NTNU – Trondheim
Norwegian University of
Science and Technology

Analysis and comparision of Transcription Factor - Target Gene databases using network analysis.

The gene regulatory network documented on the basis of experimental evidence is analysed and compared by the graph based analysis method of various resources of gene

Pratik Bikram Poudel

Biotechnology
Submission date: August 2014
Supervisor: Martin Tremen R. Kuiper, IBI

Norwegian University of Science and Technology
Department of Biology

Pratik Bikram Poudel

Analysis and comparison of Transcription Factor – Target Gene databases using network analysis.

Master's thesis in Biotechnology

Trondheim, August 2014

Supervisor: Martin Kuiper

Norwegian University of Science and Technology
Faculty of Natural Science and Technology
Department of Biotechnology

Preface:

This Master thesis project on network science is performed under the Department of Biotechnology in the Norwegian University of Science and Technology (NTNU). I sincerely vow gratitude to the Department for providing me the earnest opportunity to accomplish my Master level studies.

Beside a successful student there is always an inspirational mentor to let him unveil his full potential. I am inspired by my Supervisor Prof. Martin Tremen R. Kuiper, during every thick and thin on my stint as his disciple. Those constructive suggestions and criticism as well, were the driving force for my performance. Without his proper guidance and timely emails, it would have been a daunting task for me.

I remember my family, my mother, my father and brothers for their everlasting support and caress, which motivated me to do something for us.

And my friend Khusboo, who remain with me as a strong support throughout, is also one who cannot remain unmentioned in my acknowledgements.

Abstract:

The transcriptional regulatory network is formed by interaction between nodes and edges. The nodes are transcription factors and target genes whereas, edges are the interaction between them. It can be binding technique, cell-type or species as an interaction type. The transcription factor binds to the cis regulatory site of the DNA, thus enabling further processing through the bottom up approach model of dealing with network, is a way to elucidate the related information from the whole system.

The different four dataset were taken to compare, and analyze that network with each other. Thus already curated database of Tf-Tg relationship is studied and analyzed. The database is taken as the working field. The several parameters like node degree, clustering coefficient and sub networks are the working tools. The various tools are used for the network analysis; and an open source software platform called Cytoscape is used for purpose of visualizing different molecular interaction in networks and biological pathways and later integrating these networks with the gene annotation data and gene expression profiles. Basically the plugins of the Cytoscape enables it to perform the wide range of analysis.

After the analysis of the networks, with the assistance of different Plugins like Network analyzer, BiNGO, MiMI and MCODE, different shades of a complex network is unravelled and explored. The node degree distribution of different nodes namely transcription factor and target gene was done. The distribution provided different types of network models like Scale free, Random, small world and hierarchical network model. The clustering coefficient is found to decrease with increasing value of clustering coefficient. Similarly, several biological processes were linked to the sub networks, since the genes involved in those sub networks, hold a regulatory roles in those processes.

Through this way of network analysis by Cytoscape and its various plugins, a researcher can estimate and infer from the topological analysis of biological networks, which is very relevant to the field of current bioinformatics and systems biology.

List of abbreviations

DNA	De-oxy ribonucleic acid
RNA	Ribonucleic acid
TRRD	Transcription regulatory Regions Database
TRED	Transcription regulatory element database
YeaSTRACT	Yeast Search for Transcriptional Regulators And consensus Tracking
ChIP	Chromatin Immunoprecipitation
STKE	Signal transduction Knowledge Environment
ChIP-chip	Chromatin Immunoprecipitation coupled with microarray
ChIP-Seq	Chromatin Immunoprecipitation coupled with deep sequencing
HTRIdb	Human Transcriptional Regulation Interaction database.
EMSA	Electrophoretic mobility shift assay
SPR	Surface Plasmon Resonance
SDS-PAGE	Sodium dodecyl sulphate- Polyacrylamide gel electrophoresis
ESCAPE	Embryonic Stem Cell Atlas of Pluripotency Evidence
ChIP-PET	ChIP Paired en Ditags
MiMI	Michigan Molecular Interaction
BiNGO	Biological Network gene ontology tool
Tf	Transcription factors
Tg	Target genes
FFL	Feed forward loop
NAR	Negative auto regulation
PAR	Positive auto regulation
DOR	Dense overlap Regulons
MIMs	Multi input motifs

List of Figures

List of Tables

Table of Contents

1 Introduction

The transcriptional regulatory network describes the association of regulatory proteins with genes in the genome. The expression of those genes is characterized by the recognition of specific promoter sequence with the help of transcriptional regulatory proteins (Orphanides, Lagrange et al. 1996). One of the most important transcriptional regulatory protein is the DNA binding transcription factor which binds with the specific *cis* regulatory element of the gene's promoter that mediates activation or repression in the process of RNA polymerase II dependent process of transcription. The effect of association of regulatory protein with genes across genome is illustrated by the transcriptional regulatory network.

The network model to represent the transcriptional regulatory network can be represented as directed graphs which consist of `nodes` to represent the bioactive macromolecules like Proteins, DNA, RNA and metabolites and `edges` represents regulatory interaction between them(Babu, Luscombe et al. 2004, Barabasi and Oltvai 2004). There are two types of regulatory networks, namely transcription regulatory network and post-transcription regulatory network. Both networks can be subdivided into the physical and functional networks. The different macromolecular interaction like Protein-Protein, Protein-DNA, Protein-RNA and RNA-RNA interactions form the physical network than can be observed. On the other hand, the Functional network is the consequence of these physical interactions i.e. activation or repression of gene expression(Walhout 2006). The cumulative picture of both transcription regulatory network and post transcription regulatory is prerequisite for the comprehensive understanding on every aspect of differential gene expression in complex system.

Analyzing the transcriptional regulatory dataset helps to determine and obtain the information from the structure of the network by calculating different network topological parameters. The interaction between the target gene and transcription factor in the network provides important understandings of the biological processes. However the interacting elements of system is then represented by Network graphs (Bollobás and Cockayne 1979).

1.1 Project goal

The goal of the project is to compare public resources database containing datasets on relationship between transcription factors and target genes and perform analysis with the help of graph analytical methods. The expectation is that the graph metrics like node degree and clustering coefficient would be useful to identify global differences between datasets, which

then should be explained by considering the origin from of the different datasets. The various datasets are used for the purpose of analysis. The dataset retrieved from HTRIdb webpage has the information about transcriptional regulation between the Tf and Tg of verified by different techniques like Electrophoretic mobility shift assay(EMSA), Chromatin immunoprecipitation(ChIP), Chromatin immunoprecipitation followed by deep sequencing(ChIP-seq) and many more. Out of four different resource of dataset, I downloaded the catalogue file from TFactS datasets which has both sign-less and sign-sensitive format. In Sign-less dataset, we obtained relationship between Tf and Tg through different knowledge base resource like PubMed, TRRD, TRED etc. The observations were taken for Human, Mus musculus, Rattus norvegicus and Zebra fish and another sign-sensitive dataset from similar knowledge base resource gives the additional information about Up and Down regulation of Tf-Tg interaction. The third dataset was retrieved from YeaSTRACT database, which is the documented regulation between all transcription factor and genes described in YeaSTRACT database. Finally I took last dataset from Embryonic Stem Cell Atlas of Pluripotency Evidence webpage also, I retrieved the Protein/DNA interaction table extracted form ChIP-X studies under Ma`ayan laboratories.

Emphasis will be to derive the similarity and difference between them with the help of topological approach to understand the network functionality through different parameters like node degree and clustering coefficient. The likelihood of appearance of noisy data in ChIP sequence based at Tf-Tg relationship of the network is estimated to suggest the inflated connectivity in Tf-Tg network. The further sub network analysis of the motifs in the network is performed to determine whether the motifs can assist us in understanding the densely connected network of higher organisms.

The result of the project is expected to indicate potential inconsistencies in the data and explain these by looking carefully at how these datasets have been built and provide the proper guideline for the future transcription regulation related datasets formulation to unravel more useful information.

1.2 Background

Eukaryotic gene regulation consists of transcriptional and post transcriptional regulation. Transcriptional regulation is a process that is processed by the collective action of sequence specific DNA binding transcription factors along with core RNA pol II transcriptional machinery, also the involvement of co regulators that bridge the DNA binding factors to the transcriptional machinery, a number of chromatin remodeling factors that mobilize nucleosomes and variety of enzymes that catalyze covalent modification of histones with other proteins(http://www.blackwellpublishing.com/allison/docs/sample_ch11.pdf). Activation or repression of gene transcription result from the binding of transcription factors to the promoter and enhancers of genes, and affect subsequent gene expression at transcription levels (Chen and Rajewsky 2007). At the post transcriptional level micro RNAs have an active role in the repression of expressed genes by inhibiting or degrading the translation of their target mRNAs. Transcription factor and miRNAs acts by recognizing cis-regulatory region of the DNA and RNA respectively; hence they both are the trans-acting elements. The proper coordination of transcription factors and miRNAs assures the efficient and accurate gene expression (Cheng, Yan et al. 2011).

1.3 Transcription regulation

Transcription is the process of gene expression when a gene's DNA sequence is used as a template and transcribed into mRNA by an RNA polymerase. It is a very complex process. The detail description of the process enables us to understand the role of chromatin which results from the packaging of eukaryotic DNA. The importance of the regulation of transcription can be evaluated more, when it is understood that a considerable part of the coding capacity of the genome is devoted to this regulation, i.e. some 10 % of the genes in mammals/humans encode transcription factors.

1.3.1 Chromatin remodeling

Chromatin remodeling is a vital modification in the chromatin architecture, which allows the regulatory transcription machinery proteins to 'find' their genes and thereby control the gene expression. Initiation of transcription in chromatin template which is open requires promoter bound to enzyme RNA polymerase and transcription factor to enhancers. The open chromatin structure refers to the nucleosome octamer being removed from the promoter before RNA polymerase can bind. The modifications in chromatin structure are an important factor for controlling transcription in eukaryotic cells. The activation of gene structure is through the acetylation of chromatin structure to make it available. However the chromatin remodeling is

accomplished through the process like a) histone modification through specific enzymes, like histone acetyl transferase, deacetylase, methyl transferase and kinases and ATP dependent chromatin remodeling complex that move, eject or restructure nucleosomes.

1.3.2 Initiation of transcription

The specific protein called transcription factor is needed for RNA polymerase II to initiate transcription. In addition, other gene specific regulators also bind to the DNA sequences to control the expression of individual genes and involvement in regulation of gene expression. The initiation of transcription is by the binding of the transcriptions factor along with several other regulatory proteins to the DNA. The binding of the regulatory protein to the *cis* - regulatory sequence of DNA controls the transcription of adjacent genes. Thus, the *cis* acting sequence is cardinal for the expression of the eukaryotic genes. It serves as the binding site for the wide variety of regulatory proteins that controls the expression of individual genes.

1.3.3 Pre initiation complex formation

The important phase in the transcription initiation is the pre initiation complex formation on promoters and proper positioning of RNA polymerase at the site. The recognition of promoter in the sequence is recognized by large number of interacting proteins called general transcription factors. It along with RNA polymerase and mediator complex forms the pre initiation complex and has the cardinal role in the transcription initiation/ activation. The complex plays productive role for the regulation of RNA polymerase II functioning. The general transcription factors which are important factor for transcription initiation by RNA II polymerase are highly conserved between species. They enable RNA polymerase II to recognize promoters.

a) TFIID

   It is major transcription factor within preinitiation complex binds with TBP (TATA binding protein).

b) TFIIA

   It is one among the important transcription factor which binds immediately to the upstream of TATA box. It has contribution to the stability of TFIID at promoter.

c) TFIIB

   This transcription factor is found at upstream of initiator element. They are important for proper setting of DNA direction in the preinitiation complex.

d) TFIIF

TFIIF plays a significant role in preinitiation complex formation and is prerequisite for initiation and start site selection.

e) TFIIE

TFIIE is the transcription factor which is basically involved with late events of transcription initiation and formation of preinitiation complex.

f) TFIIH

The TFIIH along with TFIIE arrives at the later end of preinitiation complex formation and are involved in the formation of open promoter structure before elongation starts.

And

- Activators

These are the proteins that bind to the enhancers and silencers.

- Co-activators

These proteins are required for activator mediated transcription simulations.

- Mediator complex

Mediator is a very important constituent which share the function as major coactivator to transmit the diverse regulatory signals from gene specific DNA binding transcription factor to general transcription factor at the promoter.

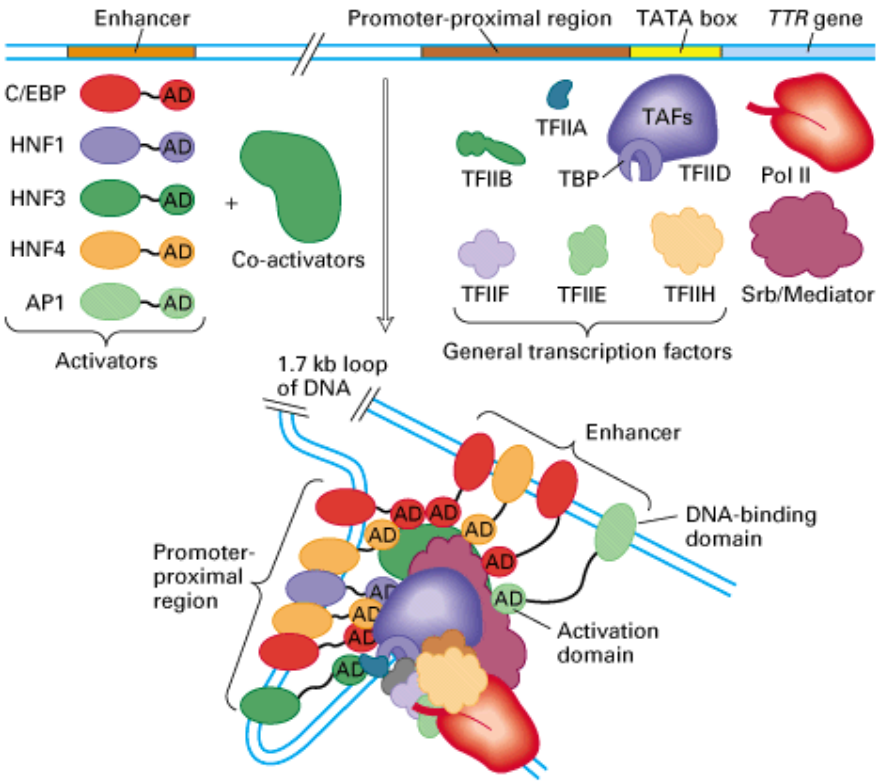


Figure 1Preinitiation complex formation. Reprinted form source without permission.

(Source:http://xray.bmc.uu.se/lars/Practicals/Transcription/Preinit.html)

The promoter of the gene to be transcribed by RNA polymerase II contains several other sequence elements surrounding the transcription site. The sequence similar to TATAA located 25 to 30 nucleotides upstream of transcription is called TATA box. According to the inference drawn from prokaryotic transcription, initially it was assumed every core promoter contain TATA box, however it was later found that only 10-20 % of mammalian core promoter carry a functional TATA box. Thus the alternative for TATA box was found in Initiator (Inr.) element. Initiator element directs the formation of preinitiation complex, by locating the start site for general transcription factor. The sequence elements in the promoter of genes transcribed by RNA polymerase II includes the initiator (Inr) elements in transcription start site (TSS) , TFIIB recognition elements (BRE), its location 35 nucleotide upstream of TSS. However it is apparent that the promoter of different genes has different combination of these core promoter elements, to bind general transcription factors. The subsequent binding of general transcription factor TFIID (TF is transcription factor and IID is polymerase II) to the promoter is the initial step in the formation of transcription factor (REINBERG and Zawel 1993). TFIID is composed of multiple subunits including TATA binding protein (TBP) and other 14 polypeptides called, TBP associated factors (TAFs). In the formation of transcription complex TBT binds to TATA box mostly, while other subunits of TFIID (TAFs) binds to Inr. Then the bindings of TFIID is followed by involvement of second transcription factor (TFIIB), which binds to TBP as well as BRE sequences. TFIIB primarily serves in forming the bridge to RNA polymerase II, which ultimately binds to the polymerase. The RNA polymerase II binds to the TBT –TFIIB complex in association with third factor, TFIIF. At last after the addition of TFIIE and TFIIH to the complex formed, completes the formation of pre initiation complex(Lewin, Krebs et al. 2011).

Then specific transcription factor binding is followed after the series of bending effect in DNA. The TBP binds to the minor groove of DNA which is bent by 80°, and the TATA box is on the major groove. The deformation, hence observed is not meant for separation of DNA, since the base pairing is maintained. The TATA sequence is related to the extent on which the DNA should be bent, which directly correlates to the efficiency of promoter. This structural change has the functional implications, since this change allows the specific transcription factors to form closer proximity with RNA polymerase(Orphanides, Lagrange et al. 1996).

1.4 Bottom-up approach

However, the understanding of a complex system demands bottom-up approaches, as it is dealing with the constituents of the complex system, the elementary and small members of the system and their mapped out interaction between these members of the system(Wuchty, Ravasz et al. 2006). Bottom up approach is the progressing from small units to larger size as an important member of any organization or system. This forms the complexity pyramid of any system.

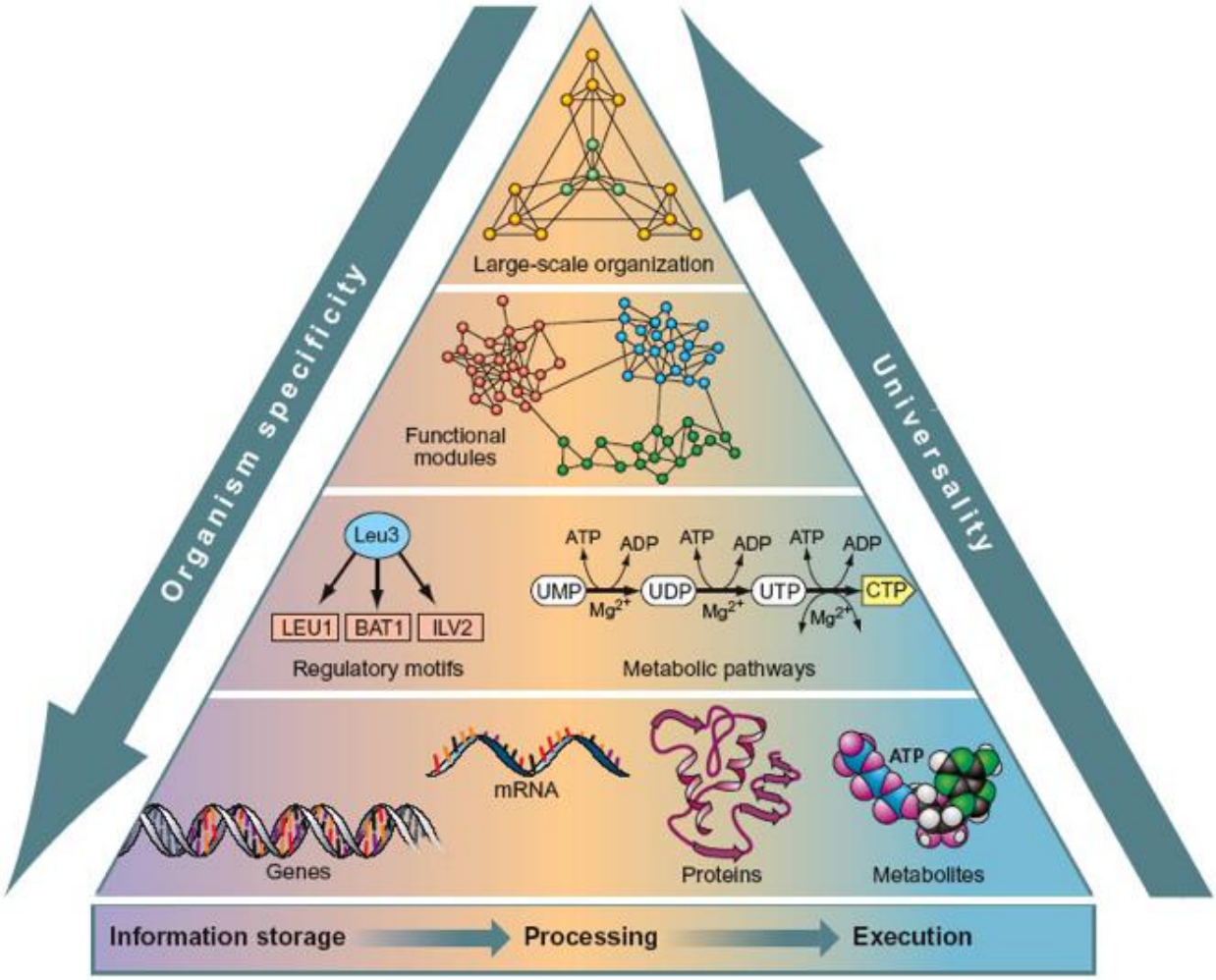


Figure 2The bottom up approach. Reprinted from Source(Wuchty, Ravasz et al. 2006).

The level 1 of the pyramid shows the schematic representation of the cell´s functional organization: genome, transcriptome, proteome and metabolome. The level 2 is the functional components linked by some functional relationships, such as regulatory motifs and metabolic pathways. The level 3 are the operational building blocks forming modules and the last level is a nested and complex binding forming a scale free hierarchical architecture. This is the

basic organizing principle behind every biological complex network(Wuchty, Ravasz et al. 2006).

At the bottom level the interaction between target gene and Transcription factor enables one to have understood basic level of the biological processes. The preliminary objective of system biology is to fuse and analyze the diverse molecular, cellular, tissue level and higher level data source to deduce how subsystems and whole organisms work from this network of interactions. (Shih and Parthasarathy 2012). Gene repertoire and regulatory apparatus perform dynamic interaction at several levels. At the genomic level, transcription factor can activate or inhibit the transcription of genes in presence of various regulators. However, these transcription factors are themselves the product of genes, the ultimate effect is that genes regulate each other's expression as part of gene regulatory networks. The elements of system that do interact or regulate each other can be represented in graph for the understanding of the particular network formed by the nodes and edges, where nodes represent genes / protein and edges represent interaction between them (Bollobás and Cockayne 1979).

The overall network is meant to be a complex and multilayered that can be revealed by the four stratum of evaluation to the network. The basic level is the collection of transcription factors, targets genes and binding sites of the DNA. In another level, the term motif is introduced to study the network. It is the organized and statistically significant pattern or sub graph of interaction. The motif is found in appreciable number in the network to reflect a deep insight to the network functional abilities. The third stratum is of modules, which are formed by the clusters of motifs as semi-independent transcriptional unit. Final level is the overall regulatory network allocating all the interconnecting interactions among the modules to build the entire network (Babu, Luscombe et al. 2004).

"Always the structure affects the function" (Strogatz 2001), the structural parameters of the network theory are basis for understanding cells internal organization and interrelationship between different nodes. After the study of topological structure of network in different species like *Escherichia coli*, *Saccharomyces cerevisae*, the developmental transcription network of Drosophila melanogaster, signal transduction Knowledge environment network (STKE) and neural connection map of *C. elegans* , various analysis performed finally led to the useful ideas relating structures to their function(Prill, Iglesias et al. 2005).

1.5 Network metrics

The consistent integration of the various processes that generate mass, energy, information transfer and cell fate specification in a cell or microorganism can be represented in the form of complex network of cellular constitutions. These network interaction discloses the fundamental principle of its design, and thereby provides the common underlying structure and function in all cell and microorganisms (Ng, Wei et al. 2005). It is understood that the robustness of the cellular dynamics in an organism is correlated with interactions of constituents like Protein, DNA, RNA and small molecules. The amount of information generated by data production from several scientific developments like large scale sequencing project enriches the field of network study. For example the complete genome sequence information of organisms provides the platform for integrated pathway/genome databases which provide the organism specific connectivity map for metabolic networks. The map of these processes are extremely complex to understand; that is why only limited insights can be obtained about the organizational principle of these complex systems. Therefore networks formed by this complex system are dealt with the quantitative approach. The better understanding about the generic properties of the complex network is possible to gain by dealing those networks in the language of topological parameters of Network science.

According to the United States Research Council, ``Network science is the study of network representation of physical, biological, and social phenomena leading to the predictive models of these phenomena (Lederman and Morikis 2014).`` Network science has an approach of simplifying complex systems, and summarizing them as components (nodes) and interaction (edges). In cellular system the nodes are the macromolecules like protein, gene sequence and RNA molecules, whereas the edges are the physical, biochemical and functional interactions observed by the various technologies. The nodes of the built network loses its functional richness, despite this loss of functional value of these nodes - many useful discoveries can be observed by mapping the interaction between those nodes, which are representing functional components, and the interactions. This is performed using systematic and standardized approaches and assays. The main purpose behind this network evaluation by the precise modeling of those nodes at the scale of whole cells is to gain the information about the pattern of development.

Thus the interactome data for the network has three distinct approaches to create, they are listed below:

1) The first approach is by Curation or compilation of existing data from already available literature, under the information retrieval system which is a process to retrieve a document that a user my finds useful to his her information need. It is usually observed by one or just a few types of physical or biochemical interaction.

