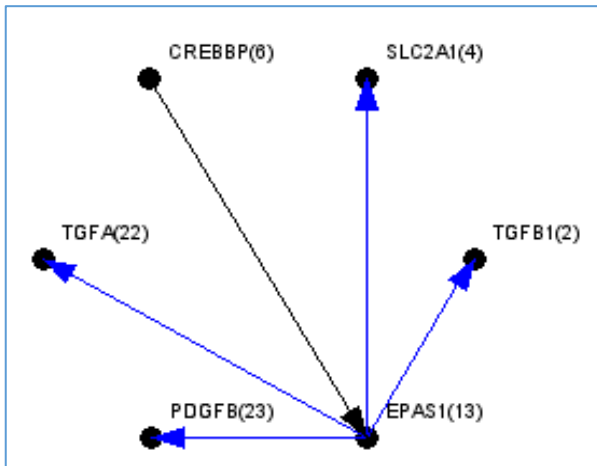
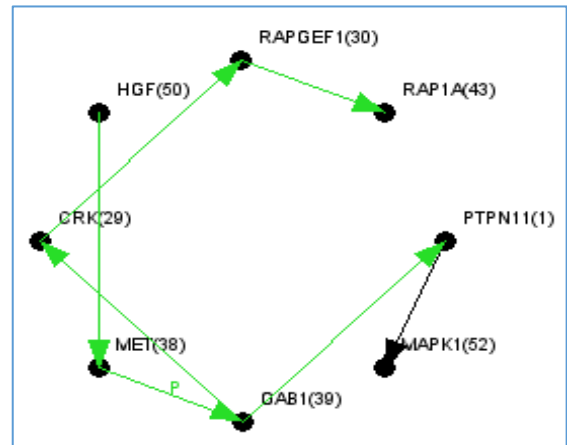
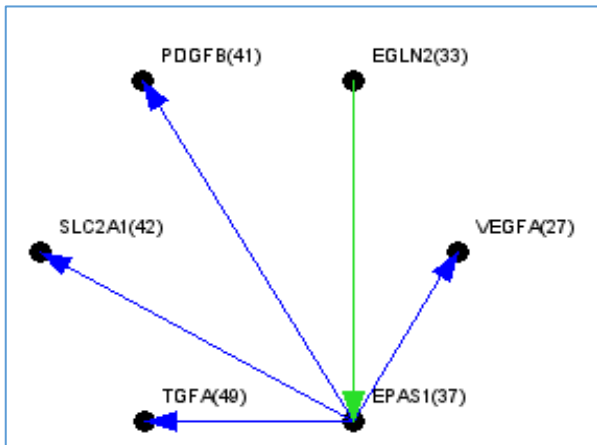
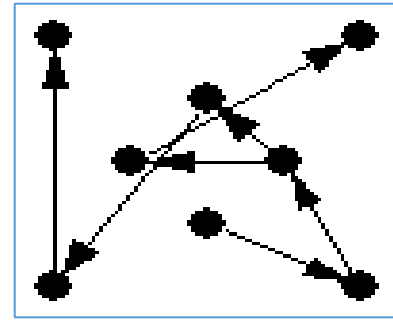
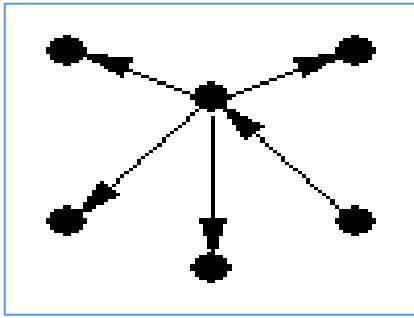


Figure 4.3: Network motif images for Renal Cancer.

**Column 1:** A generalized network motif of size 6 and its corresponding 2 instances.

**Column 2:** Another generalized network motif of size 8 and its corresponding 2 instances.

**Legend:** → Activation → Expression → Repression → Unknown p: Phosphorylation

# ProTaS: Probable Target Search

## Search Results

DISEASE NAME	DESCRIPTION	KEGG LINK
Prostate Cancer	<p>Prostate cancer constitutes a major health problem in Western countries. It is the most frequently diagnosed cancer among men and the second leading cause of male cancer deaths. The identification of key molecular alterations in prostate-cancer cells implicates carcinogen defenses (GSTP1), growth-factor-signaling pathways (NKX3.1, PTEN, and p27), and androgens (AR) as critical determinants of the phenotype of prostate-cancer cells. Glutathione S-transferases (GSTP1) are detoxifying enzymes. Cells of prostatic intraepithelial neoplasia, devoid of GSTP1, undergo genomic damage mediated by carcinogens. NKX3.1, PTEN, and p27 regulate the growth and survival of prostate cells in the normal prostate. Inadequate levels of PTEN and NKX3.1 lead to a reduction in p27 levels and to increased proliferation and decreased apoptosis. Androgen receptor (AR) is a transcription factor that is normally activated by its androgen ligand. During androgen withdrawal therapy, the AR signal transduction pathway also could be activated by amplification of the AR gene, by AR gene mutations, or by altered activity of AR coactivators. Through these mechanisms, tumor cells lead to the emergence of androgen-independent prostate cancer.</p>	<a href="#">KEGG Pathway</a>

Figure 4.4: Search page results; disease name, description and link

### Significant nodes

Node Name	Node Number in KEGG Pathway	Threshold Value	Description Link
Cell cycle	49	21.5188	<a href="#">NCBI</a>
MAP2K1	44	13.3648	<a href="#">NCBI</a>
CASP9	23	9.6607	<a href="#">NCBI</a>
C05981	33	9.4489	<a href="#">NCBI</a>
RXRA	34	7.6678	<a href="#">NCBI</a>
STK4	24	7.6678	<a href="#">NCBI</a>
RASSF1	35	7.6678	<a href="#">NCBI</a>
CCND1	25	7.6678	<a href="#">NCBI</a>
BAD	9	7.6678	<a href="#">NCBI</a>
KRAS	46	7.6678	<a href="#">NCBI</a>

Figure 4.5: Search page results; list of probable nodes, node number in pathway, threshold value and external description link.



### Statistical Parameters

Adjacency Matrix	Z-Score	P-Value	Significance Profile
'000000000001110'	2.7668	0.005	0.77698
'0000100001000110'	4.7348	0.006	1.32964
'0000100010001100'	2.0854	0.015	0.585627
'00000000000000000111100'	3.6156	0.003	0.672637
'000000000000000000011110'	2.9227	0.006	0.543732
'000000000000101100000110'	6.5487	0.001	1.2183
'0000010000010000000101100'	5.4219	0.002	1.00868
'0000010000010000100001100'	6.107	0.003	1.13613
'0000000010010001000001100'	4.2775	0.018	0.795776
'00000000000000000000000001111100'	3.5929	0.004	0.416686
'000000000000000000000000000111110'	2.9276	0.008	0.339528
'00000000000000010001000000001101000'	2.0242	0.027	0.234756
'0000000000000000000000010011000100100'	2.7487	0.018	0.31878
'000000000000100000000010011000000110'	7.5386	0.004	0.874289
'000000000000000100000010010000101000'	2.2809	0.027	0.264527
'000000000000010000000010010000101000'	2.1008	0.045	0.24364

Figure 4.6: Search page results; Adjacency matrix and statistical parameters viz. Z-score, P-Value and Significance Profile.