2) Another approach is of computational approach of prediction which is based on the orthogonal information available, such as sequence similarity, gene order conservation, copresence and co absence of genes in completely sequenced genome and protein structural information. Therefore the approach of integrating this large data set of available gene expression data and sequence data to gather some inferences about the function of uncharacterized genes is done by this process. (Spencer, Sun et al. 2003)

3) High throughput experimental mapping strategies applied at the scale of whole genome or proteomes is also an approach, which is unbiased and systematic in nature.

The different cellular level of interaction is shown in the figure below:

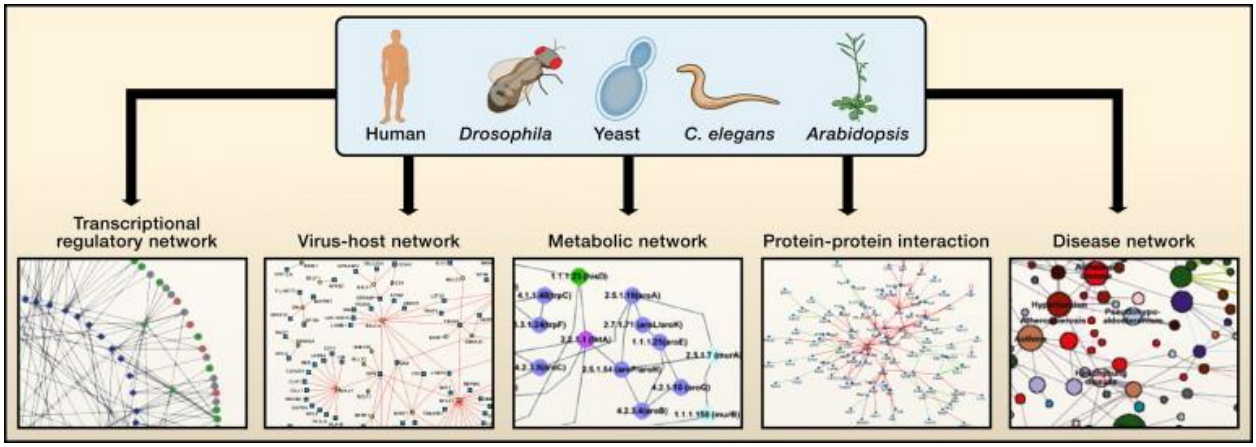


Figure 3The different networks at the cellular level where interactome of cellular systems(Vidal, Cusick et al.)

The interaction on the system is principally represented as networks of nodes and edges. Extensive networks, however, are only made for some species, including yeast, worm, fly, human and plant. In case of a protein interaction network, proteins represent nodes and physical interactions between them are edges. In transcriptional regulatory network, nodes can be either transcription factors or gene sequences, and edges represent the physical bond between them. The disease network has different disease as the nodes and the gene mutations which are associated with the linked disease is termed edge. In viral host network, viral proteins and host protein are the nodes and the edges represent the interaction between them. In metabolic network, the node represents the enzymes, and edges represent the metabolites that are the product or substrates of the enzymes(Vidal, Cusick et al. 2011)

1.5.1 Node degree (K)

Degree of a node is the number of adjacencies of that node in the network, which is the number of nodes to which focal node is connected to (Shen-Orr, Milo et al. 2002).Node may be the distribution or end point of any data transmissions. In nature, nodes are programmed or abled with capability to recognize and process or forward transmission to other nodes. The basic unit of indicator to characterize a network is the degree of the nodes involved.

$$K_i = C_{D(i)} = \sum_j^N x_{ij} \ldots\ldots\ldots\text{(i)}$$

Where i is the focal node, j represent all adjacent nodes, N is the total number of nodes, and $X$is the adjacency matrix, in which $X_{ij}$is defined as 1 if node i connected to j, and 0 other wise(Shen-Orr, Milo et al. 2002).

As we know that elementary characteristic of the node is its degree. Node degree signifies the number of links a nodes has to other nodes. In undirected network, the degree is calculated on the basis of number of edges adjacent on that node. In directed network, the functionality of the node degree depends on the direction. The Incoming nodes and Out-going nodes the nodes may be directed towards the node. If a node is directed outward then it has Out-going degree (K_{out}). Similarly in coming nodes is estimated as incoming nodes (K_{in}). The number of neighbors of a node is called the connectivity of particular node. In undirected network, the average degree ($< K >$) of nodes for the whole network N nodes and L links is calculated as average degree$< K >$ =$2L/N$ 2L/N. As already mentioned, in every directed network, each link has a selected direction, incoming (K_{in}) and Out-going (K_{out}).

1.5.2 Node degree distribution

The node degree distribution $p(K)$ is the probability distribution of these degrees over the whole network. It gives degree distribution for the every node (n) in network with having some degree (K), (nK) is the number of nodes at a particular degree, whereas (n) is the total number of nodes in that network.

Degree distribution is calculated as $p(K) = nK/n$

The degree distribution pattern differentiates different types of networks.

1.5.3 Clustering coefficient (C)

It is the measure of degree of interconnectivity in the neighborhood of a particular node. In a network to form a cluster, neighbors of the node are connected to each other. In this case, the clustering coefficient is 1. The value of clustering coefficient ranges from 0 to 1.

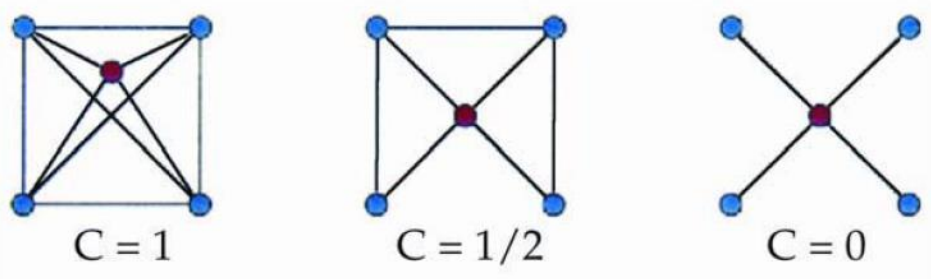


Figure 4Different clustering coefficient values in three different cases. The Figuree is reprinted from (Wuchty, Ravasz et al. 2006).

For example red node which has its neighbor connected to each other has value 1(left) and the red which has its neighbor but not connected to each other has the clustering coefficient value 0.

There are two approaches of defining and calculating the clustering coefficient: One is through the actual no of links in the neighborhood of a node and through the number of triangles in the network.

General clustering coefficient for a node A in a undirected network is calculated as

$C_A = 2n_A/K_A(K_A - 1)$……(i)

In case of directed network it is

$C_A = n_A/K_A(k_{A-1})$……(ii)

where n_A is the number of links connecting the neighbors of node A to each other.

In both the cases, the clustering coefficient gives the ratio of N/M, in which N is the number of edges between the neighbors of n, and M is the maximum number of edges that could possibly exist between the neighbors of n.And the network clustering coefficient C_A is the average of the clustering coefficient for all the nodes in the network. Therefore definition for the network clustering coefficient is in network A is

$C_A = 3 * Number\ of\ trianlges\ in\ A\ /$
$No\ of\ Connected\ node\ triplets\ in\ A)$ ….……..(iii)

The average clustering coefficient $< C >$ is the measurement of overall tendency of nodes to form clusters or groups thus forming the network structure. If $C(K)$is independent of K, the network is dominated by the numerous small tightly linked clusters. The network formed has the small sub networks within.

1.6 Network Models:

The various systems in world are relatively meaningful, if we carry out some study on it. The nature of study can be led by the measure of information that can be obtained. Some study the individual components of the system – like how a mobile phone works or how a big Airbus manages to cover the thousands of miles with that huge size. Next study can be on the connection between the phones or the interaction between the traffic towers to the Air bus to run it properly. However the much cardinal aspect of study is left behind most of the time. It the pattern of those interactions that fetches the important behaviors of the system which is studied between different phone or the airbuses. Thus the pattern of connection is studied by forming networks between, of the cellular system or the Airbus traffic management system(Ng, Wei et al. 2005)

Thus the pattern of the connection between the provided systems is studied via network. The characteristic of the dataset can be derived as a network. The obtained network model facilitates the explanation for the emergence and behaviors of the important network characteristics. The network is the very useful in understanding the complex networks. There are several models of the networks depending on the design of its formation. The node and edge distribution in the network differentiates them.

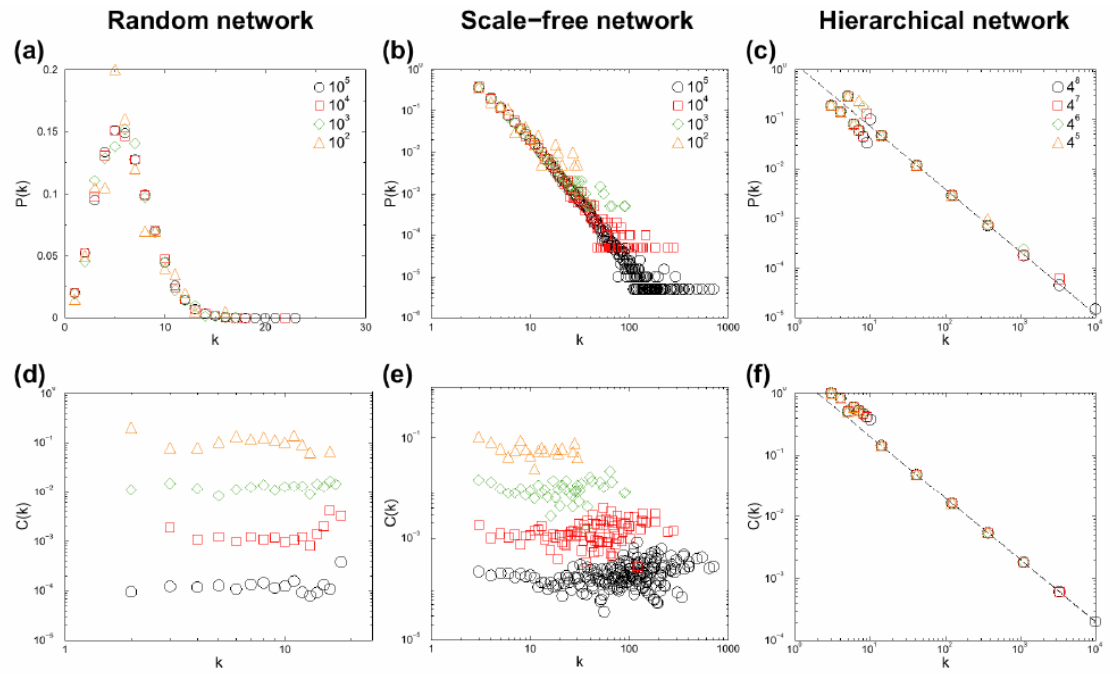


Figure 5The different properties of three network models.The image is reprinted from the source(Wuchty, Ravasz et al. 2006)

(a) The ER model follows the Poisson degree distribution$p(K)$, it is the probability of a randomly selected node having degree range within average degree$< K >$ of the network has the highest peak. (d) The clustering coefficient distribution is found independent in the random graph as well in the scale free network (e). The Scale free network (b) and Hierarchical network (c) do not have the peak as random network instead they have the decay according to the $P(k) = K^{-y}$. The clustering coefficient (f) of the hierarchical network is dependent in the degree, thus shows the graph in$C(K) = K-^{\gamma}$.

1.6.1 Random Network

Random network are the large network with no proposed design principle in its formation, a simplest and straightforward form of the complex network. It is a probabilistic model where N nodes connects to the another node, hence creating the graph with approximately $p(N) = \frac{N-1}{2}$ randomly distributed edges(Almaas, Vázquez et al. 2007). The distribution of the node follows the Poisson distribution, thus the average degree of the nodes in the network illustrating the properties of that network. In this network the clustering coefficient is found independent of the node degree k. There is exponential pattern of degree distribution, exhibiting the small world effect. The most of the nodes in the network has the similar number of the links.

1.6.2 Scale free Network

The network that follows the Power law in its degree distribution is the Scale free network.

$$P(k) = K^{-y}$$

This Network was proposed by the Barabasi and Albert (BA) is markedly different to the classical random model. The difference lies in the nature of growth of nodes linking to node in the network. The nodes grow by the addition, linking to the node. The concept of preferential attachment also makes the networks dissimilar to random one; there is higher probability to link to a node with large number of connection. Every m new link are then preferentially attached to a node i (having Ki neighbors) which is represented with probability

$$\prod i = Ki/\sum Kj$$

Where Ki is the degree of node i,

The growth pattern exhibited by the scale free network has the degree exponent $\gamma = 3$ for the Figuree of scale free network proposed by Albert and Barabasi. In the network the node is

highly connected, it is observed that there are few highly connected nodes than average (k>><k>). These highly connected nodes most often determine the network characteristics. Another important feature of scale free network is that it is highly tolerant to the random perturbations but it is highly sensitive to the targeted attack to the highly connected nodes.

1.6.3 Hierarchical network

The real network are expected to be fundamentally modular, which means that networks can be subdivided as the collection of modules where each one follows identifiable task than the other modules present. Modularity refers to that group of physically and functionally linked molecules (nodes) that are working together and each module are destined for distinct functions. The signature feature of network's modularity is high clustering coefficient (http://fenchurch.mc.vanderbilt.edu/bmif310/2009/6-B-Motifs-Modules.pdf) . In order to be evident about the hierarchical nature of such networks, the distribution of the clustering coefficient distribution can be taken into account. The behavior of $C(K)$ is assumed to show different clusters within networks in combined way in an iterative manner, generating the hierarchical network. The coexistence of modularity, local clustering, and scale free topology in real systems can be observed in the hierarchical networks. Such type of network is produced by the repeated duplication and integration process of clustered nodes.

In addition the nature of hierarchical networks to integrate the scale free topology with an inherent modular structure to generate the network follows the power law degree distribution with degree exponent ($\gamma = (2 \leq 3)$). The fact that differentiate the hierarchical network with the scale free and random network is that the clustering coefficient in hierarchical network is dependent to degree where the relation between $C(K)$and (K) is independent in both previous networks.

1.7 The datasets

The datasets were obtained from public databases holding information about transcriptional regulation that involves Tf-Tg interactions. The mode of interaction and experimental evidence suggesting their interaction were the ground for the retrieval of the database. The primary purpose of that database study is to compare the different types of datasets using network analysis. The understanding of interaction between transcription factor and target gene is important for the complete understanding of the biological system since the network formed by these interaction as a single interaction network model provides insight in the principle and properties that control the differential gene expression at system level (Walhout 2006).

The study of different dataset from different databases of transcriptional regulation was performed to analyze the different nature of database, based on the comparative study of the networks formed by them.

1.7.1 Human Transcriptional Regulation Interaction Database (HTRIdb)

HTRIdb is public database for experimentally verified interactions between transcription factor and their target genes. The Network formed by the interaction between transcription factors and target genes is important for the complete understanding of regulation of biological processes. This repository is of experimentally verified human Tf-Tg interaction, specifically physical interaction between the transcription factor and their respective Tg promoters. It can be directly extracted via user friendly web interface. The obtained database after downloading can be interactively visualized as a Network. One of the popular web versions of network visualization tool is Cytoscape.

It is populated with the 284 transcription factors that are involved in regulation with 18302 targets genes, however total of 51871 Tf-Tg interactions with 14 distinct techniques. The network built from the dataset will provide the researcher to decipher the regulation of biological processes (Bovolenta, Acencio et al. 2012). The technique used to observe the binding of the Tf-Tg are Electrophoretic mobility shift assay (EMSA), Chromatin immunoprecipitation (ChIP), Chromatin immunoprecipitation coupled with microarray (ChIP-chip), CpG chromatin immunoprecipitation, Chromatin immunoprecipitation coupled with deep sequencing(ChIP-seq). However, out of 51871 interactions, 2283 interaction were identified by the small and mid-scale techniques like chromatin immunoprecipitation, concatenate chromatin immunoprecipitation, CpG chromatin immunoprecipitation, CpG chromatin immunoprecipitation, DNA affinity chromatography, DNA precipitation assay, DNase I foot printing, electrophoretic mobility shift assay, southwestern blotting, streptavidin chromatin immunoprecipitation, Surface Plasmon resonance, and yeast one-hybrid assay) and 49588 interaction were done by chromatin immunoprecipitation coupled with microarray (ChIP -ChIP) or chromatin immunoprecipitation coupled with deep sequencing (ChIP sequencing).

Out of 2283 interaction based in small and mid-scale technique, search strategy yielded 2471 articles and the Tf-Tg interaction were refined from only 893 articles. The remaining was disapproved in the lieu of fact, caused by gene name ambiguity and lack of clear Tf-Tg interaction in those articles. The search strategy was basically performed in PUBMED,

delimited to the title or abstract focused on the Human- using the complex Boolean query with the words like “bind” and “interact”

Additionally the introspection of presence of Tf-Tg interaction and associated detection technique in articles was facilitated by the annotation tool developed by the group (i.e. Mathematica note book available upon request) that is meant to highlight the gene name or abstracts for Tf or Tg and names of the techniques. The gene name and symbol are taken for Tf and Tg are obtained from the list of gene official and gene names for genes that we built from the Homo sapiens gene information file, downloaded from the National center for Biotechnology Information (NCBI) ftp site (ftp:/ftp.ncbi.nih/gov/gene/). The Tf list was obtained from the high confidence dataset of 1391 Tfs produced by Vaqueriza and colleagues(Li and Kim).

In case of ChIP –ChIP and ChIP –seq experiments initially the referral to the hmChIP database was performed. Later from list of Gene Expression Omnibus series corresponding article were downloaded. On the other hand, PPI of Tfs and Tg s were extracted from the integrated network of human gene interactions published by the group.

The configuration of HTRIdb dataset has the observation ID, Gene ID, Symbol of Transcription factor, Gene ID of Target gene, Symbol of Target gene, Technique and PubMed Id as the reference for the interaction between Tf and Tg.

Some of the major techniques that have been employed in maximum number of observations to detect the binding between the transcription factor and target genes are briefly mentioned;

1.7.1.1 Electrophoretic Mobility shift assay

Electrophoretic mobility shift assay is the technique based on the observation that the complex formed by protein-DNA migrate move slowly than free linear DNA in the non-denaturing polyacrylamide or agarose gel electrophoresis. Since after the protein bound to DNA ultimate shift in rate of DNA migration or the retardation in the speed of migration is observed, which is termed as gel shift or gel retardation assay. In totality the determination of whether protein or mixture of protein as a binding complex is able to bind with provided DNA or RNA sequence observed through this process. It is also used to obtain quantitative information about the site distribution, equilibria and kinetics of protein-DNA interactions. The dependability of this application is on the ability of the electrophoretic system to resolve the reaction time and on their stabilities during separation process(Walhout 2006). The advantage of studying protein-DNA interaction by EMSA is the ability to differentiate the

complexes of different stoichiometry or conformation (http://www.piercenet.com/method/gel-shift-assays-emsa). These gel shift assay are performed in vitro along with DNase foot printing, primer extension and promoter probe when studying promoter probe experiments when studying transcription initiation, DNA replication, DNA repair or RNA processing and maturation.

1.7.1.2 Chromatin immune precipitation

Chromatin immune precipitation is an in situ technique which is performed for a protein of interest, selectively immuno precipitated from a chromatin preparation to know about the DNA sequence associated with it. This is a popular tool to map the localization of post translationally modified histones, histone variants, transcription factors, chromatin modifying enzymes on the genome or on a given locus(Collas 2010). It offers the physiological representation of nuclear events that occur during the processing of DNA(Spencer, Sun et al. 2003). In this assay the cells are first incubated with formaldehyde briefly in order to make sure there is no separation of DNA-associated protein in from the target DNA sequence in subsequent step. Those cells are then meant to be sonicated to fragment the DNA, and later the lysate is centrifuged to remove the cellular debris. Thus, cleared lysate is incubated with primary antibody, and antibody-antigen complexes is then subjected to protein A- or protein G- sepharose depending on the nature of antibody. Then sepharose is washed several times with buffers containing different detergent and salt concentrations, and the antibody-antigen complexes are eluted from protein A/G with help of high detergent elution buffer. At last the protein complement of immuno complexes is digested and finally the DNA is isolated by ethanol precipitation. However the quality of primary antibody, the degree of sample sonication, amount of protein A/G sepharose and the types of salt and/or detergent washes is directly proportional to the accurate determination on distribution of protein along a specific DNA sequence(Spencer, Sun et al. 2003).

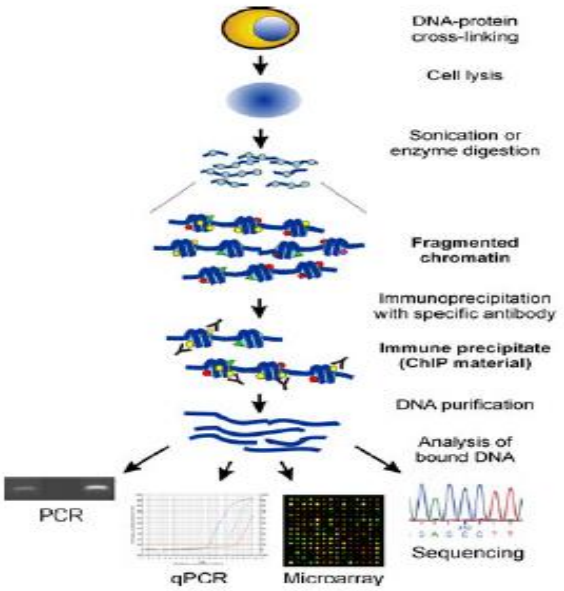


Figure 6The chromatin immunoprecipitation (ChIP) assay and post ChIP various methods of analysis, image reprinted from the source.(Collas 2010)

1.7.1.3 Chromatin immunoprecipitation coupled with microarray (ChIP-chip)

The combination of chromatin immunoprecipitation (ChIP) and DNA microarray (chip) is used to analyze the protein DNA interaction that occurs in living cells. It has enabled researcher to demonstrate high resolution genome-wide maps of the in vivo interaction between DNA associated protein and DNA(Buck and Lieb 2004). In this process protein interaction are captured in vivo by chemical crosslinking. Then the cell lysis will subsequently result in DNA fragmentation and then immunoaffinity purification of the desired protein will co-purify DNA fragments that are associated with that protein. The ChIP enriched DNA population will be amplified by PCR and fluorescently labelled (Cy5). Whereas the other set of cross linked DNA sample, is reverse cross linked and purified, and then labelled (Cy3). The both samples are mixed and hybridized into microarray containing genomic probes covering the whole or parts of the genome. Finally the binding of the DNA to the respective protein in the chip is noticed with the intensity of ChIP DNA significantly exceeding than the purified DNA sample in the array(Collas 2010).

1.7.1.4 Chromatin immunoprecipitation coupled with deep sequencing (ChIP-seq)

The powerful strategy to identify protein binding sites across the genome comprise of directly and quantitatively sequencing ChIP products. The ChIP sequence data present the potential to utilize ChIP for genome wide annotation of novel promoters and primary transcript, active transposable elements, imprinting control regions and allele specific transcription(Mikkelsen, Ku et al. 2007). The basic aim behind the ChIP-seq data analysis is to find the genomic

regions that can be identified as the pool of specifically precipitated DNA fragments. Region of high density of precipitated DNA fragment are referred to as peaks. However the growing relevance of ChIP-seq in the field vows for the transparent and accurate statistical approach that enables a researcher to fully utilize the advantages of ChIP-seq(Valouev, Johnson et al. 2008).

1.7.1.5 DNA affinity chromatography

DNA affinity chromatography is a technique used for the purification of polynucleotides and polynucleotide binding protein including polymerases, restriction endonucleases and proteins involved in recombination and various transcription factors, that control various cellular processes(Chockalingam, Jurado et al. 2000).

1.7.1.6 CpG Chromatin immunoprecipitation

As we know that the property of Chromatin immunoprecipitation with microarray analysis is very useful with a cause to increase both the speed of analysis and number of target isolated. For instance this type of research has been performed in yeast system using yeast genomic microarray(Ren, Robert et al. 2000). They worked with microarray analysis method that ascertains the genome wide position of DNA bound protein and followed this technique to monitor binding of gene specific transcription activators in yeast. However, similar kind of analysis in mammalian cells was not possible to act on. Similar analysis by microarray had to represent a major portion of intergenic region which gets hindered due to vastly greater size of mammalian genomes. The failure in this type of analysis led to the utilization of microarray technique to focus on human genomic fragments that were isolated due to their high CpG content. Often CpG island correlates with promoter region, and are regarded as the most reliable measure for promoter prediction(Ioshikhes and Zhang 2000). Thus the formulatin of CpG island microarray chromatin immunoprecipitation with an antibody to a human transcription factor enables high throughput method of knowing target promoters in vivo(Weinmann, Yan et al. 2002).

1.7.1.7 Yeast one hybrid

Yeast one hybrid screening is a method to rapidly identify the interaction between heterologous transcription factor (prey) and regulatory DNA sequence (bait). The method is dependent on two proteins (bait and prey) interaction detected through in vivo reconstitution of transcriptional activator that turns on expression of a reporter gene. The bait DNA sequence is upstream of a reporter gene.