## Chapter 5

### Annotations

Of no work the sole aim should be to generate results but to extract something useful, something non trivial or still something less apparent, from the study. Same implies for this work, and thus in addition to generation of such big datasets of images, relationship files and adjacency matrices we have devised an approach to work out the most significant players (genes or proteins or else) from the disease pathways. The more a node occurs in the network motifs the more chances are that it may be the central regulator or the regulated around which the fate of the whole network lies. So far no approach has been devised to annotate the network motifs at the level of frequency of the nodes; here we have defined 3 terms to deal with it. These terms are divided into 2 categories:

Network motif specific:

- $TO_G^P\text{CNM}$ : Total occurrences of all Genes/Proteins participating in the conserved Network Motif with respect to the corresponding Network Motif.
- $T_G^P\text{CNM}$ : Total number of all Genes/Proteins participating in the conserved Network Motif with respect to the corresponding Network Motif.

Entity specific:

- $TF_G^P\text{allNM}$ : Total Frequency of a particular Gene/Protein in all Network Motifs.

To understand the concept of these terms let's consider an example of 5 kids and 5 pencil brands. Here the kids correspond to the network motifs and the pencil brands correspond to the biological

entities of the pathway. Say, each kid has certain number of pencils of different brands which are represented in the table 5.1.

Kids	Kid 1	Kid 2	Kid 3	Kid 4	Kid 5	Total pencils of a single brand ( $TF_G^P$ allNM)
Pencil Brands						
ABC	2	0	1	5	0	8
GHI	0	0	1	0	0	1
MNO	3	3	1	0	0	7
PQR	1	3	1	0	3	8
XYZ	1	0	1	0	4	6
Total number of pencils for a kid ( $TO_G^P$ CNM)	7	6	5	5	7	
Total number of pencil brands for a kid ( $T_G^P$ CNM)	4	2	5	1	2	

Table 5.1: Kids and Pencil example to illustrate three terms:  $TO_G^P$ CNM,  $T_G^P$ CNM &  $TF_G^P$ allNM

Now the total number of pencils for a kid is the sum of frequencies of different pencil brands and corresponds to  $TO_G^P$ CNM, likewise total number of pencil brands for a kid corresponds to  $T_G^P$ CNM and total number of pencils of a single brand for all kids corresponds to  $TF_G^P$ allNM. The significance of these terms can be concluded from the fact that if a child has a total of 5 pencils and each is of different brand (as in case for kid 3) i.e. the value of  $T_G^P$ CNM and  $TO_G^P$ CNM are both equal to 5; from such a frequency distribution no significant inferences can be made out. Consider another case of kid 5 who has a total of 7 pencils from only 2 brands out of 5, which shows his liking for the pencil brands and thus puts the 2 pencil brands at a higher linking index for the kids. Now for the last parameter let's stretch the example of kid 5 a little further; the two pencil brands that are at a higher liking are PQR and XYZ now before coming out at the final conclusions let's check their liking for all the kids. Here comes in action our third parameter i.e.  $TF_G^P$ allNM, which gives us a total count of a pencil brand for all kids and if the two brands (in case of kid 5) still have a higher frequency, then they are our target brands (nodes, in case of our study), which needs to be studied further for establishing their shew.

The ratio of  $T_G^P\text{CNM} : TO_G^P\text{CNM}$  should be as less as possible because lower value implies that the number of biological entities that are participating in the given network motif are having a higher frequency in that particular network motif; which implies that, that a certain pattern of the pathway is being controlled by some genes and proteins having a higher frequency. Secondly, a threshold on the value of  $TF_G^P\text{allNM}$  is set after a thorough analysis. This value is for all the genes and proteins involved in the corresponding network motif selected by the least ratio of  $T_G^P\text{CNM} : TO_G^P\text{CNM}$ . Now, all the entities having the value of  $TF_G^P\text{allNM}$  higher than the set threshold are further analyzed for their biological activity and their role in the present diseased pathway and also other pathways.

The resultant set of genes and proteins for, in study, 5 cancers are given in the table 5.2. Only for non-small cell lung cancer the threshold is set to 5 % for all other diseases it is kept as 7% because in case of non-small cell lung cancer at threshold 7% very few significant nodes are present so we have to lower down the bar for it. A total of 10 significant genes or proteins are given for non-small cell lung cancer, 9 for small cell lung cancer, 7 for pancreatic cancer, 9 for prostate cancer and 7 for renal cancer. The threshold value can be varied across different values to get the most optimized set of probable targets. The higher the threshold-value of the entity the more chances of it to be the prime target. An important distinction that needs to be made here is the higher frequency of certain nodes because of falsified connections; so a thorough analysis of the pathway and the final result sets need to be made with inordinate precision.

The main point to note here is the presence of PIK3CA, AKT3& GRB2 for both non-small cell lung cancer and small cell lung cancer as the significant gene or protein; EGF& FOXO3 in non-small cell lung cancer and renal cancer; BAD & CASP9 in non-small cell lung cancer and prostate cancer; RASSF5 in small cell lung cancer & renal cancer; RXRA in small cell lung cancer and prostate cancer; C05981 in prostate cancer and renal cancer; CCND1 & RASSF1 in small cell lung cancer, renal cancer and prostate cancer. Beforehand such kind of similarities between related (as small cell lung cancer and non-small cell lung cancer) or unrelated (small cell lung cancer and prostate cancer) cancers was not known. With the advent of this approach it became just a matter of few pipelined computations to get to the results. Such kind of similarities can be of immense help and need to the medical sciences to visit

Name of Cancer	Threshold value of $TF^G_{pallNM}$ (%)	List of probable targets	Actual $TF^G_{pallNM}$
Non-small cell Lung cancer	5	PIK3CA	18.5364
		AKT3	15.4903
		ERBB2	10.7355
		EGFR	9.3115
		GRB2	9.2249
		EGF	6.5750
		TGFA	6.5750
		FOXO3	6.3150
		BAD	6.3150
		CASP9	6.3150
Small cell Lung cancer		MAPK1	18.9221
		CDKN2A	16.7545
		CCND1	11.5993
		AKT3	8.7873
		RASSF5	8.7873
		PIK3CA	8.7873
		GRB2	8.7873
		RXRA	8.7873
		RASSF1	8.7873
Pancreatic cancer	7	PRKCA	16.2602
		C00076	16.1698
		'PI3K-Akt signaling pathway	15.8988
		C01245	12.1048
		PDPK1	10.0271
		E2F1	10.0271
		C00165	7.7687

Prostate Cancer	7	MAP2K1	13.3648
		CASP9	9.6607
		C05981	9.4489
		CCND1	7.6678
		BAD	7.6678
		KRAS	7.6678
		RX\RA	7.6678
		STK4	7.6678
		RASSF1	7.6678
Renal cancer	7	CCND1	13.1481
		C05981	10.6173
		RASSF1	7.1605
		RASSF5	7.1605
		FHIT	7.0988
		FOXO3	7.0988
		EGF	7.0988

Table 5.2: List of significant nodes for 5 diseases taken into study and their respective thresholds

the lands unheard of and truths to be unveiled for achieving a science which can predict the probable targets in the diseased pathways at will and that too at significantly reduced stretch of time. List of significant nodes for non- small cell lung cancer disease is given from Figure 5.1.