The prediction and assurance of DNA binding result in reporter gene activation, is entirely on how cDNA expression libraries are used to produce hybrid between the prey and a strong trans-activating domain. Benefit of One hybrid screening, compared to other biochemical technique is the procedure that does not require specific optimization and in vitro conditions(Ouwerkerk and Meijer 2001).

1.7.1.8 Southwestern blotting

This lab technique is the mixture of southern and western blotting; DNA detecting southern blotting detects DNA binding protein and western blotting is for protein detection. The upper hand value of this technique over other relevant methods like EMSA and foot printing is that it provides information regarding molecular weight of unknown protein factor. Basically the technique is about denaturing SDS –PAGE acting to separate proteins electrophoretically based on size. Then the protein is transferred to a membrane support and the renaturation of membrane bound protein and incubated with a (32) P-labelled double stranded oligonucleotide probe of specific DNA sequence. Autoradiography is later, used to visualize the interaction of probe with protein(s)(Siu, Lee et al. 2008).

1.7.1.9 Surface Plasmon Resonance

Surface Plasmon Resonance (SPR) biosensors are optical sensors using special electromagnetic waves – Surface Plasmon Polaritons, to detect interactions between an analyte in in solution and a bimolecular recognition element immobilized on the SPR sensor surface(Homola 2003). It is an exceedingly and powerful probe for interactions of variety of ligands, biopolymers and membranes, including protein-ligand, protein-protein, protein-DNA, and protein-membrane binding. It is ease of method to identify these and interaction and quantifies their equilibrium constants, kinetic constants and underlying genetics. It is employed in very sensitive and label free biochemical assays (https://depts.washington.edu/campbelc/projects/plasmon.pdf).

1.7.1.10 DNase foot printing

Deoxyribonuclease 1 (DNase 1) protection mapping or foot printing, is a technique used for mapping the specific binding sites of proteins on DNA. The assay is dependent on fact that bound protein protects the phosphodiester backbone of DNA from DNase 1 catalyzed hydrolysis. Autoradiography is used in visualization of binding sites on the DNA fragments that result from hydrolysis, following separation by electrophoresis on denaturing DNA sequencing gels. The technique has been further developed as a quantitative technique to

determine and separate binding curves for each individual protein-binding site on the DNA(Brenowitz, Senear et al. 1989).

1.7.2 TFactS

TFactS is a transcriptional regulation database. It attempts to predict the transcription factors, like regulated transcription factors, activated and inhibited transcription factors in a biological condition based in the list of genes it has in its repository obtained from transcriptome experiments. The database also offers catalogue where the list of up regulated and down regulated genes, along with only regulated gene list can be retrieved. Input list for up and /or down regulated (query genes), is compared with a catalogue of annotated target genes, and returns the three lists of transcription factors with annotated target gene showing a significant overlap with the query genes. There are three list of showing regulated transcription factor in first list of sign less catalogue and second list is about activated TF and the third one is about repressed TF. The latter two activated and repressed list are produced using sign sensitive catalogue (http://www.TFactS.org/TFactS-new/TFactS-v2/index1.html).

The organism primarily used is human, interpreting human data along with Rat and Mouse orthologous genes are included in the dataset. There is the information about 338 transcription factor and 2624 target genes and 6823 experimentally verified interactions (15). The list of Tg is obtained from literature and databases like the Transcriptional regulatory element database (TRED), PubMed, TRRD, PAZAR, NFIregulomeDB and their own experimental predictions(Jayavelu and Bar 2014).

Another sign-sensitive dataset of TFactS has the regulation information of the interaction between the interacting Tf and Tg. The Up and Down regulation of interaction is observed in this dataset. The total of 111 Target genes along with 1635 Transcription factor and 3249 interaction is observed in the dataset. The conFigureation of the dataset is Official gene name of Tf, Official Tf coding gene name, Regulation, Reference, Species and Reference accession. The organism used is same as sign less dataset. Whereas the sign-less is formed naming Official Tf coding gene name, Tg Official gene name, References, Species, Accession no.

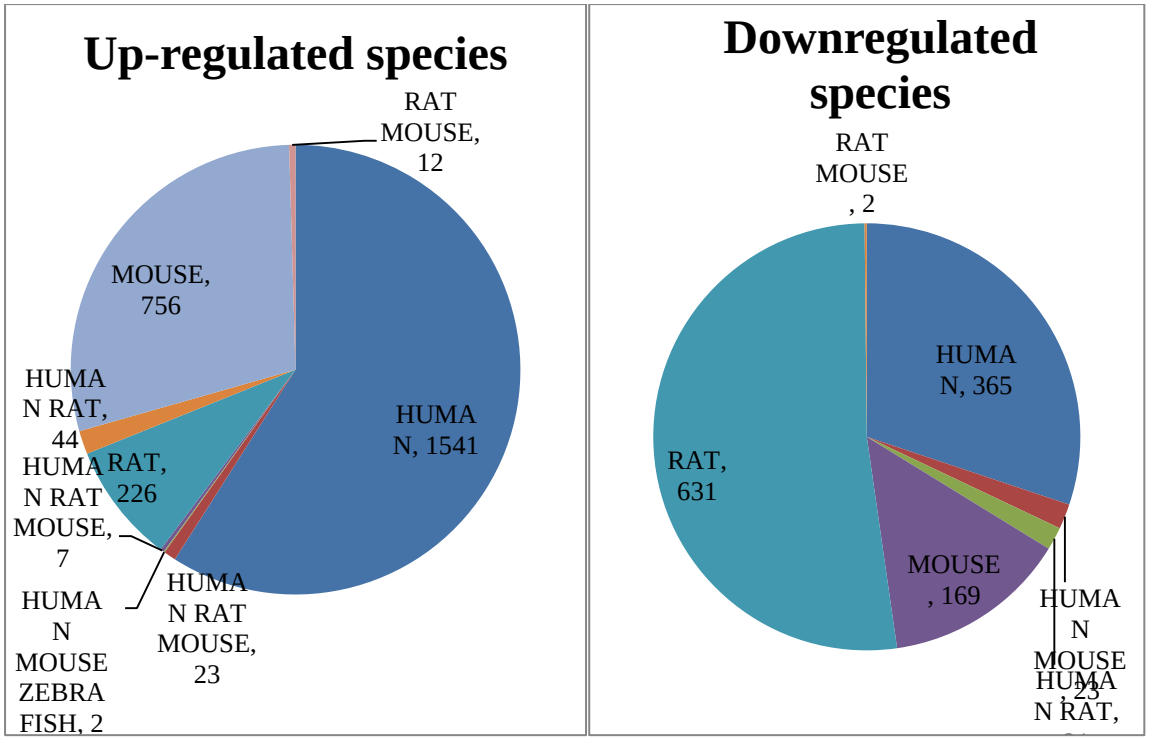


Figure 7In sign-sensitive format of TFactS dataset, the distribution of the species is shown in the piechart, in both Up regulated and Down regulated regulation

1.7.3 YeaSTRACT

YeaSTRACT (Yeast Search for Transcriptional Regulation And Consensus Tracking) is the publicly available up-to-date information on documented regulatory interaction between transcription factor and target genes, as well as Tfs and DNA binding sites, in *Saccharomyces cerevisae.* It is meant for the analysis and prediction of transcription regulatory association in *Saccharomyces cerevisae.*

The database was released in 2006 and by June 2013, it is curated repository of more than 206000 regulatory association between transcription factors (Tf) and target genes (Tg) in *Saccharomyces cerevisae*, based on more than 1300 bibliographic references. It also includes the description of 326 specific DNA binding sites shared among 113 characterized Tfs(REINBERG and Zawel 1993) . This update from the Saccharomyces genome database (SGD) further allows tools to rank the Fs controlling a gene r genome –wide response by their relative importance, based on the 1) Percentage of target gene in the dataset 2) The enrichment of the F regulon in the dataset when compared with the genome. 3) The score computed using the Tf Rank system, which selects and prioritize the relevant Tfs by accessing through the Yeast network.

The dataset is obtained through the rar. File in the webpage of the YEASTACT. The data is in the form of regulation matrix documented csv. format archived as flat files. Later the matrix format file was changed into two columns of Transcription factor and Target gene.

1.7.4 Embryonic Stem Cell Atlas of Pluripotency Evidence (ESCAPE)

Genome wide technologies such as transcriptomics and proteomics are being constantly used to get profile of mouse and human embryonic stem cells (m/hESCs).

m/hESCs centered database called Embryonic stem cell Atlas from Pluripotency Evidence integrating data from many recent diverse high throughput studies including chromatin immunoprecipitation followed by deep sequencing, genome wide inhibitory RNA screens, gene expression microarray or RNA –seq after knockdown or overexpression of critical factors, immunoprecipitation followed by mass spectrometry proteomics and phosphoproteomics. The database prepared also serve as the interactive search tool along with the visualization tool for the sub networks, to identify known and novel regulatory interaction across various regulatory layers.

In addition, database containing information in from a single regulatory layer, mostly transcriptome measurements and thus overlook other important layers as well cross layer interaction. The further data integration was enabled in the field by constructing more inclusive database called Embryonic Stem Cell Atlas from Pluripotent Evidence (ESCAPE). This database integrates numerous additional types of data ranging from epigenetics, transcriptomics, to proteomics and phosphoproteomics. The dataset are prepared in the form of gene list; gene-gene and protein-protein interaction form.

The dataset constructed from mouse/human embryonic stem cell was prepared based on numerous genome wide profiling studies, as well as loss of function/gain of function (LOF/GOF) studies. Most data sets are obtained with source as mouse with several human embryonic stem cells. The current version of ESCAPE contains 1) 206521 documented protein/DNA interaction form ChIP-ChIP/seq studies, connecting 61 transcription factors (Tfs) to their respective target genes 2) 153920 LOF/GOF interaction connecting 28 Tfs from LOF KD/ knockout studies followed by genome wide expression. These interactions directly or indirectly connect to the target gene to an upstream Tf regulator.

ESCAPE database is a user friendly web interface that allows an easy browsing and querying. The dataset retrieved has conFigureation like source name, Source ID, Target name, Target ID, PMID, ChIP type, Cell type, Modification date.

The different ChIP types used in the process are

1.7.4.1 ChIP-PETs

Rather than using the entire insets, the technique of ChIP paired end Ditags (PETs) is based on the sequencing portion of precipitated DNA. The survey on cloned ChIP DNA fragment library indicated inefficiency in distinguishing genuine binding sites and noise, without any molecular validation(Collas 2010). On the other hand, ChIP-PET enhance information content and increase accuracy of genome mapping (Wei, Wu et al. 2006).

ChIP PET technique is based on the strategy of Gene Identification Signature (GIS) where 5` to 3` signatures of full length cDNA are extracted into PETs and are joined later for efficient sequencing and mapping to genome sequences to delimit the transcription boundaries of every gene. The analysis method is 30 fold more efficient than standard cDNA sequencing approaches for transcriptome characterization(Ng, Wei et al. 2005).

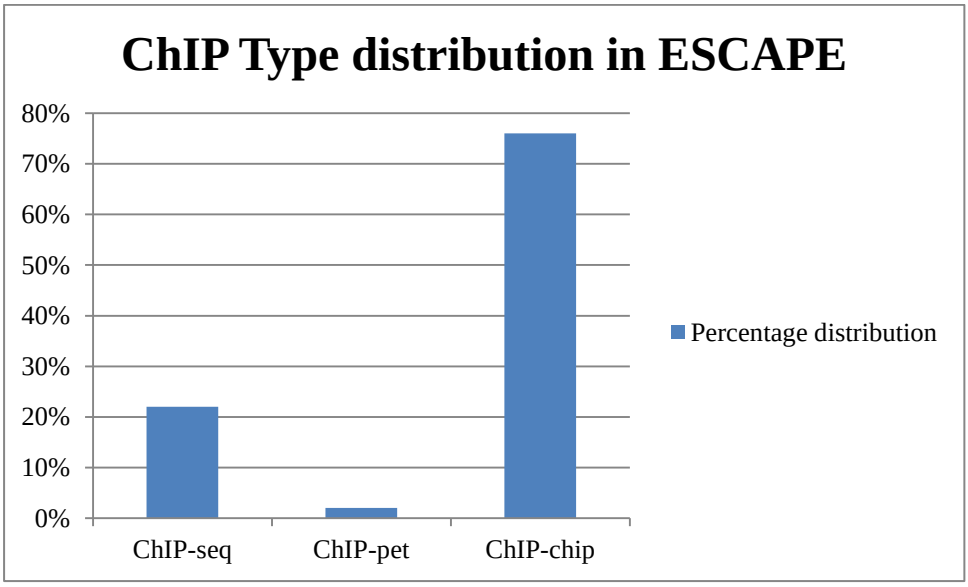


Figure 8The Bar-chart showing the distribution of the ChIP type in the ESCAPE dataset.

2 Material and Methods

The functional genomics and proteomic technique is used for the measurement of expression profiles and functional interactions from the cell and tissues of many different organisms (Barabasi and Oltvai 2004). The estimation/ measurement have the acknowledgeable potential to map the cellular potential and their various dynamics provided the computer software to filter and interpret the large amount of data. The biological and theoretical viewpoint, establishes a notion that different interaction among biomolecules forms different types of network or graph like PPI network, metabolic network, signal transduction network, gene transcriptional networks and gene regulatory networks(Zhang, Jin et al. 2007). The easily accessible software Cytoscape is used for these complex network visualization, with default parameters for the analysis and mapping. Cytoscape is open source bioinformatics software for visualizing molecular interaction networks also integration of gene expression profiles data and other various state data to analyze. The Cytoscape applications, previously called Plug-in, are provided in the store for molecular and network profiling analyses. It has different layouts and additional file formats for the support and connection with different database. The application can be developed by using the Cytoscape API based on Java™ technology and the application development is highly encouraged. Thus the already available application can be downloaded for free from Cytoscape App Store (http://apps.cytoscape.org/) . The combination of cellular network information, including protein-protein and protein-DNA interaction and expression profiles, can explain the control mechanisms underlying the observed changes in activity of a biological process. For example we can relatively know which transcription factor is to interact and regulate with the particular set of genes (Cline, Smoot et al. 2007).

2.1 Network analyzer

Network analyzer is versatile and customizable, plugins of Cytoscape. This enable user to operate the Network without any expert knowledge in graph theory. It calculates and displays a comprehensive set of topological parameters and centrality measures for directed and undirected networks. The parameters in which Network analyzer works are; Number of nodes, edges and connected components, diameter and radius of network, centralization, clustering coefficient, degree and characteristics path length and many more.

The reason using this plugin is to operate the graphical interpretation through Cytoscape. The network analyzer result is used in the report to validate the assumptions. Therefore table 1-8,

below are included as the source of all result, and preliminary assumptions about a network topology.

2.1.1 Simple parameters

The physical modelling of the dynamic of the network provides understanding via statistical properties of real network(Siu, Lee et al. 2008). The graphical analysis of the dataset was performed by the network analyzer of the Cytoscape. The directed and undirected form of network were observed for the calculation based on different topological parameters like clustering coefficient, average shortest path length, Degree and many more.

- The Clustering coefficient of the network is the measure of cohesive ness of the network. It has value within 0 to 1.
- Diameter of the network is the longest value of the calculated shortest path in the network. In the simple terms it is obtained after the shortest path of from a node to another node is calculated and the longest of the path is chosen as the network diameter.
- The network diameter and the shortest path may give indication to the small world property of the network.
- Characteristics path length (L) of the network is the shortest path length between two nodes averaged over all pairs of nodes and is calculated as

  $L = \sum i \sum j \, Li.Lj/N(N-1) \ldots$ (i)

Where, L_{ij} is the shortest path length between i^{th} node and j^{th} node.

2.1.1.1 HTRIdb dataset

A) Directed network

Table 1The Directed Network parameters of HTRIdb dataset observed from Network analyzer of Cytoscape

Clustering coefficient	0.244	Number of nodes	18310
Connected component	2	Number of edges	52324
Network diameter	10	Network density	0.0
Network radius	1	Isolated nodes	0
Shortest path	861526(0%)	Number of self-loops	0

Characteristic path length	3.850	Multiedged node pairs	463
Average number of neighbor	5.662	Analysis time (sec)	338.686

Table 1:.

B) In directed network

Table 2The Undirected Network parameters value of HTRIdb dataset observed from Network analyzer of Cytoscape

Clustering coefficient	0.398	No of nodes	18308
Connected component	1	Network density	0.000
Network diameter	6	Network heterogeneity	22.318
Network radius	3	Isolated nodes	0
Network centralization	0.532	Number of self-loops	28
Shortest paths	335164556(100%)	Multi edged node pairs	503
Characteristic path length	2.482	Analysis time(sec)	6618.547
Average number of neighbor	5.662		

Table 2:.

2.1.1.2 TFactS dataset

A) Directed network

Table 3The Directed Network parameters of TFactS dataset observed from Network analyzer of Cytoscape

Clustering coefficient	0.073	No of edges	6784
Connected components	10	Network density	0.0
Network radius	1	Isolated nodes	1
Shortest paths	35224 (4%)	Number of self-loops	0
Characteristic path length	3.992	Multi edge node pairs	34

Average no. of neighbors	4.830	Analysis time	5.236
No. of nodes	2794		

Table 3:

B) Undirected Network

Table 4The Undirected Network parameters value of TFactS dataset observed from Network analyzer of Cytoscape

Clustering coefficient	0.128	No. of nodes	2794
Connected component	10	Network density	0.002
Network diameter	8	Network heterogeneity	3.995
Network radius	1	Isolated nodes	1
Network centralization	0.203	No. of self-loops	42
Shortest path	7692332(98%)	Multi edged node pairs	34
Characteristic path length	3.355	Analysis time	193.755
Average no of neighbor	4.830		

Table 4:.

2.1.1.3 YeaSTRACT dataset

A) Directed network

Table 5The Directed Network parameters of YeaSTRACT dataset observed from Network analyzer of Cytoscape

Clustering coefficient	0	Number of nodes	541
Connected component	1	Network density	0.0
Network diameter	1	Isolated nodes	0
Network radius	1	Number of self-loops	0
Shortest path	7876(2%)	Multiedged node pairs	0
Characteristic path length	1.0	Analysis time(In sec)	0.403
Average number of	29.16		

neighbors			

Table 5:

B) Undirected dataset

Table 6The Undirected Network parameters value of YeaSTRACT dataset observed from Network analyzer of Cytoscape

Clustering coefficient	0.0	Number of nodes	541
Connected component	1	Network density	.054
Network diameter	5	Network heterogeneity	0.949
Network radius	3	Isolated nodes	0
Network centralization	0.314	Number of self-loops	0
Shortest path	292140 (100%)	Multiedged node pairs	0
Characteristic path length	2.501	Analysis time(In sec)	22.508
Average number of neighbors	29.116		

Table 6:.

2.1.1.4 ESCAPE dataset

A) Directed network

Table 7The Directed Network parameters of ESCAPE dataset observed from Network analyzer of Cytoscape

Clustering coefficient	0.340	Number of nodes	23107
Connected component	1	Network density	0.0
Network diameter	5	Isolated nodes	0
Network radius	1	Number of self-loops	38
Shortest path	1338017 (0%)	Multiedged node pairs	17509
Characteristic path length	2.224	Analysis time (In sec)	1137.607
Average number of neighbors	15.466		

Table 7: Undirected Network

Table 8The Undirected Network parameters value of ESCAPE dataset observed from Network analyzer of Cytoscape

Clustering coefficient	0.540	Number of nodes	23107
Connected component	1	Network density	0.001
Network diameter	5	Network heterogeneity	13.520
Network radius	3	Isolated nodes	0
Network centralization	0.485	Number of self-loops	38
Shortest path	533910342 (100%)	Multiedged node pairs	17509
Characteristic path length	2.432	Analysis time(In sec)	24840.84
Average number of neighbors	14.466		

Table 8:.

(Source: Network analyzer of Cytoscape at 18.00, 19.6.2014)

2.2 MCODE

Extracting all the sub networks in the huge network appears practically un-doable. Therefore there are different algorithms are being developed. The different algorithm focusses on estimating densities of sub networks and detects various motifs from complex network despite of huge network size(Zhang, Jin et al. 2007). Thus these numbers of clusters within the bigger network, later divided are useful in revealing specific local properties.

The MCODE detection method of different repeating motifs within the complete network enabled to produce different clusters (Sub networks). In the research, a basic criterion to differentiate is the MCODE score.

2.2.1 Node scoring

The Node scoring assigns the high score to nodes whose immediate neighbors are more interconnected. Node score is dependent on the K-core value of a particular cluster. K-core of the node(vertex) in that cluster of any nodes is value of highest number of degree of node in that cluster.

Therefore, $Node\ score = K - core * core\ density$

Where core density is equal to the number of edges divided by the possible no edges to the vertex with highest K-core value.

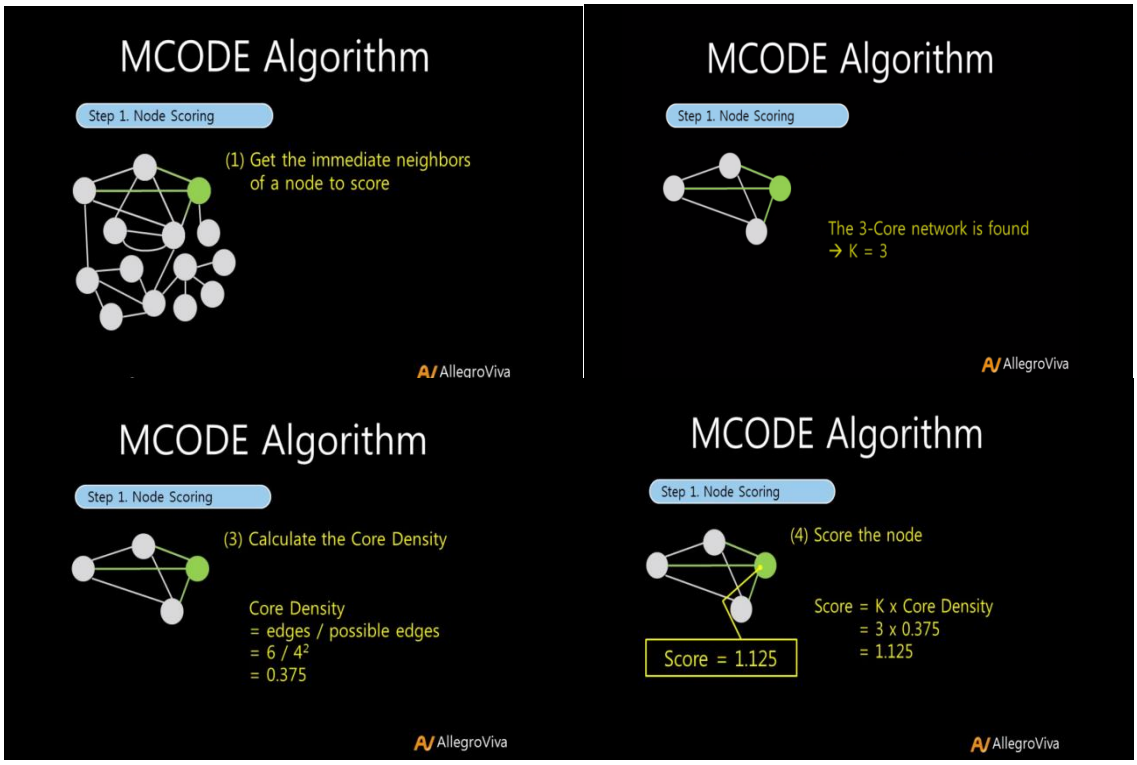


Figure 9The Screenshots obtained illustrating the MCODE algorithm calculations

(Source: http://allegroviva.com/allegromcode/mcode-algorithm/)

2.2.2 Cluster finding:

The cluster with the MCODE algorithm is to find the complex within network with the help of highest scored node (seed), and include all the nodes which have the scores above a given threshold. The procedure is repeated throughout the network; the obtained clusters are filtered out and make sure that they do not contain k-core networks. K-core filtering is the technique to remove the cluster that do not contain at least K-core network. K-core of network is the maximum induced graph where every vertex has at least K degree or it is anchored.

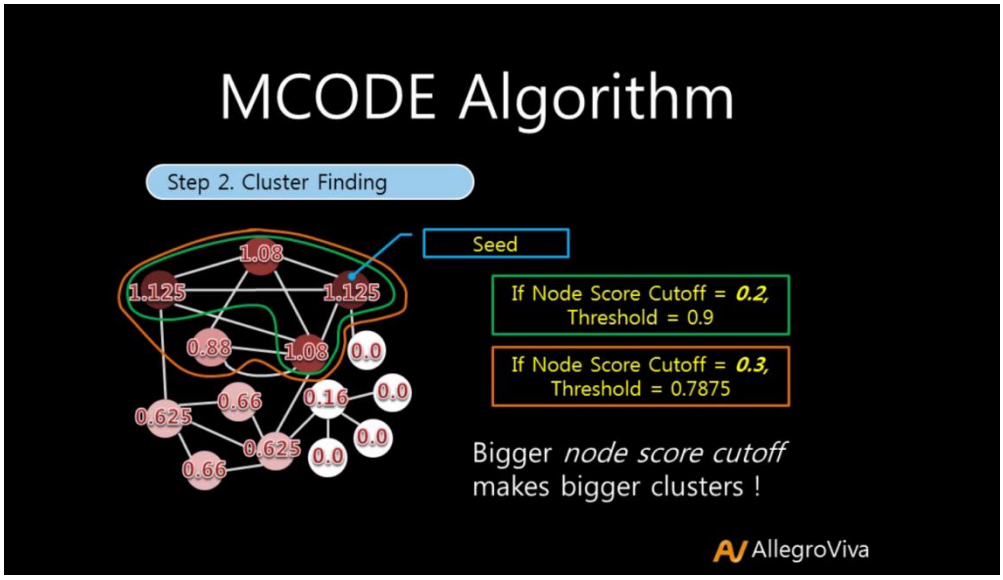


Figure 10The Screenshot obtained illustrating MCODE algorithm calculations

(Source: : http://allegroviva.com/allegromcode/mcode-algorithm/)

2.2.3 Post processing:

The process involves the Haircut technique and Fluff. Haircut removes the singly connected nodes from the selected cluster. Fluff is dependent of the selection based on the node density. Node density is calculated by$\frac{edges}{possible\ edges}$, when density cutoff is fixed as 0.2, all the other nodes having node density below 0.2 is removed.

(Source: http://allegroviva.com/allegromcode/mcode-algorithm/)

2.3 BiNGO

The overrepresentation of genes in the set of genes is performed through BiNGO in this report. The Biological Networks gene Ontology tool (BiNGO) is an open source java tool which is used to extract information about the gene, which is represented by gene ontology term. The BiNGO enables to get information about over representation of a gene in the set of genes. It can be used on the list of genes available, pasted as text, or sub graphs of biological networks visualized in Cytoscape. The later in the list i.e. using sub graphs of the different transcription regulatory network visualized from Cytoscape is used. BiNGO basically performs by mapping the dominant functional themes of the tested gene set on the GO hierarchy. The Cytoscape as a versatile visualization environment has enabled BiNGO to produce an intuitive and customizable visual representation of results.

Gene ontology is an initiative to correlate the genes by unification, i.e. representing gene and gene product attributes across all species by maintaining controlled and universal vocabulary

of gene and gene products. The gene ontology tools offer one to functionally interpret experimental data using it, for example via enrichment analysis. The major advantage is that proper, uniform and controlled vocabularies of GO which is applicable to all organisms. The Gene ontology (GO) project was initiated in late 1990`s with a purpose of capturing the increasing knowledge on gene function in a form of controlled vocabulary which is applicable in all organisms in a controlled way. Basically Gene ontology consist of three hierarchical structure of vocabularies, that gene products based in terms of their associated biological processes, molecular functions and cellular components.

BiNGO is the source of annotation for the wide range of organisms. The default annotations are exacted from GO information available from NCBI formal webpage. The numbers of gene identifier can be supported but NCBI`s EntrezID is most stable one and mostly used.

Source: (Maere, Heymans et al. 2005).

2.4 MiMI

MiMI is a Cytoscape plugin which retrieves molecular interactions from a database named Michigan Molecular Interactions (MiMI) and later displays the molecular interaction network in Cytoscape. The database gathers all the information from and merge from well-known protein interaction database including BIND, DIP, HPRD, RefSeq, Swissprott, IPI and CCSB-HI1 etc.(http://apps.cytoscape.org/apps/mimiplugin).

3 RESULTS

3.1 Node Degree Distribution

Node degree distribution analysis gives the global level of analysis to the regulatory network, where the regulatory network mostly characterizes itself by the scale free topology, with the presence of regulatory hubs (Babu, Luscombe et al. 2004; Janky, Helden et al. 2009). The degree of a node allows immediate evaluation of the regulatory relevance of that particular node. For instance, in gene regulatory network, a transcription factor of high degree is understood to be interacting with higher number of target genes. Hence enabling that node to play the central role, that is likely to be in regulatory hub. There are few regulatory transcription factors that create a regulatory hub within the network. In this thesis, we will analyze different transcription factors from four datasets. The transcription factor forming regulatory hubs will be further dealt to find out the biological significance in both cases.

Node degree analysis and distribution is performed from the dataset to understand the pattern of distribution of node degree. There are two types of nodes in this analysis: transcription factors and target genes. This means that one may identify Out-going degrees for Tfs (directional links from a Tf to a TG, representing a Tf that provides output to a Tg) and incoming degrees for Tgs (representing Tgs that receive input from a Tf). The node degrees in any biological network are distributed in fashion where few of the nodes in the network have higher degree, whereas other has fewer. The nodes having higher degree are in less number compared to those nodes which have less degree. Thus in this way, we can identify few regulatory hubs in the network which are central to the network topology. The regulatory network forms the sub network within the network. The poorly connected nodes are most of the time linked to any of these regulatory hubs.

3.1.1 HTRIdb dataset

The node degree distribution of the network determines the behavior of the transcription factor to the respective target genes. In the dataset the number of interacting transcription factor is 284 and target genes is way higher up to 18028.

- **Node degree distribution of Undirected Network**

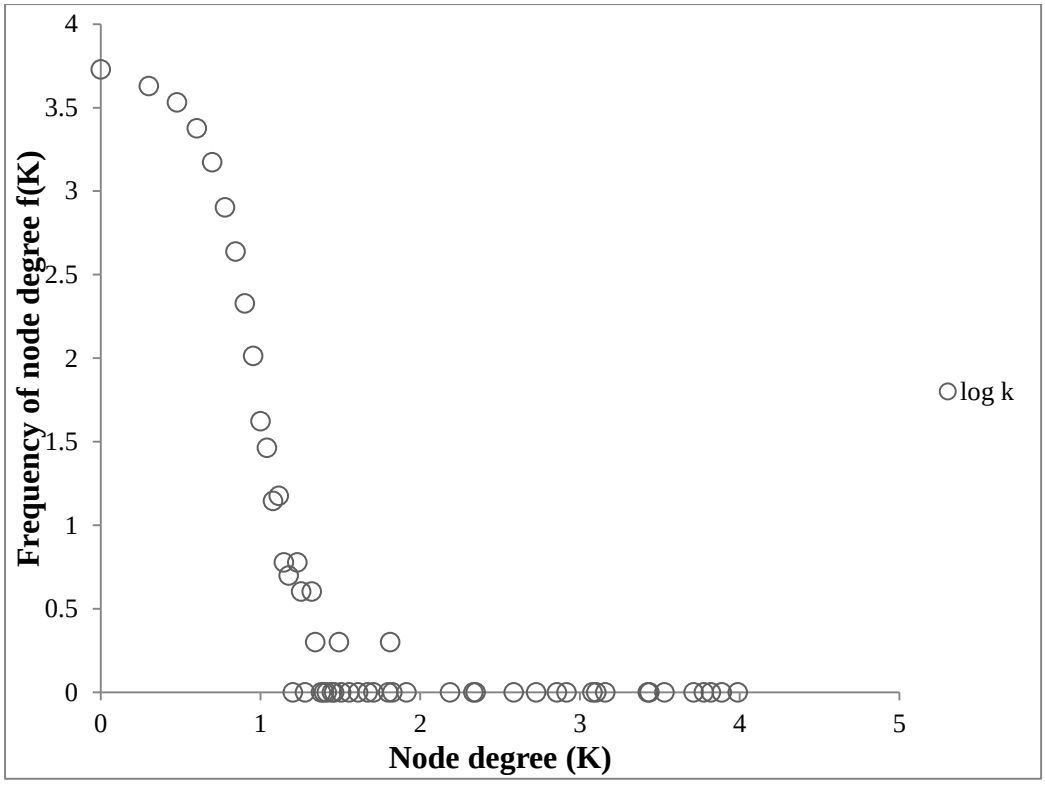


Figure 11The double log plot of Node degree distribution of undirected network on HTRIdb dataset.

Out of 18,308 nodes, constituting both transcription factors and target gene in the evaluation of the node degree, I found highest distribution equal to the log of 3.72, which is equivalent to 5335 nodes for node having degree (K) 1. Similarly, it is observed that there are 12 nodes which have degree of more than 1000 but the frequency is only 1. The power law definition of the scale free network speaks about sparsely connected nature of most nodes, and the few nodes with highly connected attribute, hence those highly connected nodes is expected to play an important role in the functionality(Ouwerkerk and Meijer 2001). The degree (K) distribution of the network is scale free in nature; however it forms the hierarchical network with the modular structure present. Thus it is the network for scale free modular hierarchical network.

- **Out –degree Distribution of Transcription factor**

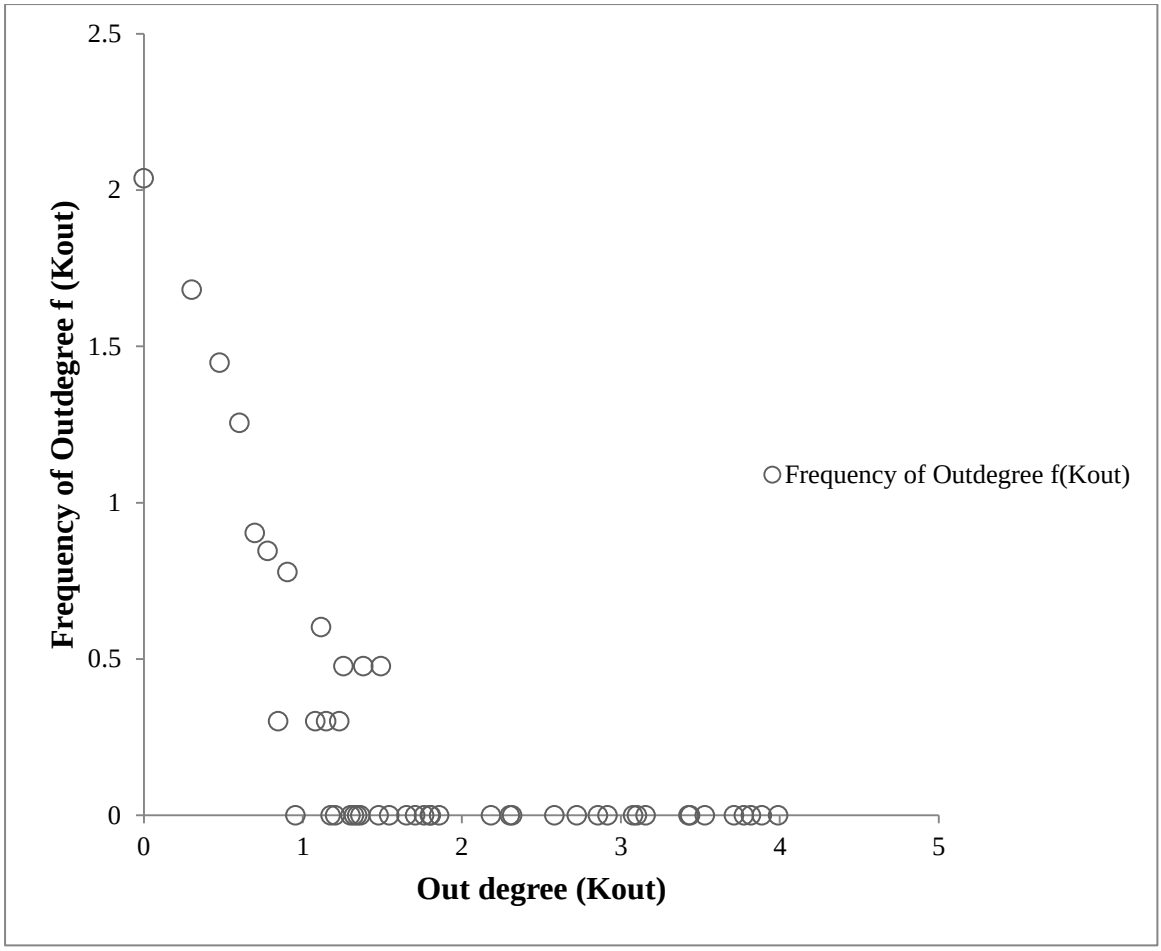


Figure 12Double log-plot of Outdegree distribution of Transcription factor on HTRIdb dataset.

It reflects the scale free hierarchical graph with modules present. The hierarchical network formed of the 283 transcription factors shows the tendency on frequency of transcription factors with minimum of $\log 0$ to $log2.1$. Transcription factor which has the highest involvement in the network is ETS1, with Outdegree $(Kout)$ of 9759. Similarly there are other 10 more transcription factors with heavy involvement in the different biological process in Human. Some of them are GATA2, AR, YBX1, FOXP3, GATA1, PRDM14, E2F4, ESR1, GATA3, and TFAP2C. The total of 109 vertices (nodes) is found with Outdegree $(Kout)$ of only 1.

The observed case of some transcription factor like AR which has mode of detection of binding to the gene by chromatin immunoprecipitation coupled with microarray has highest Outdegree $(Kout)$ of 9759 with total of 9759 bindings with different target gene with same techniques. In other words: this observation implies that AR has a one to many relationships with target genes.

Mostly the transcription factors which have higher Out- degree are found to be in very minimum number compared to other transcription factors out of those 283 ones. Total of 10 transcription factors have Outdegree ($Kout$) more than 1000.

- **In-degree Distribution of Target genes**

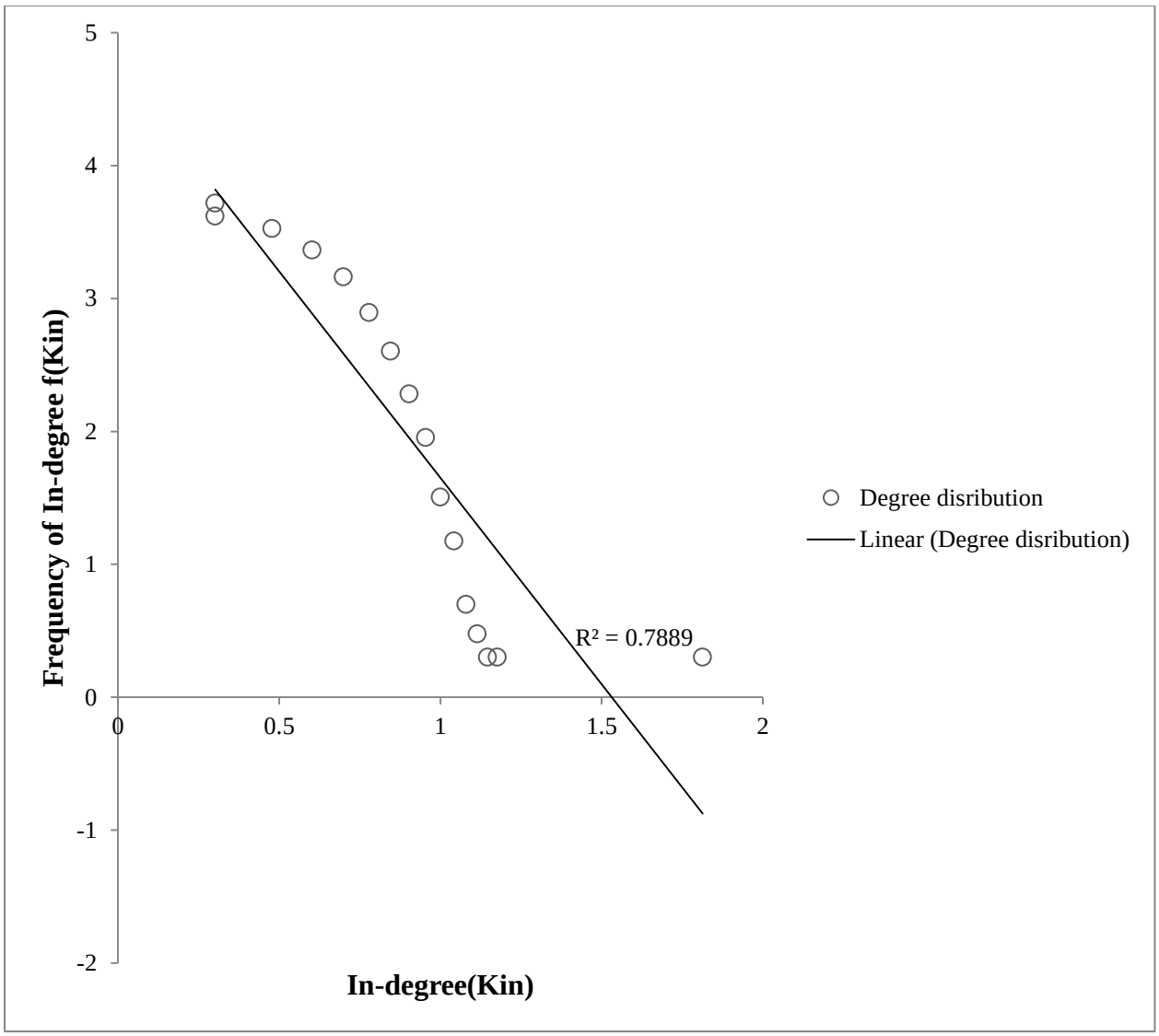


Fig 4.3: The double log plot of In-degree distribution of target genes in HTRIdb dataset.

The distribution of target genes in HTRIdb dataset is observed with scale free distribution for indegree of target genes. In this case the highest In-degree (Kin) of target gene is 65 which is followed by one having indegree of 19.It is supposed that many Tf arrive at transcription start site of these genes. However after there are target genes similar Indegree (Kin) like those ranging from 17 up to 14. The distribution of indegree of every target gene, shows there are around 5203 target gene which have 1 Indegree value. The Indegree (Kin) of target genes in the HTRIdb dataset has the power law exponent (γ) = 3.041, with the correlation of

the graph as 0.713, thus confirming the indegree (Kin) network formed by the target genes form a scale free network.

3.1.2 TFactS data

The TFactS data has two set of data, one is Sign-sensitive whereas another is Sign-less. The sign less format of the TFactS dataset has the set of interaction between Transcription factor and Target gene in the different species, supported by the various references and accession number. The sign less format of the TFactS does not provide the level of regulation between Tf-Tg. On the other hand the sign sensitive form of the dataset has the additional information on the regulation of the Tf-Tg interaction.

3.1.2.1 Sign-less dataset

The sign-less data retrieved from the TFactS dataset catalogue contains 342 transcription factor and 2450 target genes.

- **Node degree Distribution of Undirected Network**

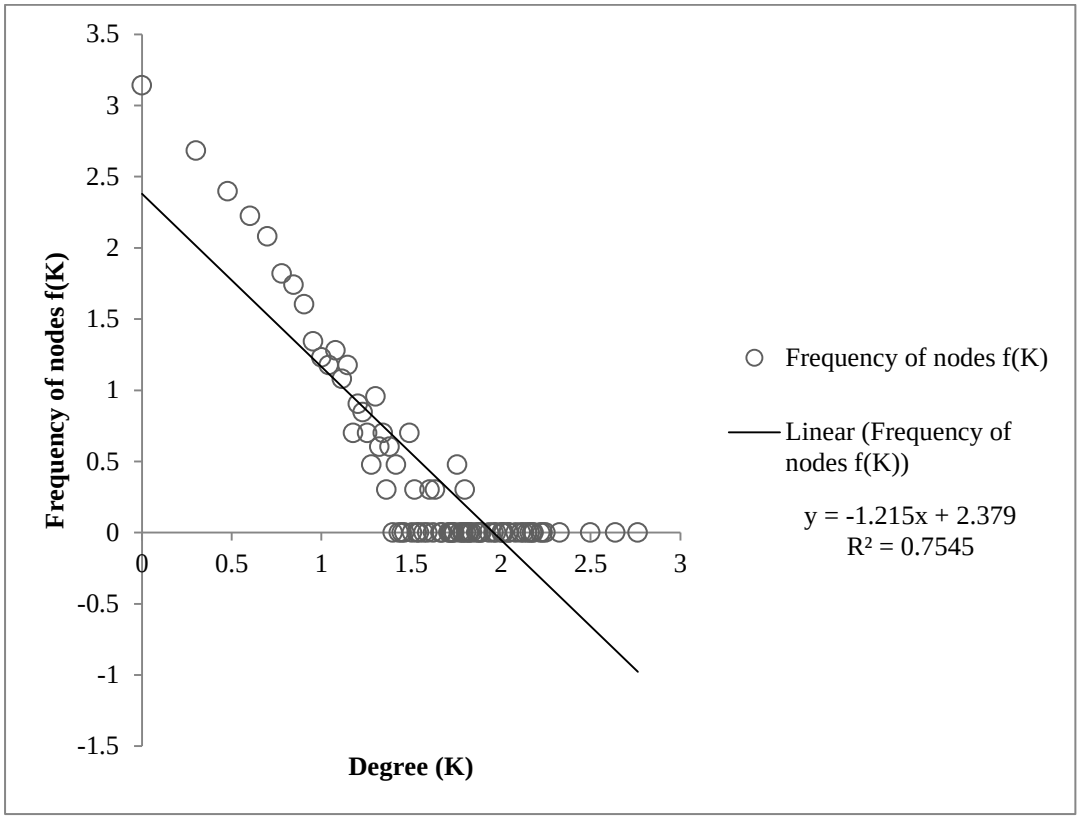


Figure 13The double log plot of Node degree distribution of undirected network In Sign-less TFactS

Out of total 2792 nodes in the TFactS dataset, which includes both transcription factor and target gene together as node, is found to be forming a scale free network. The node closest to the logarithmic of 2.75 in the graph is MYC, and the number is 579. Similarly SP1 and

CTNNB1 are also in similar range with values former being higher than the later. There are around 30 nodes with the unique degree, thus they have the frequency of 1, so they lie in the x-axis of the graph. There are 1383 nodes with the degree (K) of 1. The particular nodes with the degree ranging from 10 onwards to the range of 40 have the frequency of around 5 in average. Hence, it is observed there are many nodes clubbed around $\log of$ 0.5.that is understood as the nodes with the node degree ranging around 40 have similar distribution. The number of highly involved nodes confirms that presence of several functional modules within the network thus confirming it to be Scale free hierarchical modular network.

- **Outdegree Distribution of Transcription factor.**

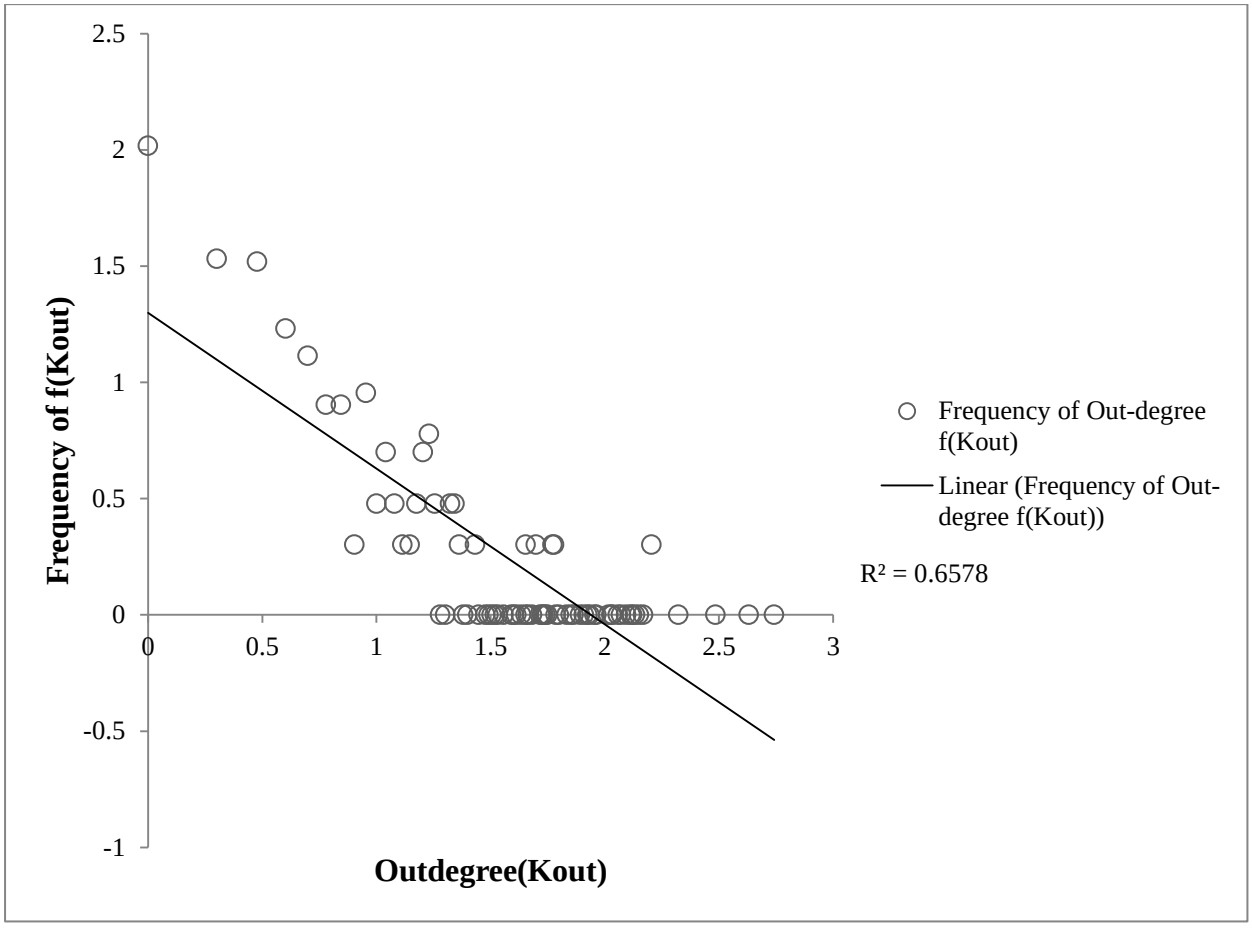


Figure 14The double log plot of Outdegree of Transcription factors on sign-less TFactS

The Outdegree $(Kout)$ of transcription factors in the TFactS dataset is similar to the overall (undirected) node degree distribution. The Outdegree of the transcription factor in the dataset is maximally influenced by the Myc, which has the Outdegree $(Kout)$ count of 552. In the same way, SP1 and CTNNB1 Outdegree are in the higher side compared to other transcription factor, in terms of their involvement with the target genes. There are 104 singly edged transcription factors, meaning that they link only to a single target gene i.e. Outdegree

(Kout) =1. Few nodes that are highly connected has major role to play, and if any destruction to the network by the targeted attack may result in loss in the essence of the network. This is the small world property in the scale free network. The hierarchical network is observed in the case of Outdegree $(Kout)$ of transcription factor.