### Significant nodes

Node Name	Node Number in KEGG Pathway	Threshold Value	Description Link
PIK3CA	40	18.5364	<a href="#">NCBI</a>
AKT3	39	15.4903	<a href="#">NCBI</a>
ERBB2	26	10.7355	<a href="#">NCBI</a>
EGFR	20	9.3115	<a href="#">NCBI</a>
GRB2	32	9.2249	<a href="#">NCBI</a>
EGF	27	6.5750	<a href="#">NCBI</a>
TGFA	19	6.5750	<a href="#">NCBI</a>
FOXO3	21	6.3150	<a href="#">NCBI</a>
BAD	22	6.3150	<a href="#">NCBI</a>
CASP9	23	6.3150	<a href="#">NCBI</a>

Figure 5.1: List of Significant nodes for Non-small Cell Lung Cancer at threshold > 5%

### Significant nodes

Node Name	Node Number in KEGG Pathway	Threshold Value	Description Link
CCND1	37	13.1481	<a href="#">NCBI</a>
C05981	6	10.6173	<a href="#">NCBI</a>
RASSF1	35	7.1605	<a href="#">NCBI</a>
RASSF5	36	7.1605	<a href="#">NCBI</a>
EGF	27	7.0988	<a href="#">NCBI</a>
FHIT	41	7.0988	<a href="#">NCBI</a>
FOXO3	21	7.0988	<a href="#">NCBI</a>

Figure 5.2: List of Significant nodes for Renal Cancer at threshold > 7%

## Chapter 6

### Inferences

In chapter 5, we saw how using this approach has brought to light the hidden relationships among different disease pathways. In many a cases, such conditions are a routine where a patient is not only suffering from just one disease but may be more than that in such cases, provided the disease pathway, we can establish the relationships between the different diseases, and that on the basis of solid grounds rather than just assumptions. Here we have mathematical parameters to hold the ground for the results that we put forward. Also, if we go through a logical reasoning, a greater involvedness of a node across the whole network (or pathway) puts some weight on the need for the system to do so. As the system has put some of its energy, resources and time into it, for it to be like the way we see it.

One main thing that has got our attention is the strange relationship between related or unrelated cancers. As in case of non-small cell lung cancer and small cell lung cancer three genes and/or proteins were found to be similar namely; PIK3CA, AKT3& GRB2. Similarly for non-small cell lung cancer and renal cancer two were found to be same namely; EGF& FOXO3 and more for other pair of diseases. This makes it clear that these elements are maintained by the biological systems as either the main regulator or the most regulated targets.

For all these probable targets a study of their roles in biological world and the ways it interact with other systems and its entities can still produce more fascinating results. The restriction on our study is implied by the number of diseases we are talking of, at that particular time, which hinders the complete knowledge about the node, which can also be involved in other diseases, and can be a prime target in those diseases, too. Another restriction is imposed by the unavailability of complete pathways of many diseases, which restricts our goal. Also in a computational model of a biological system, any compound if goes under a state transition or conformational change is denoted by a different entity on the computer system whereas in the real world it is not two different compounds but one; so in such cases the frequency of such nodes is tend to be high and is a false indication of it to be a probable target



and thus needs to be taken care of. So, the final list needs to be checked before proposing the set of probable targets.

The one major thing that needs to be done is to establish the authenticity of the inferences made from the model and to show the validity of the approach. We have to perform the analysis on real data sets from medical records to affirm its cogency. Adding on to it, the targets predicted need to be tested, whether they actually help in resolving the problem with the system. In case the model holds true for its prediction for the real data sets, the big things to come out of it are unbelievable, fascinating and amazing. The near possibilities that my knowledge and intellect makes apparent include: 1) What if, in coming years, we have a single medicine for 2 or 3 or more diseases? 2) What if, the search for the targets can be performed at fingertips? 3) What if, the drug discovery process cuts a little time out of its course? 4) What if, the fear of new diseases is not a fear anymore?

Also if the targets detected by this model are in accordance with the traditionally established facts, it can provide an easy and fast methodology to unveil the targets for new or still ambiguous pathways. It is fast and can be followed to investigate any pathway and hence increases the efficiency of our work, without pushing on cost, time and labor.

## Appendix A (I)

Script for generation of relationships file from the .xml file.

```
#!/C:/strawberry/perl/bin

print "enter the file name ";

$path =<STDIN>;

chomp $path;

unless(open (FILE,$path)){

print " cannot open the file $path ";

exit ;

}

open FH, ">nm.txt" or die $!;

@file =<FILE>;

close FILE;

$l=@file ;

$arr;$j=0;

for($i=0;$i<$l;$i++){

$w=substr($file[$i],0,20);

#print "test $w\n";

if($w eq " <relation entry1"){

# print "$w\n";

$arr[$j]=$file[$i];
```

```

$j++;
}
}

$l=@arr;

@adj;$r=0,$c=0;
for($r=0;$r<$l;$r++){
for($c=0;$c<2;$c++){
$l=length $arr[$r];
for($i=0;$i<$l;$i++){
$sub=substr($arr[$r],$i,1);
if ($sub eq '='){
$adj[$r][$c]=substr($arr[$r],($i+2),2);$c++;
#print $i;print"$adj[$r][$c]";
}
}
}
}

for($r=0;$r<$l;$r++){
for($c=0;$c<2;$c++){
print FH "$adj[$r][$c] ";
}print FH"\n";
}

```

## Appendix A (II)

Script for generation of image instances, summary of frequencies and list of nodes from the .xml file, .OUT, .dump files generated by the tool FANMOD and a .csv file generated from the images by FANMOD.