- **In degree Distribution of Target genes**

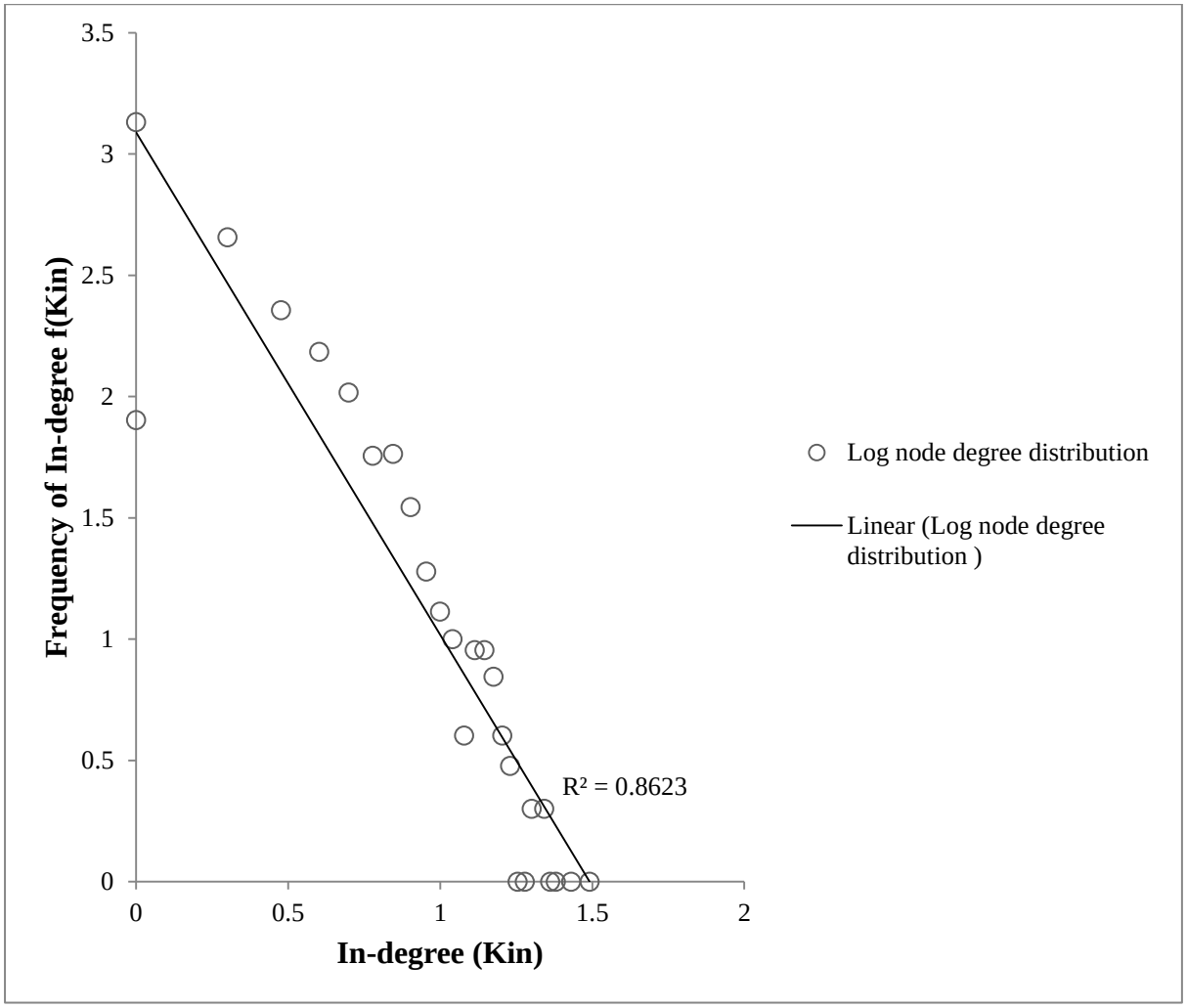


Figure 15The double log plot of In-degree distribution of Target genes in Sign-less TFactS

About In-degree distribution of the target gene in the TFactS sign lesss dataset, there are around 2450 target genes having certain number of transcription factors acting on them. The regulation of Tf-Tg in the TFactS is mainly depicted as highly involving transcription factors to the limited number of target genes. There is around 8 target gene with unit frequency and having Outdegree of around 15-40. In the similar comparison, the target genes with the Indegree (Kin) of 1 are in frequency of around 1000.

The nature of graph created by the indegree distribution of target genes in Sign less TFactS dataset is of Scale free nature. The Power law distribution of the target genes in the network can be verified by the γ= 2.42 with the correlation of 0.981 reflects it is a scale free network.

3.1.2.1.1 Sign sensitive dataset

In the sign sensitive TFactS dataset, there are 111 Target gene and 1635 transcription factor and total of 3249 Tf-Tg interaction in sign sensitive dataset of TFactS. Basic design of sign sensitive differs from another version of TFactS dataset because it includes the additional information of its regulation. There are 2616 up regulated interaction and 633 down regulated interaction. The observed interaction happens in different species and also the same interaction is seen in more than one species.

- **Node degree distribution**

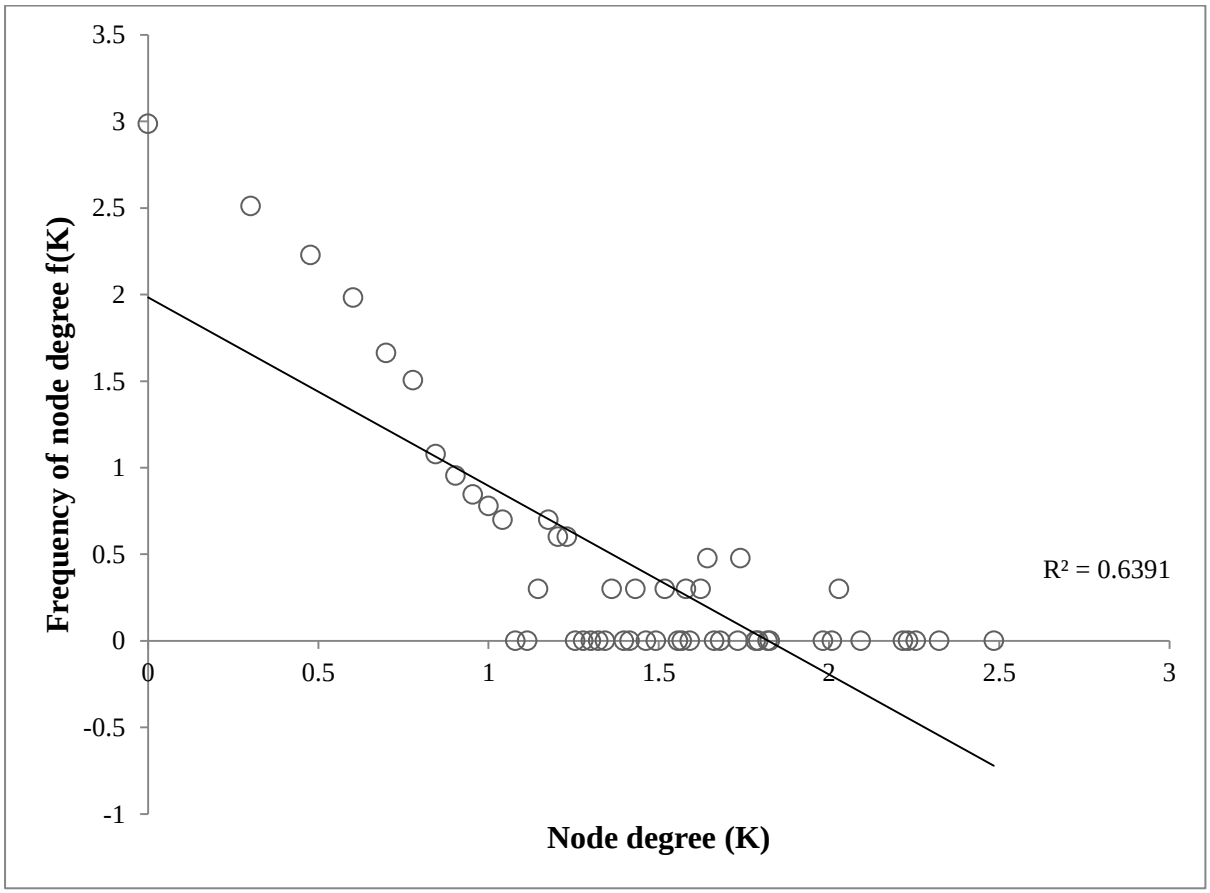


Figure 16The double log-plot node degree distribution of sign-sensitive TFactS.

The node degree distribution of the nodes of the TFactS dataset in the sign sensitive format has the distribution of nodes accordingly the scale free nature distribution in the network. The node degree of 1 has the frequency of 968 that means the most of the nodes are singly connected. In precise there are 28 nodes with the unit frequency and node degree (K) ranges

from 12 to 305. The highest node degree in the dataset is of CTNNBI of 305 and all other like MYC, SPI, E2F, and FOXO1 has the node degree of 211, 180, and 171,165 respectively. The degree exponent value of the network is γ=1.218 with correlation of 0.989, thus with the different modular features within the network the assumption of it being a hierarchical network is conferred through graphical observation.

- **Outdegree distribution of Transcription factor.**

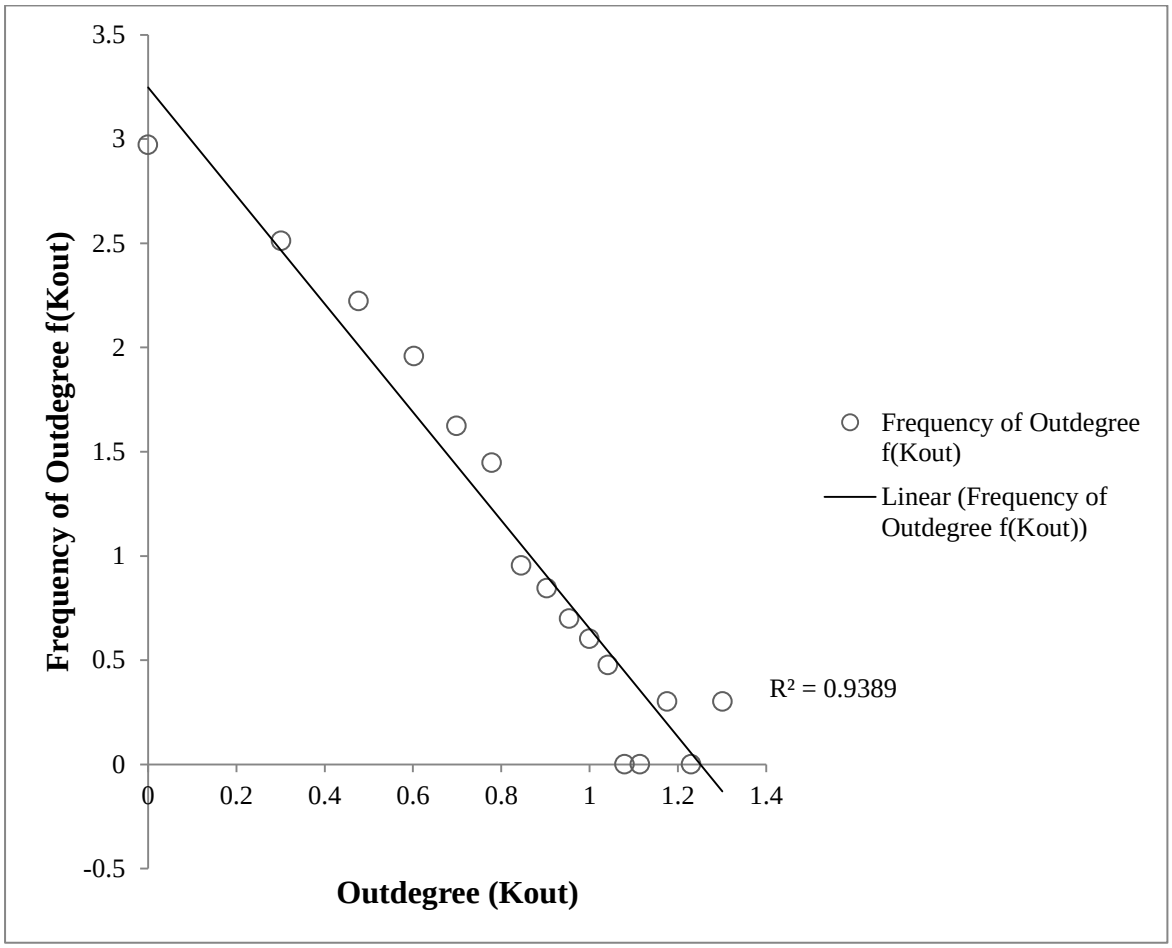


Figure 17The double log-plot Outdegree distribution of signs sensitive TFactS.

The Outdegree $(Kout)$ of transcription factor in the sign sensitive dataset of TFactS has the graphical representation for 1635 transcription factors, but most of them have small Outdegree. The highly interacting transcription factors in the dataset are SPP1, CDKN1A, BCL2, E2F1, MYC and few more. The highest of Outdegree $(Kout)$ is around logarithmic of 1.3 i.e. around 20, which is for transcription factor CDKN1A.

The power law distribution of the nodes in the network for the transcription factors of TFactS sign sensitive dataset show scale free architecture. The power law coefficient (γ) is 2.593

with the correlation of 0.978, hence confirms the network formed by the transcription factor Outdegree $(Kout)$ is a scale free type.

- **In-degree distribution of Target genes**

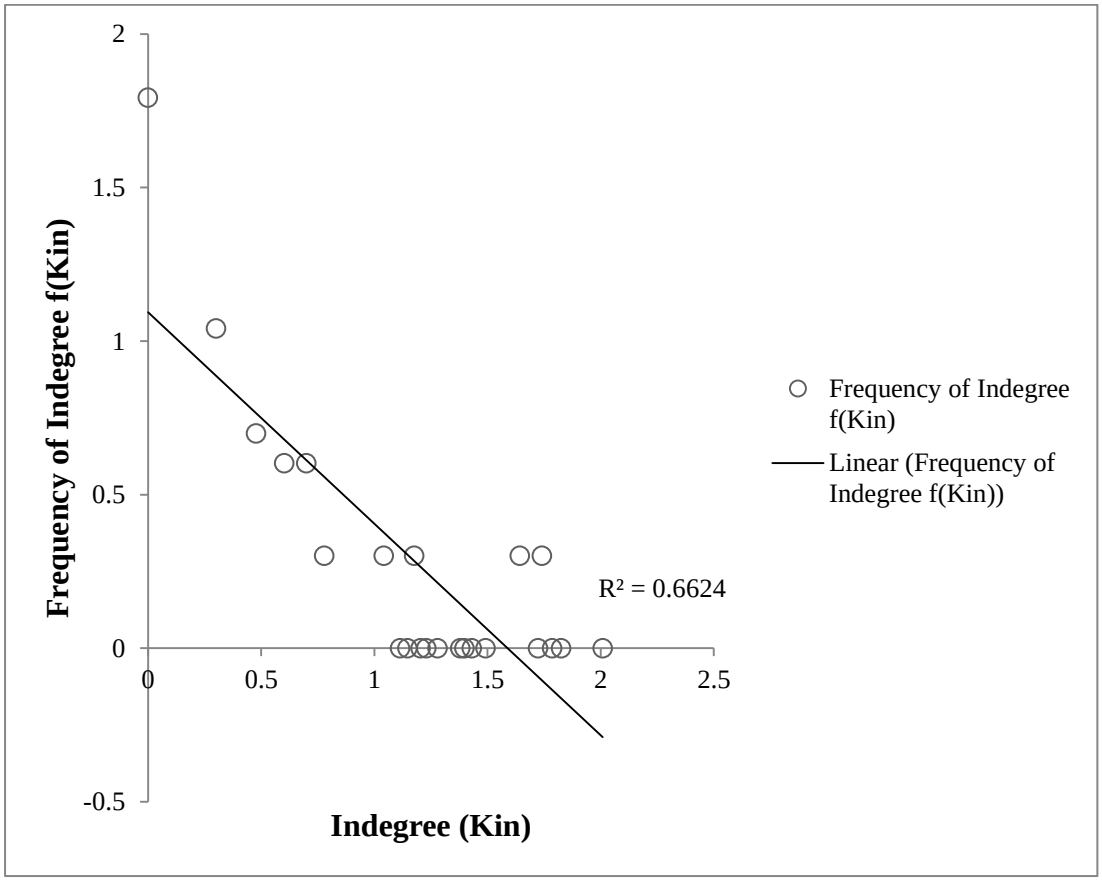


Figure 18The double log-plot of In-degree distribution of sign-sensitive TFactS

The number of target gene in the sign-sensitive dataset of TFactS is 111. The target genes in this dataset are having similar range of frequencies like the one target gene with the in degree (Kin) ranging from 5 to 8 have the similar frequency of around five. There is another range of similar frequency for target genes in the dataset which have the Outdegree ranging from the 8 to 60 with the same frequency of 5. The most value of in degree (Kin) of target gene is observed to be Kin =100.

The peak for target gene with any node degree value is observed. The graph favors the exponential distribution on the graph. With the γ being 0.5 and the no more nodes forming a bigger hubs, that should corresponds to around 60 % of total indegree, (Wuchty, Ravasz et al. 2006) the exponential nature of the network can be assured. The graph also shows the peak at log 1.8, confirming it to be an exponential graph.

3.1.2.2 YEASTRACT dataset

The data obtained from the network analyzer of Cytoscape were used to make graphs. In this dataset the 254 transcription factors and 287 target genes are involved. Nearly equal number of target genes and transcription factor is observed which is rare, since I have been working with similar datasets as well.

- **Node degree distribution of Undirected Network**

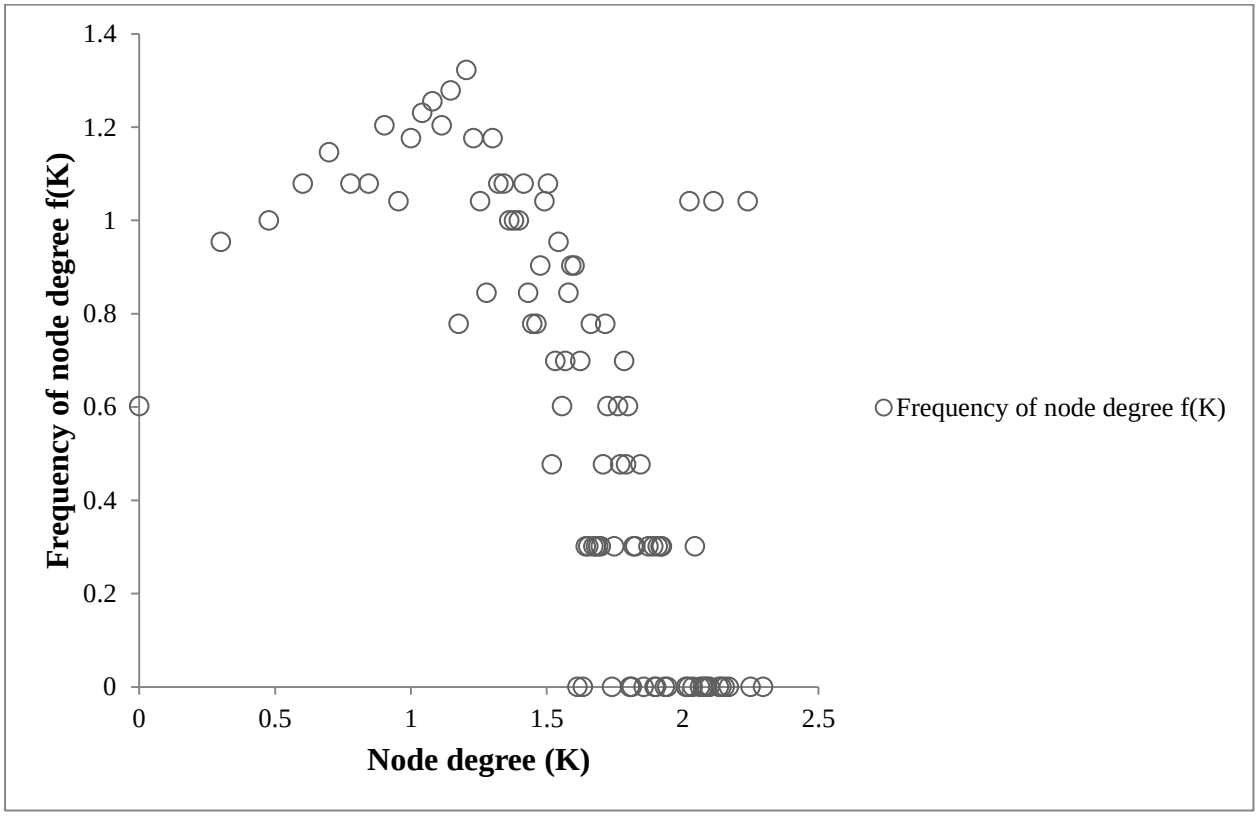


Figure 19The double log -plot of node degree distribution of YeaSTRACT dataset

The node degree distribution of the overall nodes in the dataset is observed in Figuree above. There are several nodes with degree (K) up to 300 having unit frequency. However the highest frequency of any node is 21. There are 16 nodes having the frequency of 21. The average degree of the network is equal to the degree of the most nodes $(K \approx < K >)$ (Wuchty, Ravasz et al. 2006).The nodes have the similar number of edges distributed throughout the network with very small characteristic path length in between the nodes(=). The shortest path length of the nodes with the characteristically smaller clustering coefficient are the feature small world effect in the random network(Watts and Strogatz 1998). Smaller path length is estimated by the data provided by the network analyzer of Cytoscape. The value was compared to databases. The distribution graph of YeaSTRACT is exponential in

nature in nature. The Erdos –Renyi model of random graph is observed in the YeaSTRACT dataset.

- **Outdegree distribution of Transcription factor**

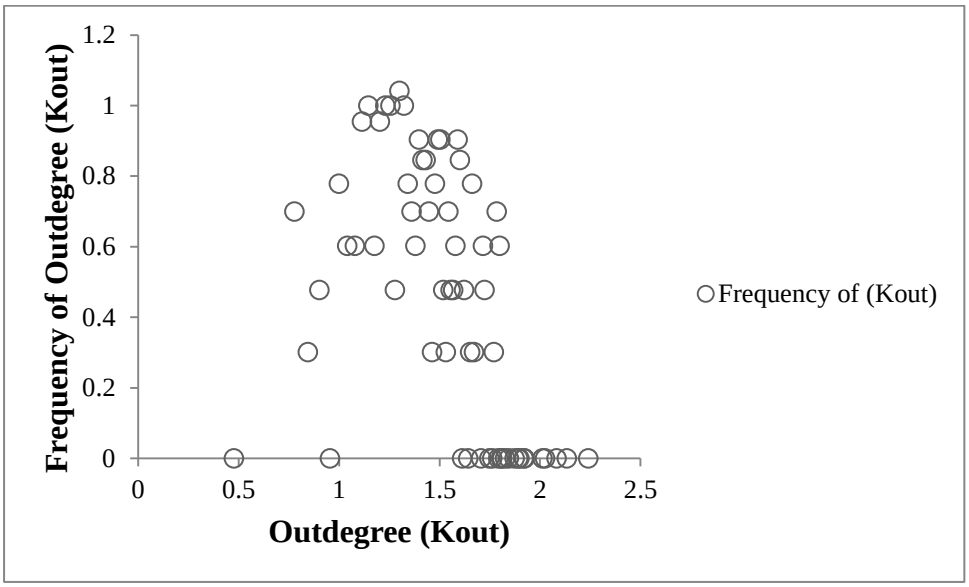


Figure 20The double log-plot of Outdegree of Transcription factor in YeaSTRACT dataset

The Indegree of target genes in the YeaSTRACT dataset has the most involving transcription factor as Ace2p with Outdegree $(Kout)$ value as 198. This protein is found to be involved in the nucleic acid binding and metal ion binding in saccharomyces cerevisae. Similarly there are other transcription factors like spf1p, ste12p, tec1p, msn2p with the out degree $(Kout)$ of 178, 148, 143, and 139 respectively. There are nodes with the out degree $(Kout)$ of around 10 with the frequencies ranging from $\log of\ 1\ to \log of\ 1.2$, that reflects the nodes of that region having average node frequency of 10. There are several nodes with similar frequency ranging from 1 to 13. Outdegree $(Kout)$ having $Kout$ = 198, for target gene Ace2p, forms a sub graph within the network, similarly there are other 15 more target genes with higher indegree, but one thing is to consider that average degree of the Indegree that happens to be around 16 is more or less equal to the Outdegree $(Kout)$ of every transcription factor in the dataset $(K \approx < K >)$. Thus recommending the network formed by transcription factors in the YeaSTRACT is also random network.

- **In-degree disribution of Target genes**

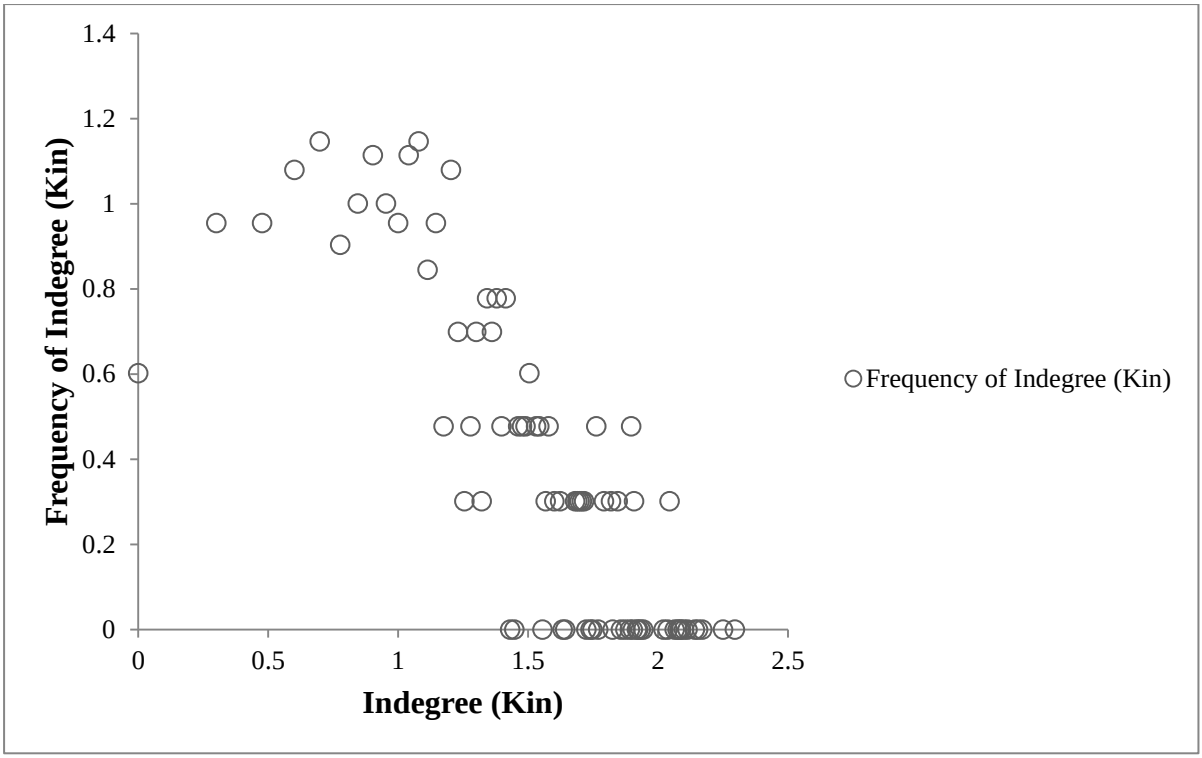


Figure 21The double log-plot of in degree distribution of target genes in the YeaSTRACT dataset.

The Indegree (Kin) distribution of target gene in YeaSTRACT shows some target genes with relatively higher indegree count. The target gene with highest indegree count is for PHO5 equaling 174; similarly GPX2 has higher indegree of 136. There are 5 more target nodes with Indegree value higher than 100. It appeals to form regulatory hubs in the network. But after looking at the pattern of distribution of target genes in the network, only the distinctive peak is observed there is no uniform flow of node distribution as for scale free networks. The double log plot of target genes of the YeaSTRACT dataset is a random network.

3.1.2.3 ESCAPE dataset

The graph is made from the data obtained from network analyzer of Cytoscape. The nodes representing target gene is higher 23047 compared to 61 transcription factor, the overall node is 23109.

- **Node degree distribution of Undirected Network**

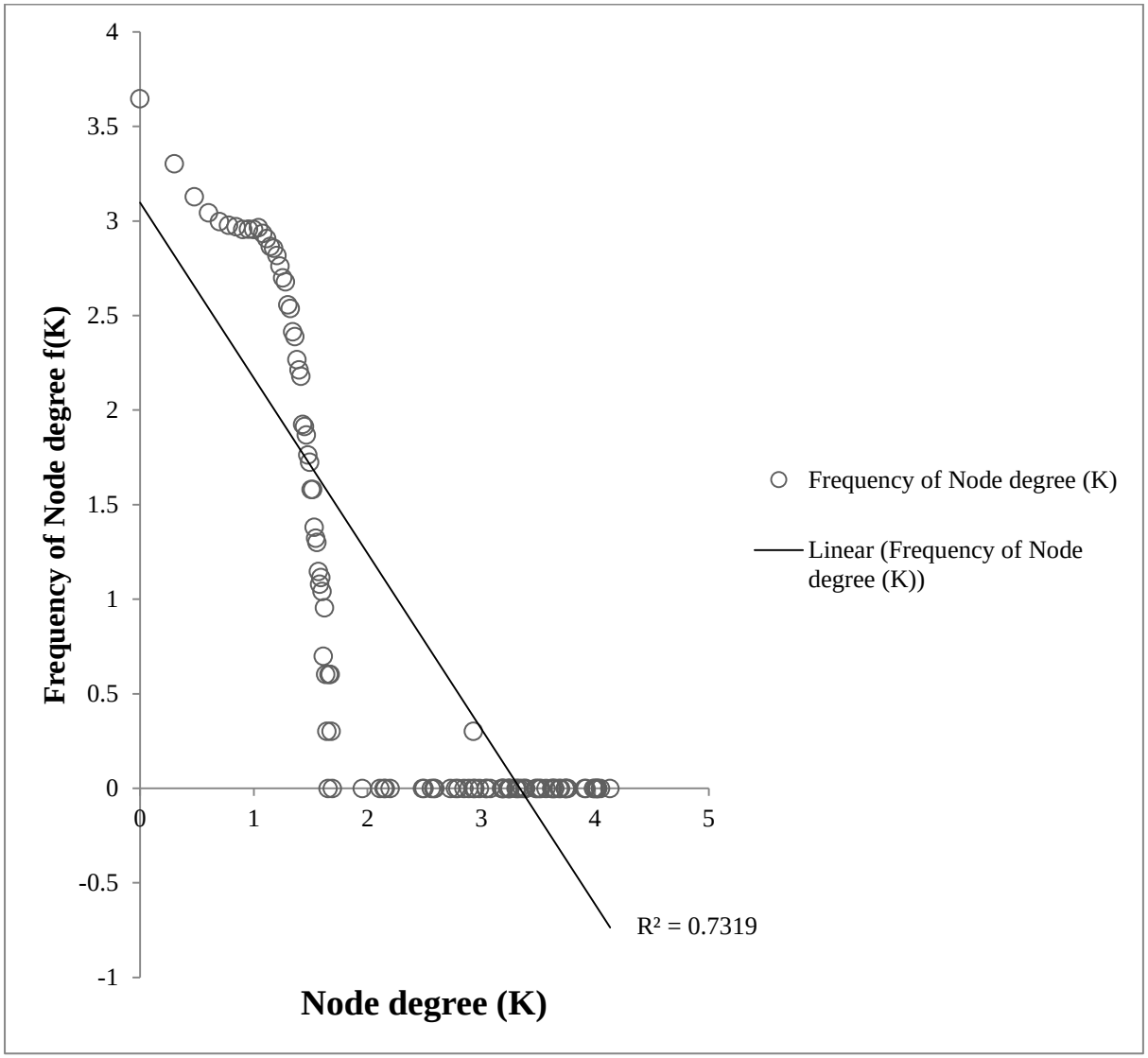


Figure 22The double log-plot of node degree distribution of undirected network of ESCAPE dataset.

The node degree (K) distribution of undirected network of ESCAPE dataset reveals the particular pattern of nodes distributed over the whole network. It gives the information what is the pattern of interaction between the different Tf-Tg in whole. There are around 42 different nodes with the node degree above 1000. The node which has the highest nodes degree (K) is MYC, it is 13,566. There are other 6 transcription factors which have the massive size in node (K) degree ranging above 10,000. It is observed that nodes with the degree (K) up to 80 have the higher frequency of above 1000 to 3500. These nodes are found predominantly in the network created by ESCAPE dataset to be of modular structures within. There are number of functional module in the network giving it a hierarchical shape.

The negative slope for the frequency of nodes with node degree reconfirms it. The degree coefficient (γ) of 0.928, with correlation of 0.969 for this network was observed.

- **Outdegree distribution of Transcription factor**

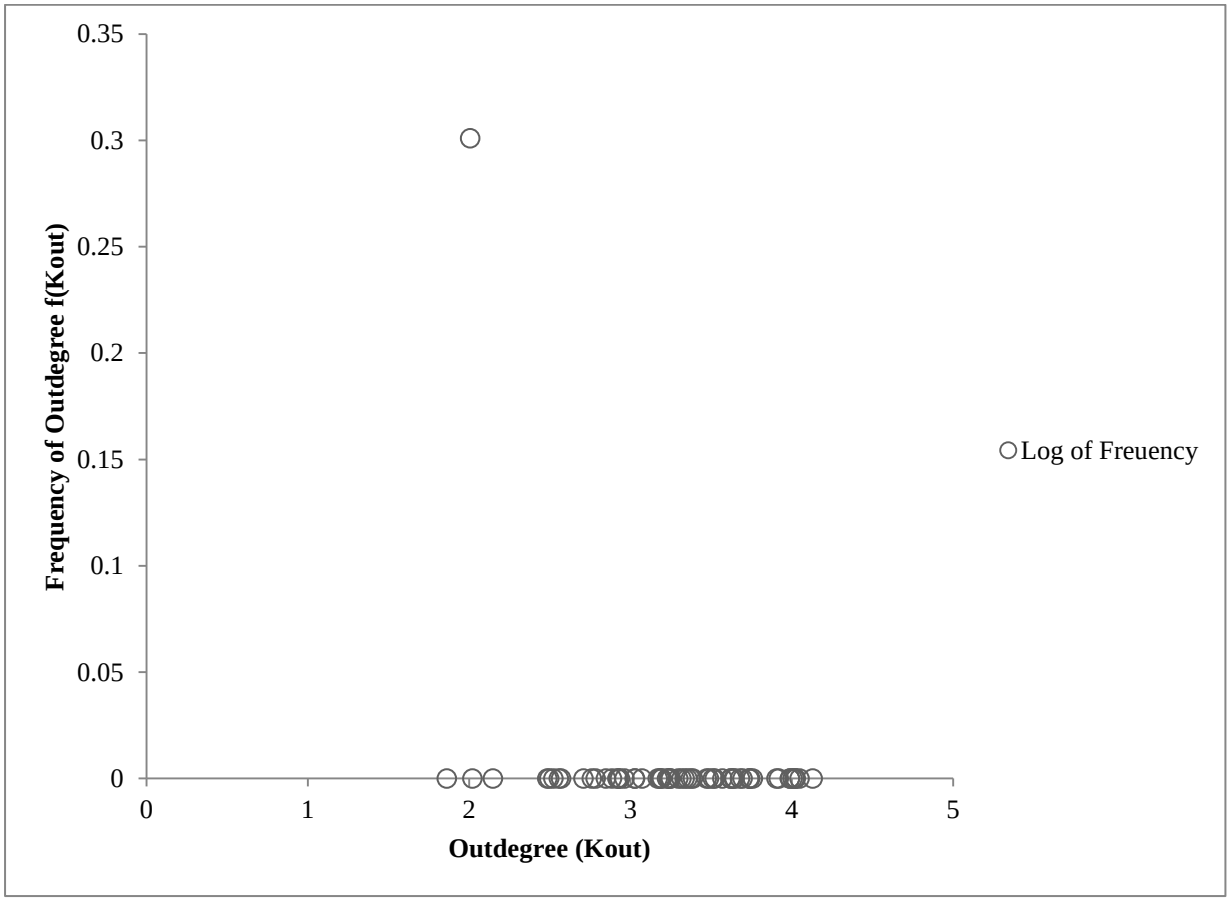


Figure 23The double log-plot of out degree distribution of transcription factor in ESCAPE dataset.

The distribution of transcription factor in the ESCAPE dataset is dominated by the transcription factors having high out degree ($Kout$) ranging from 73 to the count of 13,537. The interesting pattern of distribution is observed in the transcription factors of ESCAPE that those nodes have the similar frequency of around 1. The transcription factor with higher out degree is observed but the frequency of observation is similar. In total only two transcription factors KLF5 and KLF2 which have Outdegree ($Kout$) equal to $\log 2$ and are found in single number, whereas in terms of distribution of nodes there are two nodes with similar Out degree of 102, thus their frequency is plotted in 2. The pattern of distribution of transcriptional factors shows the small world property of the random graph in this graph.

- **In-degree Distribution of Target genes**

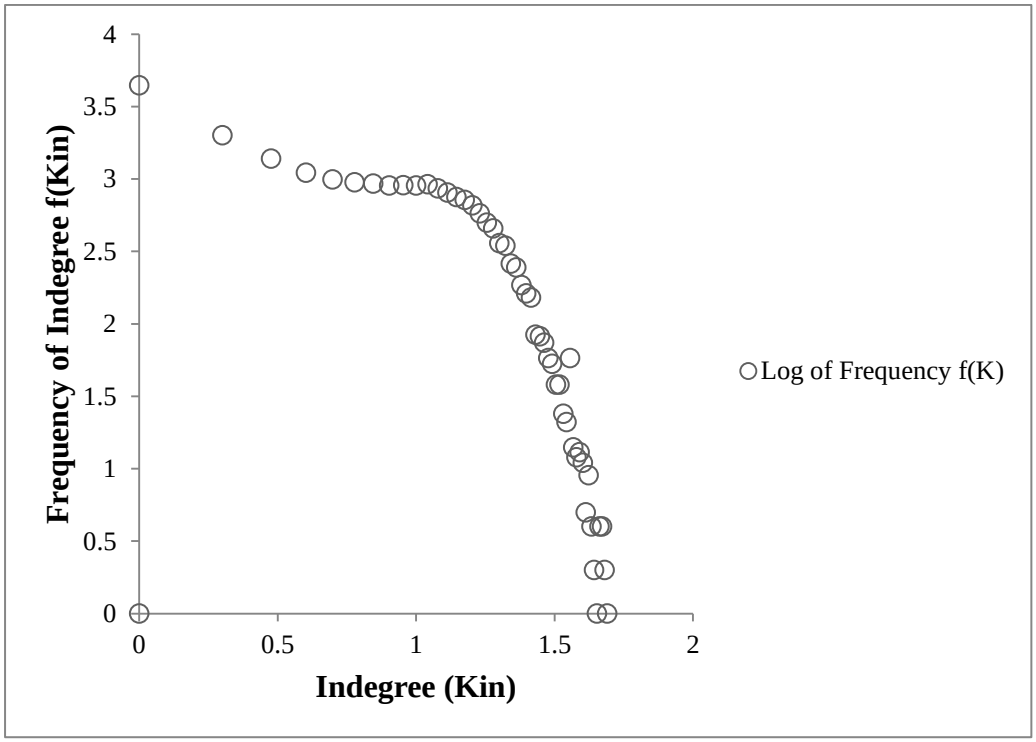


Figure 24The double log-plot of in degree of target genes of ESCAPE dataset

The distribution of target genes in the ESCAPE dataset is determined by the target gene where mostly the single target gene is involved in transcription. There pattern of distribution of the target gene do not have significantly highly involving target gene, instead there are numbers of target gene of Indegree (Kin) with higher frequency even though they have high Indegree. The total of 4438 target gene is observed with Indegree (Kin) of 1.The highest for target gene Indegree is found 49 for DIDO1.

The graph of Indegree (Kin) of target gene has similar pattern of distribution. The graph follows the scale free network with$\gamma = 2.10, corre = o.87$. The indegree for target genes of ESCAPE dataset has the scale free nature.

6.2 Clustering coefficient distribution

For the first time standard Clustering coefficients (C) was used by Watts and Strogatz that clustering coefficient C is the parameter to characterize the local transitivity in any social network(Wei, Wu et al. 2006). It is one of the global parameter which is used to characterize the topology of complex network. The distribution of clustering coefficient is the characteristic of the different nodes in the network to form several triangles with their neighbors. In calculation a triangle between enables one to estimate the clustering coefficient

distribution from 0 to 1. So the average clustering coefficient measures the global density of interconnected nodes in the network(Ng, Wei et al. 2005).

The tendency of forming cluster in the network is enunciated by the property of the scale free network where they have preferential attachment. The preferential attachment is the addition of new node to already formed cluster of the network instead of establishing itself alone. The property of preferential attachment determines many vital property of biological network. Preferential attachment in biological network can be explained by gene duplication. During evolution when the gene is subjected to duplication, it leads to the duplication of protein it encodes. This gene duplication preserves the interactions, thus newly duplicated protein get interacted with all the neighbor of the original protein. This leads to the preferential attachment, as more macromolecules gain more new neighbors this way(Jeong, Néda et al. 2003).

It is the measure of cohesiveness in the network, termed by the clustering coefficient of that network. The clusters in the network form the subset within that displays a higher level of inner connectivity. It is well observed that the value of clustering coefficient of the node is its likelihood to get connected with the nodes nearby in its neighbor, are eventually connected to each other if the cluster is formed. For example in a village two villagers who have common acquaintance is more often known to each other. The clustering coefficient is the local property that describes the network structure of nodes that are close to each other.

The clustering coefficient of the scale free network is significantly higher, since it is expected in the scale free network for a node to get connected to each other significantly. The empirical analysis of the network based on the Barabasi – Albert model shows that the clustering coefficient C of the network decreases with the increasing network size. It thus follows power law($C \sim N^{-\gamma}$). Finally it converges to the 0 when we increase the number of neighbors.

The clustering coefficient distribution of the different dataset is plotted

A) HTRIdb

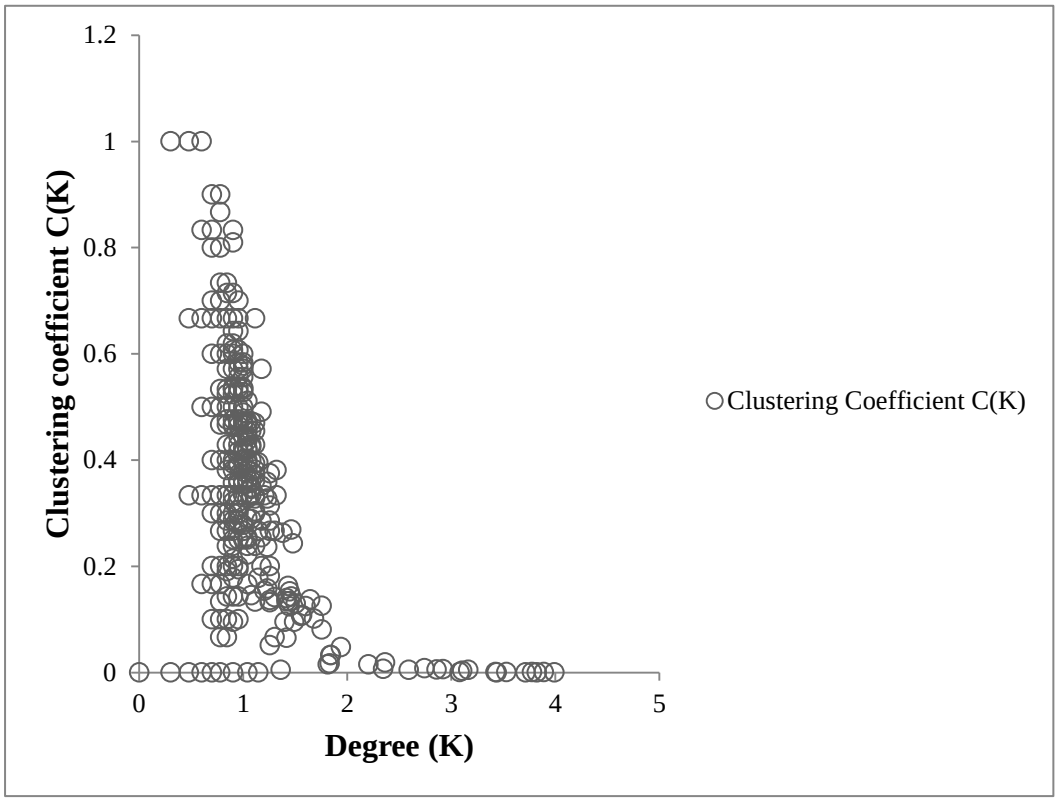


Figure 25Semi-log clustering coefficient distribution of the HTRIdb dataset

Clustering coefficient distribution of the nodes in the HTRIdb dataset makes the relative distribution of the nodes with some clustering effect. The node which has higher clustering coefficient value is supposed to be forming a clique of nodes within and transfer the information subsequently. The total of around 649 nodes has the clustering coefficient value of 1. However there are also some nodes which do not form any cluster in the network. Since the graph that is formed after the distribution analysis of clustering coefficient has the negative slope, with the increasing degree (K) clustering coefficient (C) decreases. The hierarchical architecture of the dataset is confirmed by several nodes that are having degree ranging up to 9759.

The interesting observation reveals that only target genes have the clustering coefficient of 1, by forming 649 triangle of transmission within the network.

B) SIGNLESS TFactS

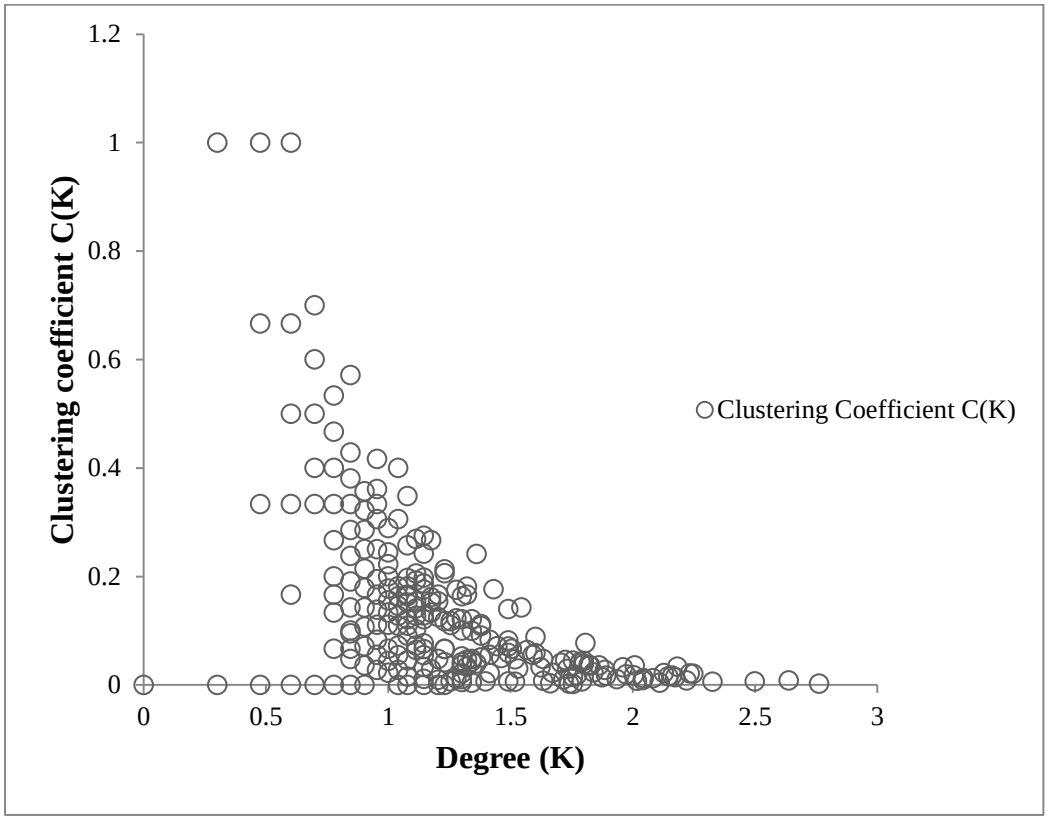


Figure 26Semi-log clustering coefficient distribution of Sign less TFactS dataset.