(For non-small lung cancer)

```
<?php
    ini_set('max_execution_time', 100000000);

    if(!isset($_GET['dis']) || !isset($_GET['csvfilename']) || !isset($_GET['xmlfilename']))
    {
        echo ' <br><br><form method="get"><br>Select Disease: (To add more diseases add
more options in the file)<br><select size="2" name="dis">

                <option>non small lung cancer</option>

                <option>pancreatic cancer</option>

                <option>prostate cancer</option>

                <option>renal cancer</option>

                <option>small lung cancer</option>

        </select><br><br>

        <br>CSV File Name: <input type="text" name="csvfilename">

        <br>XML File Name: <input type="text" name="xmlfilename">

        <br><input type="submit"></form>;

        die();
    }
}
```

```

else
{
    $disease=$_GET['dis'];
    $csvfilename=$_GET['csvfilename'];
    $xmlfilename=$_GET['xmlfilename'];
    $disease = strtolower($disease);
    if (!file_exists($disease))
    {
        mkdir($disease, 0755, true);
    }
    file_put_contents($disease.'/NamesOfIds.csv','');
    $xmlfiledata=file_get_contents($xmlfilename);
    $data=explode('<entry id=""',$xmlfiledata);
    unset($data[0]);
    $allnames=array();
    foreach($data as $d)
    {
        $nameofid="Unavailable";
        $data2=explode('<graphics',$d);
        $sid=explode('',$d,2);
        $data2=explode('>',$data2[1],2);
        if(strpos($data2[0],"name"))
        {
            $name=explode('name=""',$data2[0],2);

```

```

$name=explode("", $name[1]);

if(strstr($name[0], ","))

    $name=explode(',', $name[0], 2);

$nameofid=$name[0];

}

$line2="".$id[0].", ".$nameofid.".\\n";

$allnames[$id[0]]=$nameofid;

file_put_contents($disease.'/NamesOfIds.csv', $line2, FILE_APPEND |
LOCK_EX);

}

$totalxmlids=count($allnames);

$fileh1=fopen($csvfilename, "r");

echo 'Execution started...';

$newfile=fopen($disease."/".$disease.".txt", "w");

$allmatrices=array();

$scorematrix=array(array());

$totalmatrices=0;

while($soyline=fgets($fileh1))

{

    $soyline = strtolower($soyline);

    if(strstr($soyline, $disease))

    {

        $soyline=explode("\\", $soyline, 17);

        $soyline[15]=basename($soyline[15], ".png");

```

```

$yno=$soyline[7];

$mat=$soyline[15];

$amat = str_split($mat, $yno);

$i=0;

$adj=array(array());

foreach($amat as $val)

    $adj[$i++]=str_split($val,1);

if (file_exists("nm".$yno.".txt.OUT.dump"))

{

    $handle = fopen("nm".$yno.".txt.OUT.dump", "r");

    $g_count=1;

    $font='arial.ttf';

    if (!file_exists($disease.'/'.$yno))

    {

        mkdir($disease.'/'.$yno, 0755, true);

    }

    if (!file_exists($disease.'/'.$yno.'/'.$mat))

    {

        mkdir($disease.'/'.$yno.'/'.$mat, 0755, true);

    }

    if ($handle)

    {

        $array1=array_fill(0,150, 0);

        while (($line = fgets($handle)) !== false)

```

```

{
    if(strstr($line,$mat))
    {
        $vcoords=array(array());
        $graph=imagecreate(400,400);
$white=imagecolorallocate($graph,255,255,255);
        $x=250;
        $y=220;
        $black=imagecolorallocate($graph,0,0,0);
$green=imagecolorallocate($graph,40,222,40);
$red=imagecolorallocate($graph,255,0,0);
$blue=imagecolorallocate($graph,0,0,255);
$line=trim(preg_replace('/\s\s+/', '', $line));
        $sids=explode(",",$line);
        $sidies=array_shift($sids);
        $i=0;
$vertices=getpoints(400/2,400/2,100,$vno);
        $vpos=0;
        foreach ($sids as $sid)
        {
            $sarray1[$sid]++;
            //echo 'Vertex number='.(($i+1).' Id
= '$sid.' and name is '$allnames[$sid].<br>';
            imagefilledellipse($graph,$vertices[$vpos],$vertices[$vpos+1],10,10,$black);

```



```

                                imgettext
($graph,7,0,$vertices[$vpos]+9,$vertices[$vpos+1]-9,$black,$font,$allnames[$id].("$id."));

                                $vcoords[$i][0]=$vertices[$vpos];

    $vcoords[$i][1]=$vertices[$vpos+1];

                                $vpos=$vpos+2;

                                $i++;

                                }

                                for($m=0;$m<$vno;$m++)

                                    for($n=0;$n<$vno;$n++)

                                        {

$xml_file=fopen($xmlfilename,"r");

                                while(($line=fgets($xml_file))!==false)

                                    {

                                if(

preg_match('/entry1="'.$sids[$m].'" entry2="'.$sids[$n].'"', $line))

                                    {

                                fputs($newfile,$sids[$m]." ".$sids[$n]."\n");

                                                                //echo

'Relation between: '.$sids[$m].' and '.$sids[$n].'<br>';

                                while(!strstr($line, "/relation>"))

                                                                {

                                if(preg_match('/name=/', $line))

```

```
{
```

```
$relation=explode("\\", $line, 3);  
//echo $relation[1].<br>;  
  
if($relation[1]=="activation")  
connectvertex($graph,$vcoords[$m][0],$vcoords[$m][1],$vcoords[$n][0],$vcoords[$n][1],15,5,  
$green)  
  
else if($relation[1]=="inhibition")  
  
connectvertex($graph,$vcoords[$m][0],$vcoords[$m][1],$vcoords[$n][0],$vcoords[$n][1],15,5,  
$red);  
  
else if($relation[1]=="phosphorylation")  
{  
    $cx=intval(($vcoords[$m][0]+$vcoords[$n][0])/2);  
    $cy=intval(($vcoords[$m][1]+$vcoords[$n][1])/2);  
    imagettftext ($graph,7,0,$cx,$cy,$green,$font,"P");  
}  
  
else if($relation[1]=="expression")  
connectvertex($graph,$vcoords[$m][0],$vcoords[$m][1],$vcoords[$n][0],$vcoords[$n][1],15,5,  
$blue);  
  
else  
connectvertex($graph,$vcoords[$m][0],$vcoords[$m][1],$vcoords[$n][0],$vcoords[$n][1],15,5,  
$black);  
  
}  
  
$line=fgets($xml_file);  
  
}
```

```

}
fclose($xml_file);
}

imagepng($graph,$disease.'/'.$vno.'/'.$mat.'/'.$g_count.'.png',0);

        imagedestroy($graph);

        $g_count++;

        //echo '<br>_____NEXT
GRAPH_____<br>';

        }

    }

    for($idno=1;$idno<=$totalxmlids;$idno++)
    {

$scorematrix[$idno][$totalmatrices]=$array1[$idno];

    }

    $allmatrices[$totalmatrices++]=$mat;

    fclose($handle);

    }

}

}

$headerstring="";

$line="";

for($i=0;$i<$totalmatrices;$i++)

{