The clustering coefficient distribution of the Sign less TFactS dataset has 30 target genes forming 30 separate triangles having the 1 value of clustering coefficient. The architecture that clustering coefficients determines for this network is of hierarchical nature. Basically the decay of the clustering values from the highest 1 to the least 0 has the slope forming to elucidate some reasoning behind, the negative slope ranges from the nodal degree 1 to 579.

C) **SIGNSENSITIVE** **TFactS**

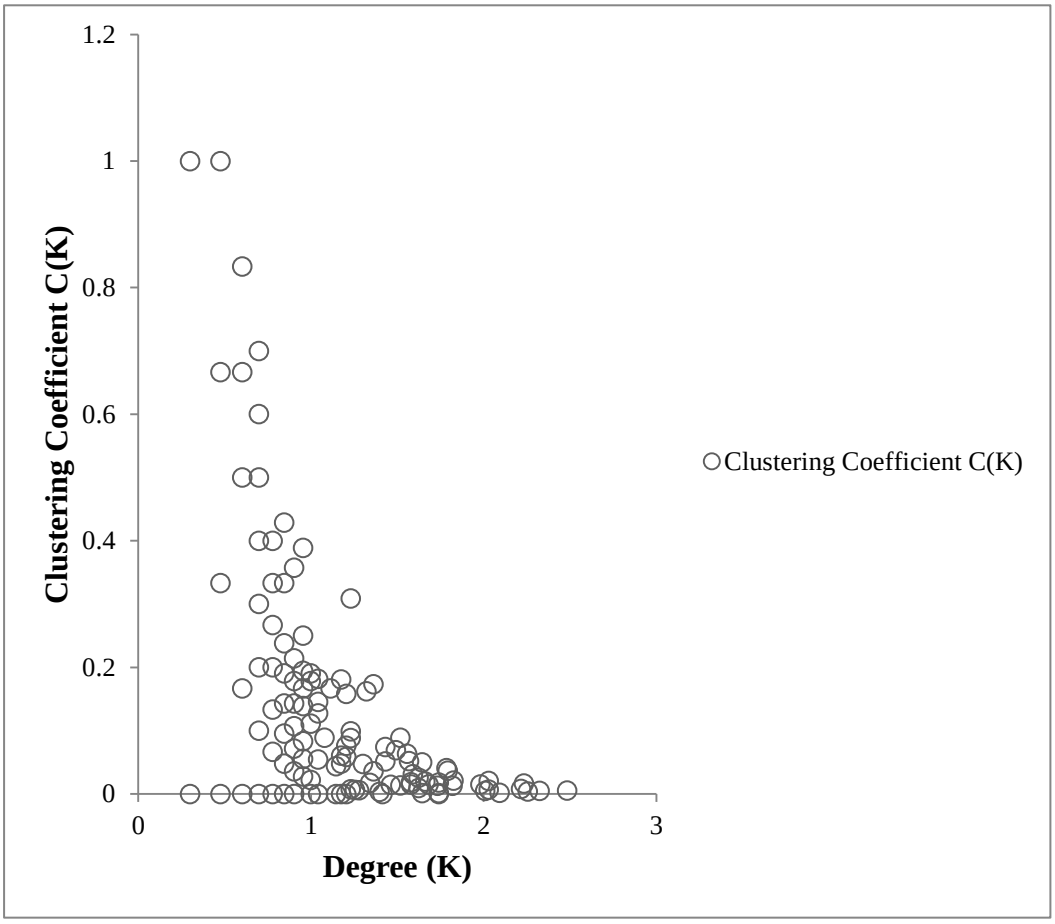


Figure 27The semi-log clustering coefficient distribution of Sign sensitive TFactS dataset

The sign sensitive dataset of the TFactS has fewer nodes forming complete clique of three nodes together, but this time transcription factors are forming the triangle valued 1. The pattern of distribution of clustering coefficient with corresponding degree has the hierarchical architecture also for the Sign-sensitive TFactS dataset.

D) YeaSTRACT

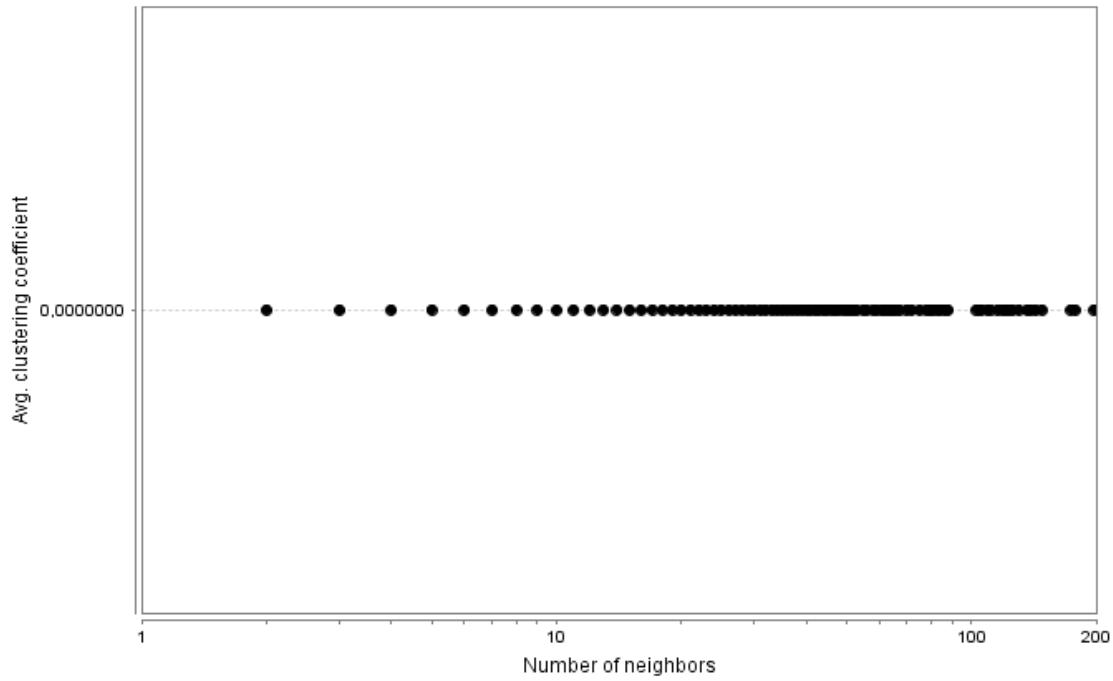


Figure 28The semi-log Clustering coefficient distribution obtained through Network Analyzer of Cytoscape (4/07/2014)

The clustering coefficient pattern observed in YeaSTRACT dataset has inference to the random model as described by Erdyos and Renyi (RM). The clustering coefficient in the dataset is independent to the degree; on the other hand this independent network has the feature of uniform distribution of clustering coefficient all over the network. However the graph show that the line created for the distribution of clustering coefficient to reconfirm its network architecture.

E) **ESCAPE**

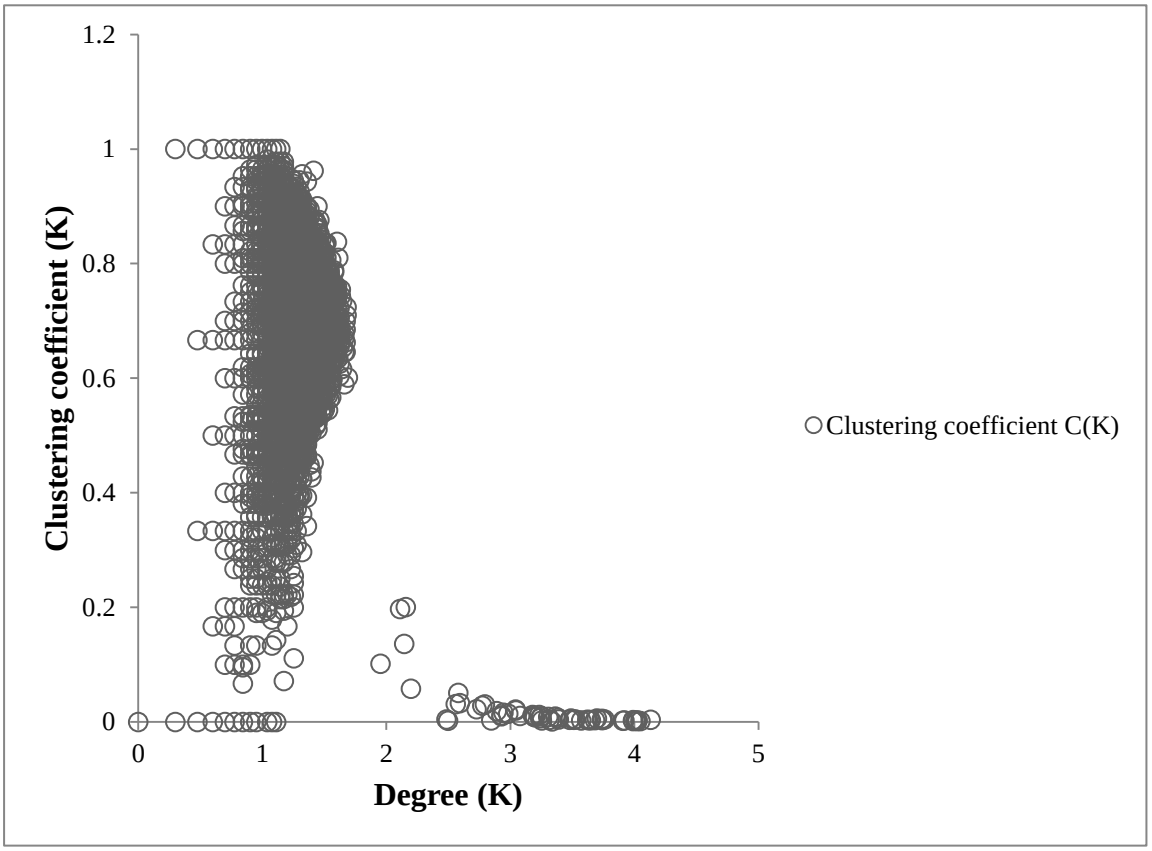


Figure 29The semi-log clustering coefficient distribution of ESCAPE dataset

It is observed that total of 608 triangles with clustering coefficient value 1, appears in ESCAPE dataset. Target genes are involved in forming triangle. The prediction of this huge dataset, with varying range of clustering coefficient, forms scale free modular hierarchical architecture of the network. There are nodes with high nodal degree (K) but very minimal clustering coefficient.

3.1.3 Sub network distribution

The sub network obtained from the different dataset will be analyzed in this section. To obtain the different sub networks from the various networks Cytoscape Application named MCODE is used. It is the graph based clustering algorithm applied to detect sub networks in an interactome(Li and Kim). MCODE is a Molecular Complex Detection application which is an automated method for finding molecular complexes in large protein interaction networks (Bader and Hogue 2003). The recurring regulation pattern of network motifs, were first studied in *Escherichia coli*, where detected pattern occurred in transcription network with higher frequency i.e. more than expected compared to random networks(Alon 2007). The different motif obtained allows one to gain easy interpretation of the entire known

transcriptional network of any organism(Shen-Orr, Milo et al. 2002). The motifs can be considered as the basic building blocks of many networks including transcription regulatory network. Each motifs of a network can carry out specific information processing function. In addition the motifs are also very much useful in making us understand the densely connected network of higher organisms(Alon 2007).

An example for representation of different motifs in regulatory networks.

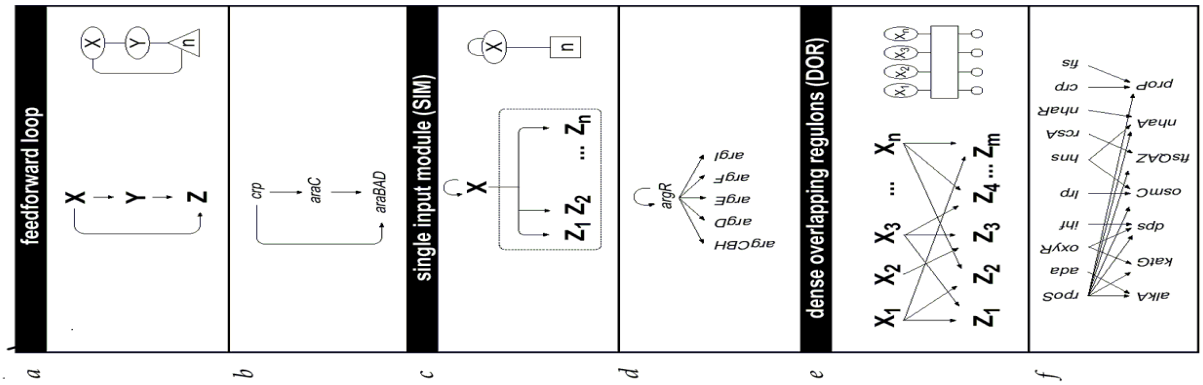


Figure 30The network motifs obtained from E.coli transcriptional regulation network reprinted from (Shen-Orr, Milo et al. 2002)| a) The motif structure called feed forward loop is shown: a transcription factor X regulates second transcription factor Y, and both jointly act on the operons Z1, Z2….Zn, its example (b) is shown as L- arabinose utilization. In Single input motif (SIM) single transcription factor acts on the set of operons Z1,Z2,..Zn and X is auto regulatory. The arginine biosynthesis (d) is an example of SIM. Dense overlapping regulon (DOR) motifs are the combinatorial mode of regulations where Z1…Zm regulates the set of transcription factors X1…Xn. These motifs can be illustrated by algorithms that detect dense regions of connections, with high ratio of connection to transcription factors. F) Stationary phase response is taken as example of DOR.

The appearance of different motif has the functional significance in the network. They carry information processing functions. Transcription network in the Eukaryotes is represented as directed graphs where nodes are the transcription factors and target genes and the edge is the interaction between them. The MCODE algorithm used to detect the different recurring patterns within the network is evaluated in this report.

The different pattern of motifs that appeared in the datasets are;

a) Simple regulation:

   Simple regulation is the basic regulation without involving additional interaction between transcription factor and target genes.

   - Negative auto regulation:

     Negative auto regulation (NAR) is observed when the transcription factor represses the transcription of its own gene

   - Positive auto regulations

When transcription factor enhance its own rate of production Positive auto regulation (PAR) occurs.

b) Feed forward loop

The motif consisting three genes i.e. a regulator, X, which regulates Y, and gene Z, which can be regulated by both X and Y. Eight different type of feed forward loop can be observed in transcriptional regulation(Shen-Orr, Milo et al. 2002).

c) Single Input Module (SIM)

Another family of motifs which is patterned like a regulator regulates group of target genes. In idealistic sense, no other regulator is involved in interaction with target gene. Regulators also auto regulate themselves.

d) Dense Overlapping Regulon (DOR)

This family of motif is a bit complex. The set of regulators combinatorial control set of output genes. These motifs are referred to as dense overlapping regulon (DORs) or multi-input motifs (MIMs).

The sub network was obtained by the MCODE application of Cytoscape. In the default parameter of MCODE, with fluff and loops not included, degree cutoff as 2, Haircut was included with node score cutoff equaling 0.2.

A) Human Transcriptional Regulation Interaction database (HTRIdb)

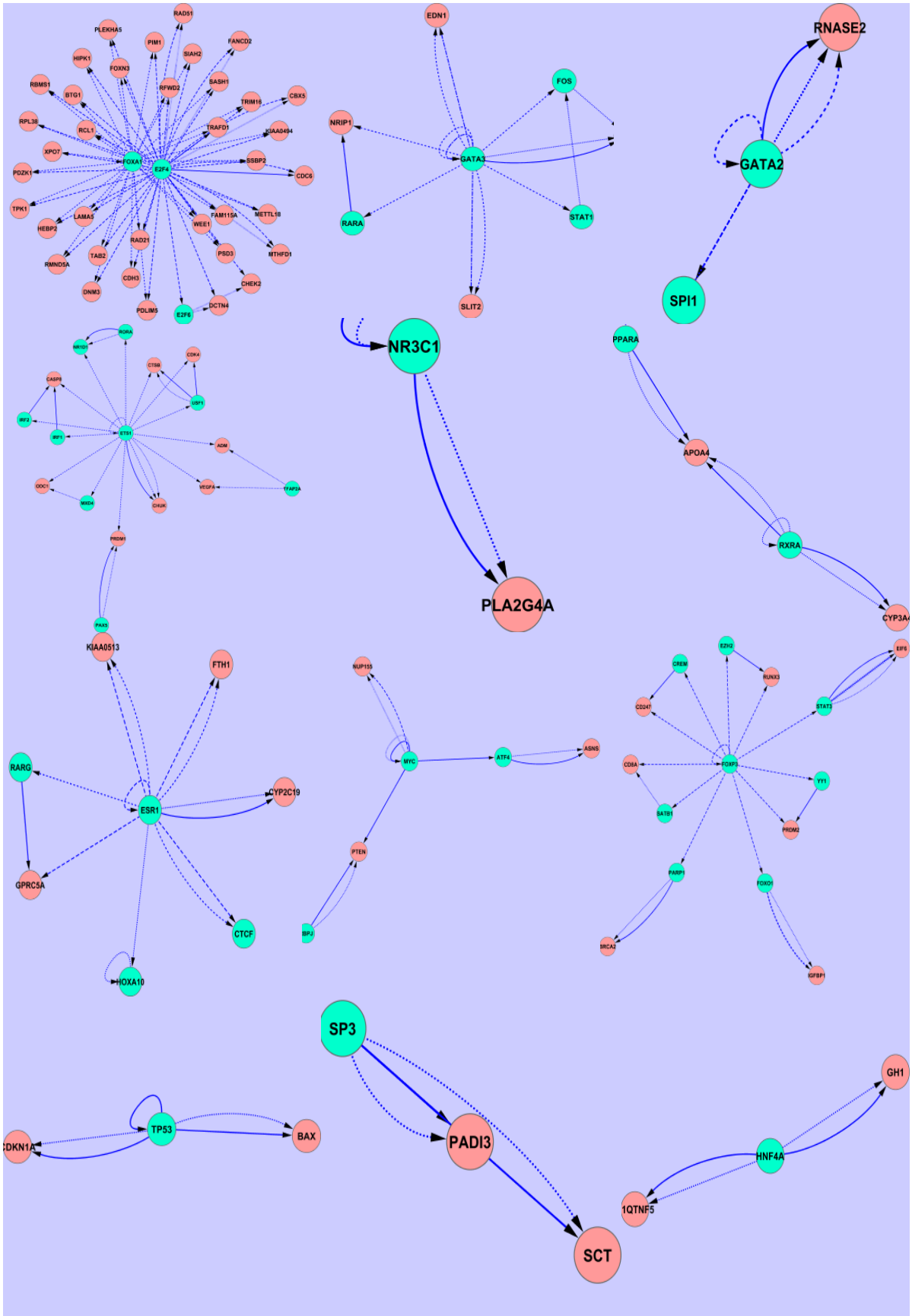
The HTRIdb dataset is subjected to MCODE analysis. The node and edge attributes of the network visualized under Cytoscape was primarily like transcription factor is colored fluorescent green and target gene is in default light pink. The edge attributes were basically worked on to make the techniques separable. In the table below techniques along with edge attributes mentioned:

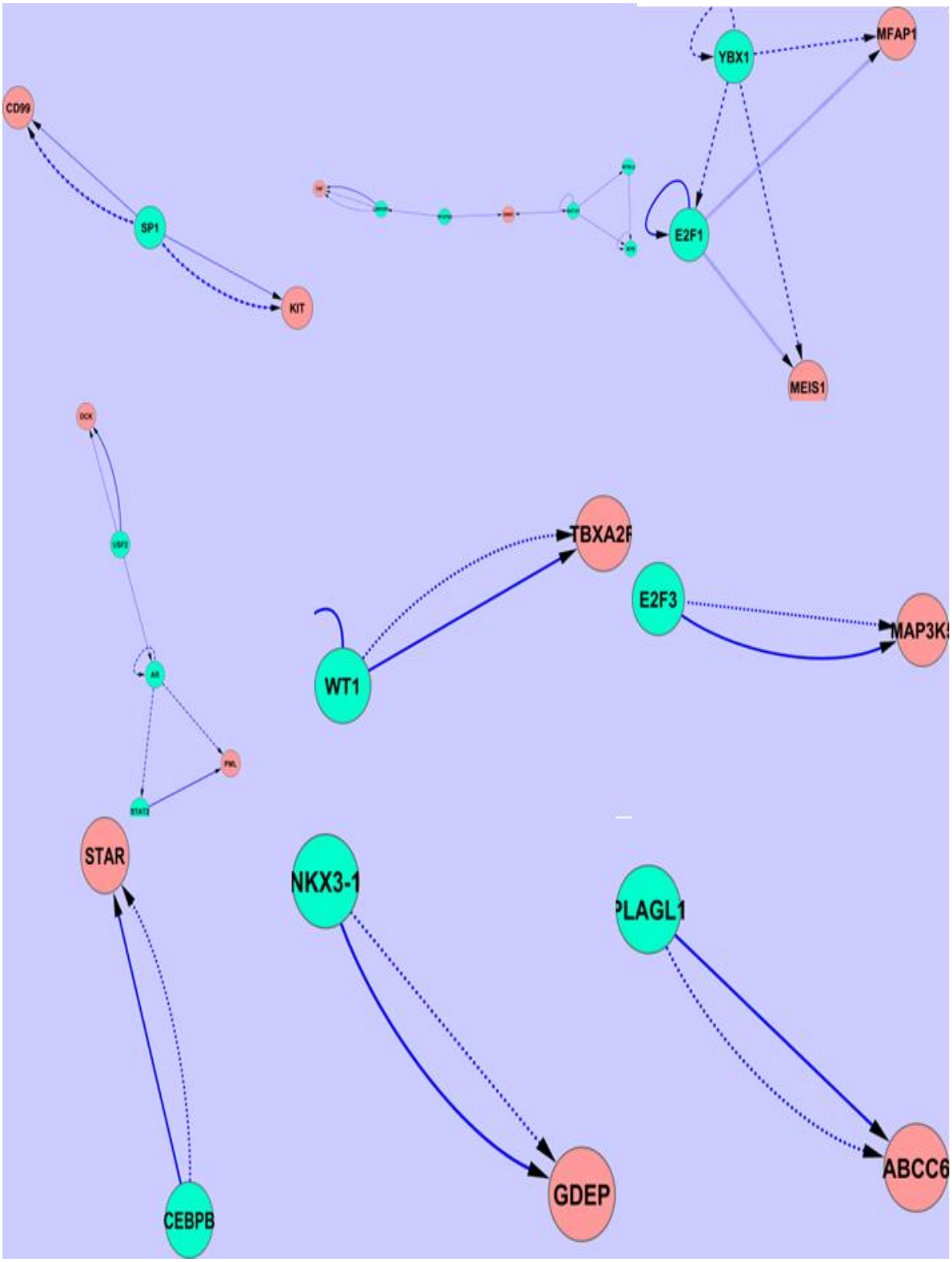
Table 9The Edge line attribute for techniques of HTRIdb in Cytoscape.

SN.	Techniques	Edge line attribute
I	Electrophoretic mobility shift assay	Solid
II	Chromatin Immunoprecipitation	Solid
III	ChIP-chip	Long dash
IV	ChIP-seq	Equal dash
V	DNA affinity chromatography	Zigzag
VI	CpG chromatin immunoprecipitation	Vertical slash

VII	DNA affinity precipitation assay	Contiguous arrow
VIII	Yeast one hybrid	Dot
IX	Southwestern blotting	Dot
X	Surface Plasmon Resonance	Dot
XI	Avidin-biotin conjugated DNA	Parallel line
XII	Streptavidin ChIP	Dot
XIII	DNase foot printing	Surface Plasmon Resonance
XIV	Concatenate ChIP	Dash Plot

The different sub networks are obtained from the HTRIdb dataset.