```

```

        $headerstring=$headerstring."\t".$allmatrices[$i]."";
    }

    file_put_contents($disease.'/Summary.xls', $headerstring."\n");

    for($j=1;$j<=$totalxmlids;$j++)
    {
        $line="".$j;

        for($i=0;$i<$totalmatrices;$i++)
        {
            $line=$line."\t".$scorematrix[$j][$i];
        }

        $line.="\n";

        file_put_contents($disease.'/Summary.xls', $line,FILE_APPEND | LOCK_EX);
    }

    fclose($fileh1);

    fclose($newfile);

    echo '<br><br>Script has been executed and completed!<br><a href="/.nm.php">Click
to Go BACK</a>';

}

function connectvertex($im, $x1, $y1, $x2, $y2, $alength, $awidth, $color)
{
    $distance = sqrt(pow($x1 - $x2, 2) + pow($y1 - $y2, 2));

    $dx = $x2 + ($x1 - $x2) * $alength / $distance;

    $dy = $y2 + ($y1 - $y2) * $alength / $distance;

    $k = $awidth / $alength;

```

```

$x2o = $x2 - $dx;

$y2o = $dy - $y2;

$x3 = $y2o * $k + $dx;

$y3 = $x2o * $k + $dy;

$x4 = $dx - $y2o * $k;

$y4 = $dy - $x2o * $k;

imageline($im, $x1, $y1, $dx, $dy, $color);

imagefilledpolygon($im, array($x2, $y2, $x3, $y3, $x4, $y4), 3, $color);
}

function getpoints($centerx,$centery,$dist,$vno)
{
    $points = array();
    for($a = 0;$a <= 360; $a += 360/$vno)
    {
        $points[] = $centerx + $dist * cos(deg2rad($a));
        $points[] = $centery + $dist * sin(deg2rad($a));
    }
    return $points;
}
?>

```

## Appendix A (III)

Reducing the summary file generated from previous from 2D data to 1D data

```
#!/C:/strawberry/perl/bin

print "enter the file name\n";

$path =<STDIN>;

chomp $path;

unless(open (FILE,$path))

    {

        print " cannot open the file $path ";

        exit ;

    }

@file =<FILE>;

close FILE;

$line=@file[0];

@arr1= split("\t",$line);

chomp @arr1;

open (OUT, ">summary.txt");

    print OUT "ImageId","\t","NodeId","\t","Frequency","\n";

    close (OUT);

my $x=1;
```

```
for($i=1;$i<@arr1;$i++)
{
  for($j=1;$j<@file;$j++)
  {
    $temp=@file[$j];
    @arr2=split("\t",$temp);
    chomp @arr2;
    open (OUT, ">>summary.txt");
    print OUT "$arr1[$i]","\t","$arr2[0]","\t","$arr2[$x]","\n";
    close (OUT);
  }
  $x++;
}
}
```

## Appendix B – Database creation commands on MySQL

1. CREATE DATABASE protas;
2. USE DATABASE protas;
3. CREATE TABLE `protas`.`diseases` (  
  
    `diseaseId` INTEGER UNSIGNED NOT NULL AUTO\_INCREMENT,  
  
    `diseaseName` VARCHAR(50) NOT NULL DEFAULT "",  
  
    `diseaseDesc` VARCHAR(2000) "",  
  
    `keggLink` VARCHAR(200) "",  
  
    PRIMARY KEY(`diseaseId`)  
  
)  
  
ENGINE = InnoDB;
4. CREATE TABLE `protas`.`g\_images` (  
  
    `generalizedId` INTEGER UNSIGNED NOT NULL AUTO\_INCREMENT,  
  
    `diseaseId` INTEGER UNSIGNED NOT NULL DEFAULT 0,  
  
    `generalizedImage` VARCHAR(200) NOT NULL DEFAULT "",  
  
    PRIMARY KEY(`generalizedId`),  
  
    CONSTRAINT `FK\_g\_images\_1` FOREIGN KEY `FK\_g\_images\_1` (`diseaseId`)  
  
        REFERENCES `diseases` (`diseaseId`)  
  
        ON DELETE RESTRICT  
  
        ON UPDATE RESTRICT  
  
)  
  
ENGINE = InnoDB;
5. CREATE TABLE `protas`.`i\_images` (



```

`instanceId` INTEGER UNSIGNED NOT NULL AUTO_INCREMENT,
`diseaseId` INTEGER UNSIGNED NOT NULL DEFAULT 0,
`generalizedId` INTEGER UNSIGNED NOT NULL DEFAULT 0,
`instanceImage` VARCHAR(200) NOT NULL DEFAULT "",
PRIMARY KEY(`instanceId`),
CONSTRAINT `FK_i_images_1` FOREIGN KEY `FK_i_images_1` (`diseaseId`)
REFERENCES `diseases` (`diseaseId`)
ON DELETE RESTRICT
ON UPDATE RESTRICT,
CONSTRAINT `FK_i_images_2` FOREIGN KEY `FK_i_images_2` (`generalizedId`)
REFERENCES `g_images` (`generalizedId`)
ON DELETE RESTRICT
ON UPDATE RESTRICT
)
ENGINE = InnoDB;

```

```

6. CREATE TABLE `protas`.`nodes` (
`nodeId` INTEGER UNSIGNED NOT NULL AUTO_INCREMENT,
`diseaseId` INTEGER UNSIGNED NOT NULL DEFAULT 0,
`nodeName` VARCHAR(50) NOT NULL DEFAULT "",
`nodeDesc` VARCHAR(200),
PRIMARY KEY(`nodeId`)
)
ENGINE = InnoDB;

```

```

7. CREATE TABLE `protas`.`freq` (

```

```

`freqId` INTEGER UNSIGNED NOT NULL AUTO_INCREMENT,
`diseaseId` INTEGER UNSIGNED NOT NULL DEFAULT 0,
`generalizedId` INTEGER UNSIGNED NOT NULL DEFAULT 0,
`nodeId` INTEGER UNSIGNED NOT NULL DEFAULT 0,
`frequency` VARCHAR(45) NOT NULL DEFAULT "",
PRIMARY KEY(`freqId`),
CONSTRAINT `FK_freq_1` FOREIGN KEY `FK_freq_1` (`diseaseId`)
REFERENCES `diseases` (`diseaseId`)
ON DELETE RESTRICT
ON UPDATE RESTRICT,
CONSTRAINT `FK_freq_2` FOREIGN KEY `FK_freq_2` (`generalizedId`)
REFERENCES `g_images` (`generalizedId`)
ON DELETE RESTRICT
ON UPDATE RESTRICT,
CONSTRAINT `FK_freq_3` FOREIGN KEY `FK_freq_3` (`nodeId`)
REFERENCES `nodes` (`nodeId`)
ON DELETE RESTRICT
ON UPDATE RESTRICT
)
ENGINE = InnoDB;

```

```

8. CREATE TABLE `protas`.`cal_values` (
`valueId` INTEGER UNSIGNED NOT NULL AUTO_INCREMENT,
`diseaseId` INTEGER UNSIGNED NOT NULL DEFAULT 0,
`generalizedId` INTEGER UNSIGNED NOT NULL DEFAULT 0,

```

```
`zscore` FLOAT,  
`pvalue` FLOAT,  
`sp` FLOAT,  
PRIMARY KEY(`valueId`),  
CONSTRAINT `FK_cal_values_1` FOREIGN KEY `FK_cal_values_1` (`diseaseId`)  
REFERENCES `diseases` (`diseaseId`)  
ON DELETE RESTRICT  
ON UPDATE RESTRICT,  
CONSTRAINT `FK_cal_values_2` FOREIGN KEY `FK_cal_values_2` (`generalizedId`)  
REFERENCES `g_images` (`generalizedId`)  
ON DELETE RESTRICT  
ON UPDATE RESTRICT  
)  
ENGINE = InnoDB;
```

## Appendix C –Data transformation commands on Pentaho Kettle

1. Populating the disease table in the database.



```
<?xml version="1.0" encoding="UTF-8"?>
```

```
<transformation-steps>
```

```
<steps>
```

```
<step>
```

```
<name>Table output disease</name>
```

```
<type>TableOutput</type>
```

```
<description/>
```

```
<distributed>Y</distributed>
```

```
<copies>1</copies>
```

```
<partitioning>
```

```
<method>none</method>
```

```
<schema_name/>
```

```
</partitioning>
```

```
<connection>protas</connection>
```

```
<schema/>
```

```
<table>diseases</table>
```

```
<commit>100</commit>
```

```
<truncate>Y</truncate>
```

```

<ignore_errors>N</ignore_errors>

<use_batch>N</use_batch>

<partitioning_enabled>N</partitioning_enabled>

<partitioning_field/>

<partitioning_daily>N</partitioning_daily>

<partitioning_monthly>Y</partitioning_monthly>

<tablename_in_field>N</tablename_in_field>

<tablename_field/>

<tablename_in_table>Y</tablename_in_table>

<return_keys>N</return_keys>

<return_field>id</return_field>

<cluster_schema/>

<remotesteps> <input> </input> <output> </output> </remotesteps> <GUI>

  <xloc>535</xloc>

  <yloc>144</yloc>

  <draw>Y</draw>

</GUI>

</step>

<step>

  <name>CSV file input disease</name>

  <type>CsvInput</type>

  <description/>

  <distribute>Y</distribute>

  <copies>1</copies>

```

```
<partitioning>
  <method>none</method>
  <schema_name/>
</partitioning>
<filename>C:\Users\Rajinder\Desktop\transformation files protas\diseases.csv</filename>
<filename_field/>
<rownum_field/>
<include_filename>N</include_filename>
<separator>,</separator>
<enclosure>&quot;</enclosure>
<header>Y</header>
<buffer_size>50000</buffer_size>
<lazy_conversion>Y</lazy_conversion>
<add_filename_result>N</add_filename_result>
<parallel>N</parallel>
<encoding/>
<fields>
  <field>
    <name>diseaseId</name>
    <type>Integer</type>
    <format/>
    <currency>$</currency>
    <decimal>.</decimal>
    <group>,</group>
```

```
<length>1</length>
<precision>0</precision>
<trim_type>none</trim_type>
</field>
<field>
  <name>diseaseName</name>
  <type>String</type>
  <format/>
  <currency/>
  <decimal/>
  <group/>
  <length>26</length>
  <precision>-1</precision>
  <trim_type>none</trim_type>
</field>
<field>
  <name>Description</name>
  <type>String</type>
  <format/>
  <currency/>
  <decimal/>
  <group/>
  <length>1269</length>
  <precision>-1</precision>
```

```
<trim_type>none</trim_type>
</field>
<field>
  <name>keggLink</name>
  <type>String</type>
  <format/>
  <currency/>
  <decimal/>
  <group/>
  <length>99</length>
  <precision>-1</precision>
  <trim_type>none</trim_type>
</field>
</fields>
<cluster_schema/>
<remotesteps> <input> </input> <output> </output> </remotesteps> <GUI>
  <xloc>322</xloc>
  <yloc>144</yloc>
  <draw>Y</draw>
</GUI>
</step>
</steps>
<order>
```



```
<hop> <from>CSV file input disease</from><to>Table output
disease</to><enabled>Y</enabled> </hop>
```

```
</order>
```

```
</notepads>
```

```
</notepads>
```

```
</transformation-steps>
```

## 2. Populating the node table



```
<?xml version="1.0" encoding="UTF-8"?>
```

```
<transformation-steps>
```

```
<steps>
```

```
<step>
```

```
<name>Add constants</name>
```

```
<type>Constant</type>
```

```
<description/>
```

```
<distributed>Y</distributed>
```

```
<copies>1</copies>
```

```
<partitioning>
```

```
<method>none</method>
```

```
<schema_name/>
```

```
</partitioning>
```

```

<fields>
  <field>
    <name>diseaseId</name>
    <type>Number</type>
    <format/>
    <currency/>
    <decimal/>
    <group/>
    <nullif>1</nullif>
    <length>-1</length>
    <precision>-1</precision>
  </field>
</fields>
<cluster_schema/>
<remotesteps> <input> </input> <output> </output> </remotesteps> <GUI>
  <xloc>281</xloc>
  <yloc>125</yloc>
  <draw>Y</draw>
</GUI>
</step>
<step>
  <name>Add sequence</name>
  <type>Sequence</type>
  <description/>

```

```

<distributed>Y</distributed>

<copies>1</copies>

  <partitioning>

    <method>none</method>

    <schema_name/>

  </partitioning>

<valuenam>nodeId</valuenam>

<use_database>N</use_database>

<connection/>

<schema/>

<seqname>SEQ_</seqname>

<use_counter>Y</use_counter>

<counter_name/>

<start_at>1</start_at>

<increment_by>1</increment_by>

<max_value>999999999</max_value>

<cluster_schema/>

<remotesteps> <input> </input> <output> </output> </remotesteps> <GUI>

  <xloc>411</xloc>

  <yloc>126</yloc>

  <draw>Y</draw>

</GUI>

</step>

<step>

```

```
<name>CSV file input</name>
<type>CsvInput</type>
<description/>
<distributed>Y</distributed>
<copies>1</copies>
  <partitioning>
    <method>none</method>
    <schema_name/>
  </partitioning>
<filename>C:\Users\Rajinder\Desktop\transformation files protas\nslcnodes.csv</filename>
<filename_field/>
<rownum_field/>
<include_filename>N</include_filename>
<separator>,</separator>
<enclosure>&quot;</enclosure>
<header>Y</header>
<buffer_size>50000</buffer_size>
<lazy_conversion>Y</lazy_conversion>
<add_filename_result>N</add_filename_result>
<parallel>N</parallel>
<encoding/>
<fields>
  <field>
    <name>nodeNumber</name>
```

```
<type>Integer</type>
<format/>
<currency>$</currency>
<decimal>.</decimal>
<group>,</group>
<length>2</length>
<precision>0</precision>
<trim_type>none</trim_type>
</field>
<field>
  <name>nodeName</name>
  <type>String</type>
  <format/>
  <currency/>
  <decimal/>
  <group/>
  <length>32</length>
  <precision>-1</precision>
  <trim_type>none</trim_type>
</field>
</fields>
<cluster_schema/>
<remotesteps> <input> </input> <output> </output> </remotesteps> <GUI>
<xloc>122</xloc>
```

<yloc>119</yloc>

<draw>Y</draw>

</GUI>

</step>

<step>

<name>Select values</name>

<type>SelectValues</type>

<description/>

<distributed>Y</distributed>

<copies>1</copies>

<partitioning>

<method>none</method>

<schema\_name/>

</partitioning>

<fields> <field> <name>nodeNumber</name>

<rename>nodeNumber</rename>

<length>-1</length>

<precision>-1</precision>

</field> <field> <name>nodeName</name>

<rename>nodeName</rename>

<length>-1</length>

<precision>-1</precision>

</field> <field> <name>diseaseId</name>

<rename>diseaseId</rename>

```

    <length>-1</length>

    <precision>-1</precision>

</field>    <field>    <name>nodeId</name>

    <rename>nodeId</rename>

    <length>-1</length>

    <precision>-1</precision>

</field>    <select_unspecified>N</select_unspecified>

</fields>    <cluster_schema/>

<remotesteps> <input> </input> <output> </output> </remotesteps> <GUI>

    <xloc>550</xloc>

    <yloc>129</yloc>

    <draw>Y</draw>

</GUI>

</step>

<step>

    <name>Table output</name>

    <type>TableOutput</type>

    <description/>

    <distribute>Y</distribute>

    <copies>1</copies>

    <partitioning>

        <method>none</method>

        <schema_name/>

    </partitioning>

```

```

<connection>Protasstaging</connection>

<schema/>

<table>node</table>

<commit>100</commit>

<truncate>Y</truncate>

<ignore_errors>N</ignore_errors>

<use_batch>Y</use_batch>

<partitioning_enabled>N</partitioning_enabled>

<partitioning_field/>

<partitioning_daily>N</partitioning_daily>

<partitioning_monthly>Y</partitioning_monthly>

<tablename_in_field>N</tablename_in_field>

<tablename_field/>

<tablename_in_table>Y</tablename_in_table>

<return_keys>N</return_keys>

<return_field/>

<cluster_schema/>

<remotesteps> <input> </input> <output> </output> </remotesteps> <GUI>

  <xloc>728</xloc>

  <yloc>130</yloc>

  <draw>Y</draw>

</GUI>

</step>

</steps>

```



<order>

<hop> <from>Add constants</from><to>Add sequence</to><enabled>Y</enabled> </hop>

<hop> <from>Add sequence</from><to>Select values</to><enabled>Y</enabled> </hop>

<hop> <from>CSV file input</from><to>Add constants</to><enabled>Y</enabled> </hop>

<hop> <from>Select values</from><to>Table output</to><enabled>Y</enabled> </hop>

</order>

<notepads>

</notepads>

</transformation-steps>


## Appendix D –MySQL commands for search option at GUI

1. `SELECT d.diseaseName, d.description, d.keggLink, c.zscore, c.pvalue, c.sp, c.adjMatrix from disease d, cal_values c where d.diseaseId = '$disease' && d.diseaseId=c.diseaseId;`
2. `create or replace view vertical_sum as select adjMatrix, count(frequency) as totalnodes, sum(frequency) as totalFreq, diseaseId from freq where frequency != 0 && diseaseId = '$disease' group by adjMatrix;`
3. `create or replace view min_adjMatrix as select adjMatrix from vertical_sum order by min(totalNodes/totalFreq) asc limit 1;`
4. `create or replace view significantnodes as select f.nodeId, f.diseaseId, m.adjMatrix from freq f, min_adjMatrix m where f.diseaseId= '$disease' && f.frequency!=0 && m.adjMatrix=f.adjMatrix group by nodeId;`
5. `create or replace view significantnodes1 as select n.nodeName, n.NodeNumber, sum(f.frequency) as totalFreq from node n, freq f, significantnodes s where f.diseaseId= '$disease' && s.nodeId=f.nodeId && s.nodeId=n.nodeId group by n.nodeName order by totalFreq desc;`
6. `create or replace view significantnodes2 as select sum(s1.totalFreq) as totalsum from significantnodes1 s1 order by totalsum desc;`
7. `create or replace view significantnodes3 as SELECT s1.nodeName, s1.NodeNumber, (((s1.totalFreq)/s2.totalsum)*100) as threshold FROM significantnodes1 s1, significantnodes2 s2;`

8. `SELECT * FROM significantnodes3 s where threshold > '$value';`

## References

1. "Cancer Facts and Figures". American Cancer Society, 2013.
2. "José Costa". Internet:  
<http://www.britannica.com/EBchecked/topic/92230/cancer/224705/Nomenclature-of-malignant-tumours>, Apr. 23, 2014 [May 2, 2014].
3. "Cancer Facts & Figures 2014". American Cancer Society Atlanta, 2014.
4. "Latest world cancer statistics Global cancer burden rises to 14.1 million new cases in 2012". December 2013
5. R. Dikshit, P.C. Gupta, C. Ramasundarahettige, V. Gajalakshmi, L. Aleksandrowicz, R. Badwe, R. Kumar, S. Roy, W. Suraweera, F. Bray, M. Mallath, P.K. Singh, D.N. Sinha, A.S. Shet, H. Gelband, P. Jha. "Cancer mortality in India: a nationally representative survey". *The Lancet*, 379(9828), pp. 1807-1816, 2012.
6. N. Weiner. "Cybernetics or Control and Communication in the Animal and the Machine". MIT Press, Cambridge, MA, 1948.
7. K. Hiroaki. "Systems biology: a brief overview." *Science*, 295.5560, pp. 1662-1664, 2002.
8. Wong, Elisabeth, B. Baur, S. Quader, C.H. Huang. "Biological network motif detection: principles and practice." *Briefings in bioinformatics* 13:2, pp. 202-215, 2012.
9. Chen, Jin, W. Hsu, M. L. Lee, S.K. Ng. "NeMoFinder: Dissecting genome-wide protein-protein interactions with meso-scale network motifs." In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 106-115. ACM, 2006.
10. Omidi, Saeed, F. Schreiber, A.M. Nejad. "MODA: an efficient algorithm for network motif discovery in biological networks." *Genes & genetic systems* 84 no. 5, 2009.
11. Kashani, Zahra RM, H. Ahrabian, E. Elahi, A.N. Dalini, Elnaz, S. Ansari, S. Asadi, S. Mohammadi, F. Schreiber, A.M. Nejad. "Kavosh: a new algorithm for finding network motifs." *BMC bioinformatics* 10 no. 1, pp. 318, 2009.
12. Kreher, L. Donald, D.R. Stinson. "Combinatorial algorithms: generation, enumeration, and search." CRC press, Vol. 7, 1998.
13. B. McKay. "The NAUTY." Internet: <http://cs.anu.edu.au/bdm/nauty>, [Mar. 15, 2014].

14. Schreiber, Falk, H. Schwöbbermeyer. "MAVisto: a tool for the exploration of network motifs." *Bioinformatics* 21, no. 17, pp. 3572-3574, 2005.
15. Schreiber, Falk, H. Schwöbbermeyer. "Frequency concepts and pattern detection for the analysis of motifs in networks." In *Transactions on computational systems biology III*, pp. 89-104. Springer Berlin Heidelberg, 2005.
16. N. Kashtan, S. Itzkovitz, R. Milo, "Network motif detection tool Mfinder tool guide. Technical report 2005". Departments of Molecular Cell Biology and Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel, 2005.
17. Kashtan, Nadav, S. Itzkovitz, R. Milo, U. Alon. "Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs." *Bioinformatics* 20, no. 11, pp. 1746-1758, 2004.
18. Wernicke, Sebastian, F. Rasche. "FANMOD: a tool for fast network motif detection." *Bioinformatics* 22, no. 9, pp. 1152-1153, 2006.
19. Barabási, A. László, and R. Albert. "Emergence of scaling in random networks." *Science* 286, no. 5439, pp. 509-512, 1999.
20. Schreiber, Falk, H. Schwöbbermeyer. "MAVisto: a tool for the exploration of network motifs." *Bioinformatics* 21, no. 17, pp. 3572-3574, 2005.
21. Chen, Jin, W. Hsu, M. L. Lee, S.K. Ng. "NeMoFinder: Dissecting genome-wide protein-protein interactions with meso-scale network motifs." In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 106-115. ACM, 2006.
22. N. Kashtan, S. Itzkovitz, R. Milo, "Network motif detection tool Mfinder tool guide. Technical report 2005". Departments of Molecular Cell Biology and Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel, 2005.
23. S. Wernicke. "A faster algorithm for detecting network motifs." *Algorithms Bioinformatics*, 3692, pp. 165-77, 2005.
24. J.A. Grochow, M. Kellis, "Network motif discovery using sub-graph enumeration and symmetry-breaking." *Res Computational Molecular Biology*, 4456, pp. 92-106, 2007.
25. McKay, D. Brendan. "Isomorph-free exhaustive generation." *Journal of Algorithms* 26, no. 2, pp. 306-324, 1998. 
26. Uetz, Peter, L. Giot, G. Cagney, T.A. Mansfield, R.S. Judson, J.R. Knight, D. Lockshon et al. "A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*." *Nature* 403, no. 6770, pp. 623-627, 2000.

27. T. Lepisto, A. Salomaa, A. Lingas, et al." A polynomial-time algorithm for sub-graph isomorphism of two-connected series-parallel graphs." In: Proceedings of the 15th International Colloquium on Automata, Languages and Programming (ICALP) 1988, Tampere, Finland, pp. 394–409.
28. <http://www.genome.jp/kegg/> [Aug. 5, 2013].
29. Kanehisa, Minoru, S. Goto, M. Hattori, K.F.A. Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, M. Hirakawa. "From genomics to chemical genomics: new developments in KEGG." *Nucleic acids research* 34, no. suppl 1, pp. D354-D357, 2006.
30. Wernicke, Sebastian. "A faster algorithm for detecting network motifs." In *Algorithms in Bioinformatics*, pp. 165-177. Springer Berlin Heidelberg, 2005.

## Posters

Poster presented on “**Deciphering Biological networks: Clues for cure**” in **Biorhythm**, Feb 2014 under the aegis of **DBT. (Bagged 1<sup>st</sup> Runner’s Up)**

Poster presented on “**Annotation of Biological Networks through Network Motifs using Top Down approach**” in **3<sup>rd</sup> IFIP International Conference of Bioinformatics**, Sep 2013.

Poster presented on “**Identification and analysis of network motifs in human disease specific pathways applying top down Systems Biology approach**” in Virtual Conference, **Bioinformatics to Systems Biology**, Oct 2013.

## Resume

### EDUCATIONAL QUALIFICATION

College/University/School	Degree	Marks Obtained	Duration
Jaypee University of Information Technology, Solan, Himachal Pradesh, India	Masters in Technology Computational Biology	8.7 out of 10 CGPA	July, 2012 - May, 2014
Dr. D.Y. Patil Biotechnology and Bioinformatics Institute, Pune, Maharashtra, India	Bachelors in Technology Biotechnology	7.4 out of 10 CGPA	August, 2007 - June, 2011
Co-operative Public School, Jammu, Jammu & Kashmir, India	HSC (Physics + Chemistry + Mathematics + Biology)	62.7%	Apr, 2006 - Mar, 2007
Co-operative Public School, Jammu, Jammu & Kashmir, India	SSC (All Subjects)	84.8%	Apr, 2004 - May, 2005

### ACADEMIC PROJECTS

#### M.TECH PROJECT

<b>Major-Project:</b> Top down approach to annotate heterogeneous biological networks through network motifs.	July,2013 to May,2014	Jaypee University of Information Technology, Solan, Himachal Pradesh, India.	Studying the biological pathways in terms of networks and elucidating the significance of certain patterns and entities in the pathway and finally developed a GUI and datawarehouse for the results and inferences generated.
<b>Minor Project:</b> Develop a datawarehouse on any biologically important entity.	Oct,2012 to Dec,2013	Jaypee University of Information Technology, Solan, Himachal Pradesh, India.	A datawarehouse was developed for Insulin providing a better look into different structure variations, polymorphism, availability, regulation etc Tools used were MySQL server, Pentaho KETTLE and CA Erwin.

#### B.TECH PROJECT

<b>Major-Project:</b> Molecular insight in modulation of immune responses with plant based fractions.	Dec,2010 to May,2011	Indian Institute of Integrative Medicine (CSIR), J&K, India.	Testing the immunomodulatory activity of <i>Tanacetum gracile</i> plant extracts on the mice's immune system (in vivo & in vitro), parameters measured: Haemagglutination Titer, Delayed Type Hypersensitivity response, Lymphocyte proliferation assays and Macrophage function assays.
<b>Minor Project:</b> Preclinical methodology to evaluate immune response in mice.	Jun,2008 to Jul,2008	Indian Institute of Integrative Medicine (CSIR), J&K, India.	In-vivo evaluation of immunomodulatory activity of lead molecules.



### **Poster in Conference(s)**

<b>Title</b>	<b>Venue</b>	<b>Organizing Body</b>	<b>Date</b>	<b>Result</b>
Deciphering Biological networks: Clues for cure.	<b>Department of Bioinformatics, GGSDS College, Chandigarh, India.</b>	Department of Biotechnology (DBT), India.	Feb 2014	1 <sup>st</sup> Runner's Up
Annotation of Biological Networks through Network Motifs using Top Down approach.	MANIT Bhopal, Madhya Pradesh, India.	3 <sup>rd</sup> IFIP International Conference of Bioinformatics	Sep 2013	Participation
Identification and analysis of network motifs in human disease specific pathways applying top down Systems Biology approach.	Virtual Conference	BSB '13 Bioinformatics	Nov 2013	Participation