3.1.3.1.1 
CD99
SP1
KIT
YBX1
MFAP1
E2F1
MEIS1
WT1
E2F3
STAR
CEBPB
NKX3-1
GDEP
ABCC6

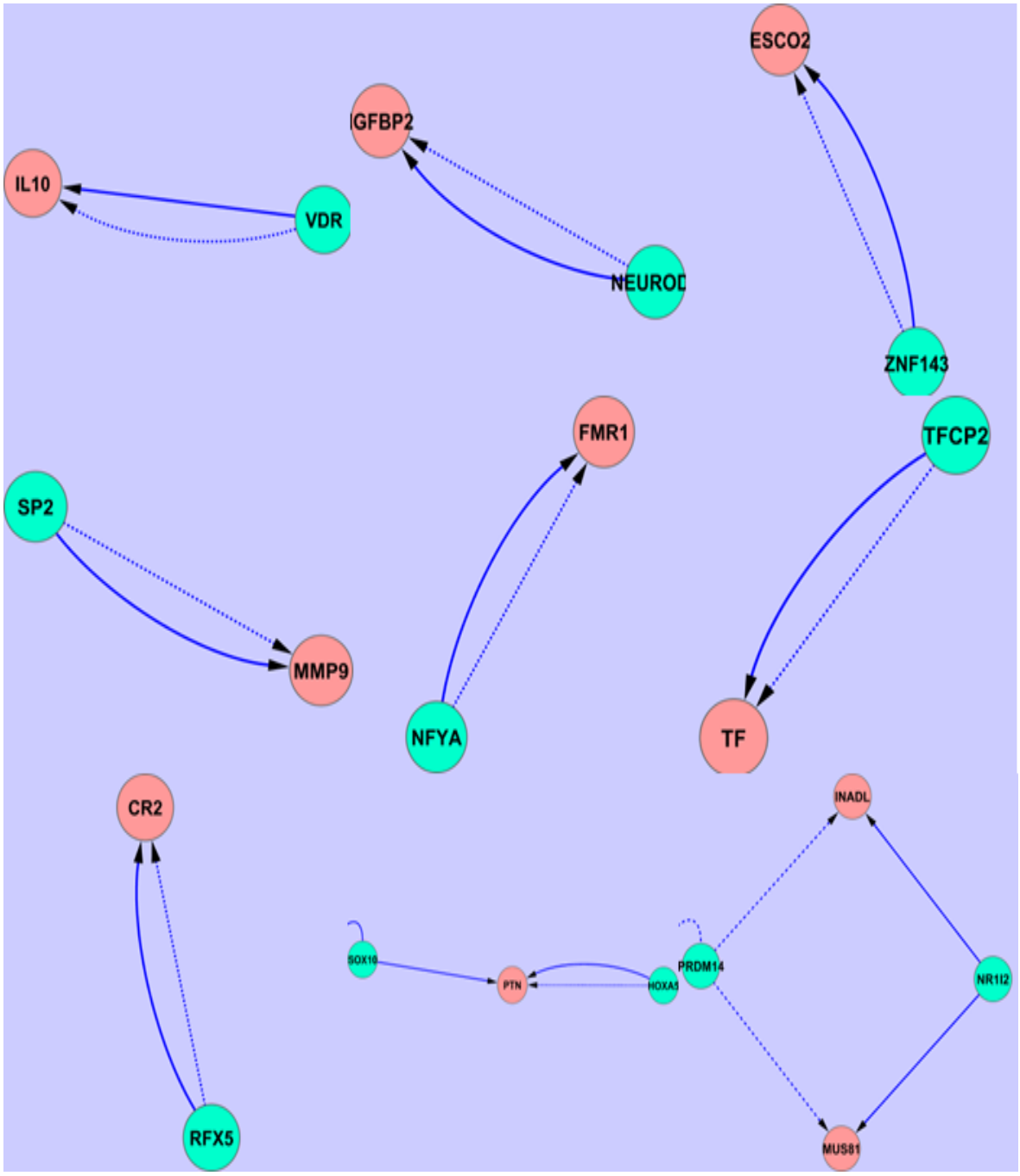


Figure 31The Subnetworks visualized by the Cytoscape, subnetwork are created through MCODE for HTRIdb dataset. The fluorescent green are the transcription factors and pink are the target genes. The motif count starts from top left of every column up to right.

Table 10The Sub networks for HTRIdb dataset from MCODE

cluster	Type	Transcription factors	Target gene	MCODE score	Node/Edge
1	SIM	3	37	1.875	40/76

2	DOR	4	4	1.75	8/15
3	FFL	2	1	1.667	3/6
4	DOR	8	9	1.647	17/29
5	SR	1	1	1.5	2/4
6	DOR	2	2	1.5	4/8
7	SIM	2	6	1.5	8/14
8	DOR	3	3	1.5	6/10
9	SIM	8	7	1.467	15/23
10	FFL	1	2	1.333	3/5
11	FFL	1	2	1.33	3/4
12	FFL	1	2	1.33	3/4
13	FFL	1	2	1.33	3/4
14	DOR	5	2	1.286	7/11
15	SIM	2	2	1.25	4/7
16	DOR	3	2	1.2	5/7
17	SR	1	1	1	2/3
18	SR	1	1	1	2/2
19	SR	1	1	1	2/2
20	SR	1	1	1	2/2
21	SR	1	1	1	2/2
22	SR	1	1	1	2/2
23	SR	1	1	1	2/2

24	SR	1	1	1	2/2
25	SR	1	1	1	2/2
26	SR	1	1	1	2/2
27	27	1	1	1	2/2
28	SR	1	1	1	3/4
29	FFL	2	1	1	4/5
30	SIM	2	2	1	1/4

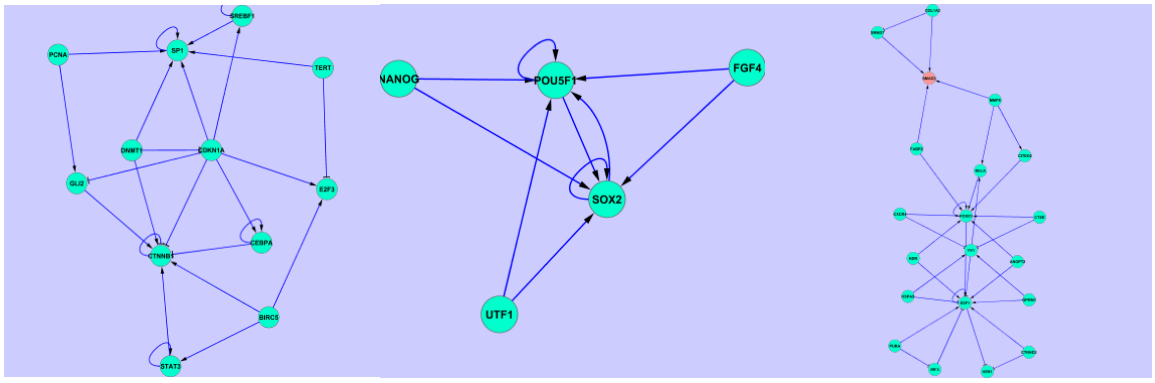
Overall 30 different sub networks was obtained with three families of motifs, Feed forward loop (FFL), Single Input modules (SIM), Dense over lapping regulon (DOR)

B) Sign Sensitive TFactS

The Sign Sensitive TFactS dataset was used to obtain sub networks within the network, keeping the default parameters of the MCODE. The transcription factor is fluorescent green and light pink is target gene.

The edge attribute, referring to the species from which the regulatory association between transcription factor and target gene is mentioned below:

The total of 8 sub networks was obtained:



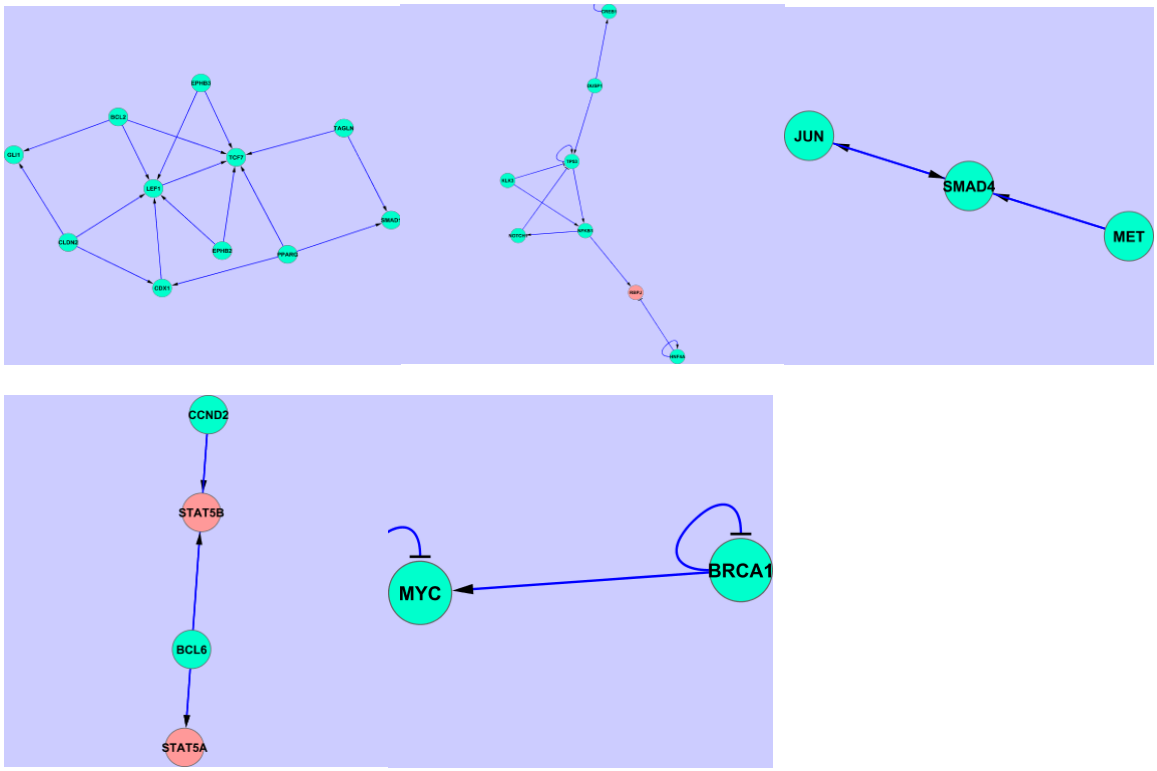


Figure 32The different sub networks visualized through Cytoscape, sub networks are created using MCODE for Sign sensitive datasets. The green fluorescent nodes are the transcription factor and pink are the target genes. The Arrow points the up regulation whereas the T shape arrow is for down regulations. The motif count starts from top left of every column up to right.

Table 11The table is the list of motifs from TFactS sign-sensitive dataset

Cluster	Type	Transcription factor	Target gene	MCODE	Node/Edge
1	DOR	12	0	1.667	12/25
2	SIM	5	0	1.6	5/10
3	DOR	19	1	1.55	20/3
4	DOR	11	O	1.545	11/17
5	DOR	6	2	1.125	8/12
6	FFL	3	0	1	3/3
7	SIM	2	2	1	4/4
8	SR	2	0	0.5	2/3

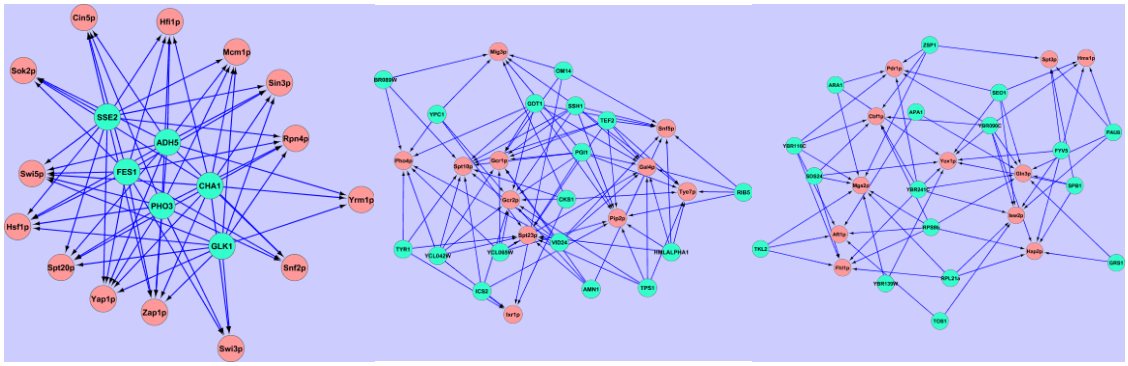
Table 12:.

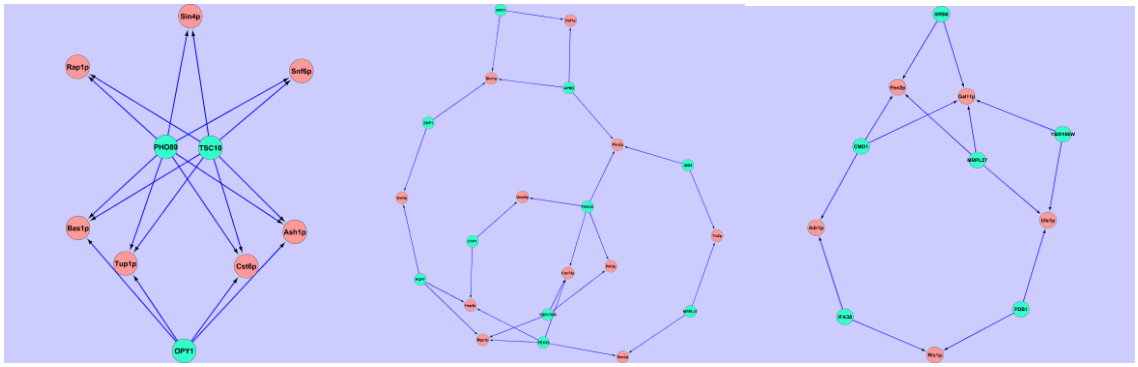
Overall 8 different sub networks was obtained with three families of motifs, Feed forward loop (FFL), Single Input modules (SIM), Dense over lapping regulon (DOR)

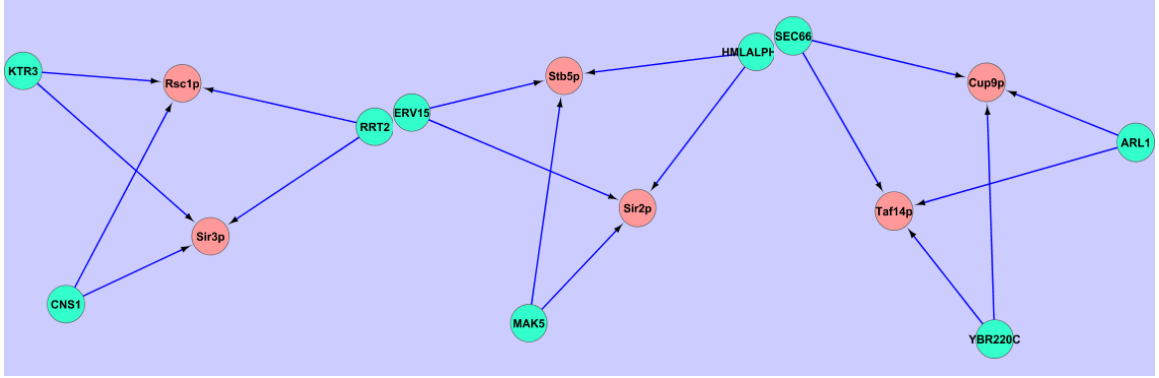
C) YeaSTRACT

Under the default condition of MCODE application, the YeaSTRACT dataset is applied through MCODE algorithm to obtain sub networks within. The transcription factor is fluorescent green and target gene is light pink.

The different sub networks are obtained from YeaSTRACT dataset.







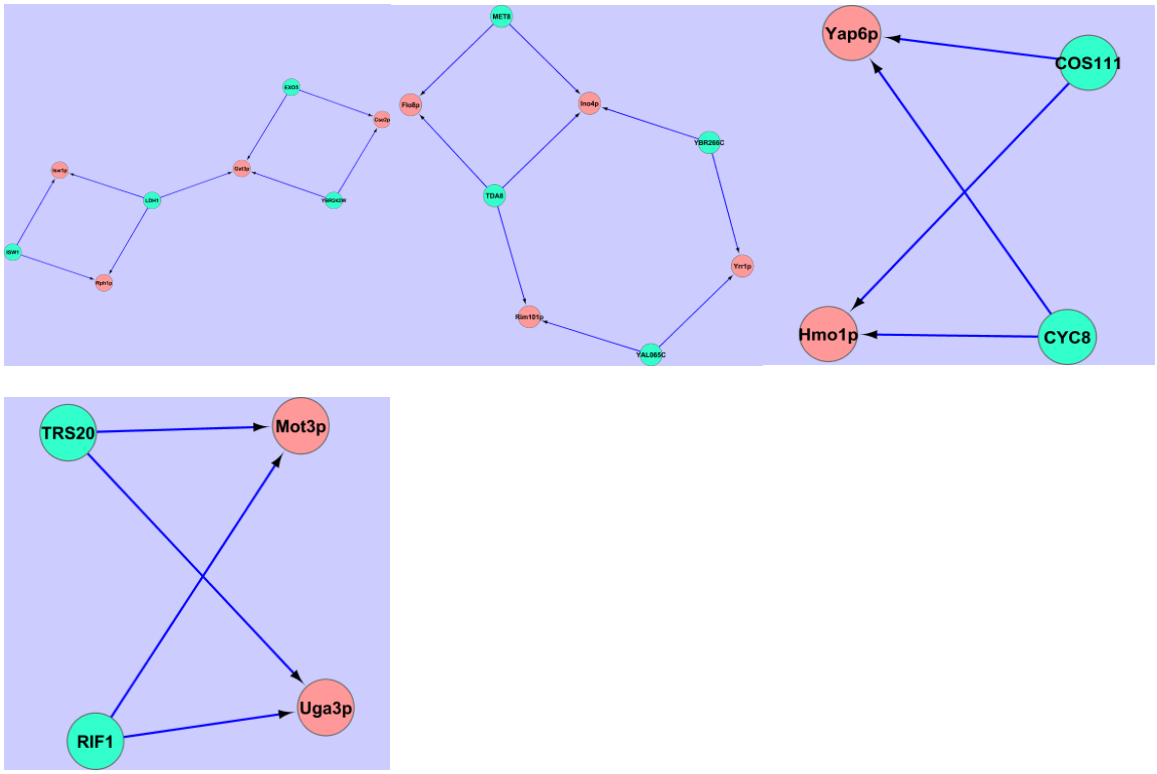


Figure 33The different sub networks visualized through Cytoscape

The topmost left to right is motif no 1-3, similarly second row of images has the same pattern, left to right is 3-6, 6-9, 9-12 and 13th motifs with fluorescent green as transcription factor and pink as target genes. The motif count starts from top left of every column up to right.

Table 12The sub networks of the YeaSTRACT dataset from MCODE

Cluster	Type	Transcription	Target	MCODE	Node/Edge

		factor	gene	score	
1	SIM	6	14	3.35	20/67
2	SIM	17	11	3.214	28/90
3	SIM	17	11	2.429	28/68
4	SIM	3	7	1.8	10/18
5	SIM	10	11	1.286	21/27
6	SIM	6	5	1.273	11/14
7	SIM	3	2	1.2	5/6
8	SIM	3	2	1.2	5/6
9	SIM	3	2	1.2	5/6
10	SIM	4	4	1.125	8/9
11	SIM	4	4	1.125	8/9
12	SIM	2	2	1	4/4
13	SIM	2	2	1	4/4

For YeaSTRACT dataset, the 13 motifs are observed but all of them with the same motif family type; Single input motif (SIM).

4 Discussion:

Despite the glorious approach of reductionism in biology, the alarming necessity for uncovering the absolute molecular world with the approach of system biology has become a new frontier in biology. System biology is inherently a study of mechanisms of underlying complex biological process as an integrated system of many interacting components (www.biocomp.unibo.it/piero/MSB/**IntSB.ppt**). An example of large Zoological and botanical expedition at the end of nineteenth century led to characterization of organism diversity and their relations. In the same way, molecular biologists are exploring the canvas of diversity inside the cell. The similar pattern of exploration was observed after the elucidation of biochemical pathways in the early-middle twenty centuries provided more complete picture of genes, proteins and metabolites by the initiating research on molecular biology.

Even though in this period of high through put researches or biotechnological advancements, the current understanding of system is only a sketch of actual relation between elements, whereas the most of biological details are still in the state of oblivion. With increasing number of research, exponentially increasing data flows towards the motive of scientific community in addressing the lacking biological details. The advent of genomic, proteomic and metabolomics approach which has offered the required ingredients for the revolution of molecular biology and biomedical facet of science and emergence of novel approach of system biology that are meant on unravelling the mystery of biology. Definitely, the approach in graph theory of system biology for the real system has been applied to append the understanding on relation between the elements of the system. The graph theory opens the possibility for global understanding of the system, against the predominant reductionism idea on the current scientific explorations.

Graph theory enables a systemic study through statistical calculation of the local interactions; however the global understanding is delimited by its sample size. By this we can estimate, in any statistical approach, the larger size of data is directly proportional to the reliability of the statistics. Basically the graph theory is a highly developed field of mathematics, which has from past three centuries been an important means to describe properties of networks. In very explicit language for understanding a network, networks can also be described as nothing more than set of discrete elements(the vertices) and set of connections (edges) that link the elements, typically in pairwise fashion(Cline, Smoot et al. 2007). This concept is rather stochastic to define a elements´ interaction to form a random network (Shen-Orr, Milo et al.

2002) therefore rather deterministic approach is recently fed into network science. So during 1999, Albert –Laszlo Barabasi, Hawoong Jong and Reka Albert came up with the idea of scale free network. The basic hypothesis of their model focused on the probability distribution. The new connections get attach to a node with probability proportional to existing number of connection. The growth and preferential attachment are the mechanisms which are prevalent to a number of complex systems(Jeong, Néda et al. 2003).

Thus obtained graph theory leads to the elucidation of all these datasets by forming a network. In a network for cellular system, cell acts as `node` in the network of molecules which are connected by the biochemical reactions as `edge`. The network is ubiquitous in nature. Societies are the network of people linked by friendships, family and professionals as relation. In large scale, food webs and ecosystems can also be represented as network. Network is present everywhere as mentioned earlier. The language we speak can also be represented as network where words are connected by syntactic relationships. Network has its presence pervading in technology, internet network, power grid and transportation networks are some of its examples.(Walhout 2006).

The network analysis in different dataset is primarily performed in this thesis to observe the pattern of distribution of transcription factor and target genes; and their clustering coefficient distribution in the network. The different sub networks that are formed in the network are also studied which enables one to access the overall information about a network. This is the bottom up approach of system biology, which has start with the separates nodes of the network in the bottom. Similarly motifs and modules were studied before visualizing and understanding the whole network. The transcription factor and target genes are shown by nodes and the interactions between them is edge. The different dataset were retrieved from various databases presenting transcription regulation between transcription factor and target genes.

- HTRIdb dataset
- TFactS Sign less/ Sign sensitive dataset
- YeaSTRACT dataset
- ESCAPE dataset

The HTRIdb dataset is primarily an open access database for experimentally verified human transcriptional regulation interactions. The interaction pattern based modelling of transcription factor and target gene is useful in achieving the complete understanding of the

biological process. It is basically a repository of human transcription regulation interactions. The advantage of this dataset to other is that it gives the information on different experimental techniques which has been used to extract the information(Bader and Hogue 2003). The different techniques used have the certain biochemical feature of it that determine the interactions. The database allows the user to upload their set of interactions and increase the database quality. The inconsistencies, that can be identified in the database is also another aspect where one can work on(Alon 2007).

TFactS dataset has the catalogue of 2720 target genes and 6401 experimentally validated regulations. In order to decipher the regulated transcription factor´s network, which were obtained from microarray data are compared with the well characterized target genes in TFactS. TFactS has the validated published list of regulated genes which were compared to tools based on *In silico* promoter analysis(Zhang, Jin et al. 2007). *In silico* gene mining strategy is an excellent tool for identification of key genes and gene clusters whose expression is changed in disease tissue. The data generated by this investigation offers delineation for molecular basis of diseases(Vidal, Cusick et al. 2011). The prediction of transcription factors which are regulated, inhibited or activated in biological condition for TFactS is entirely based on the lists of regulated, up regulated and down regulated genes resulted from the transcriptomics experiment (http://www.TFactS.org/). The nature of transcription factors in the catalogue is determined to be correlated with the target genes in the set. The different species under which the experiments are performed is also separately dealt. The different data source like PAZAR, NFIregulome, curated TRED and TRRD along with PubMed are used to obtain the Tg for all those Tf. The catalogue of Tf-Tf interaction in TFactS is the repository of those transcription factor and its respective target genes, with additional information on the regulation level, whether they are up regulated or down regulated.

The genome of Saccharomyces cerevisae was sequenced by 1996, after that period this model eukaryote has been used to understand complex biological networks that controls the cellular processes. The Yeast Search for Transcriptional Regulators And Consensus Tracking (YeaSTRACT) is the consequence of release of genome sequence of Saccharomyces cerevisae, now it has been transformed into a curated repository for the 20, 6000 regulatory associations between transcription factor and target genes in Saccharomyces cerevisae. The database information is based on the huge number of bibliographic references. The site has the updated gene information from the Saccharomyces Genome database, and use of updated

gene ontology term is another feather attached to the database. Along with PBM- MITOMI changes, MITOMI is the versatile platform useful for different bio molecular interaction measurements including protein –protein, protein-DNA, protein-RNA and protein small molecules(Geertz, Shore et al. 2012), the database is prospering under newly appended curation regulatory information on environmental condition, association type and evidence code(http://www.YeaSTRACT.com/).

The ability of the cell to give rise to all the cell of the embryo and the adult organism is referred to as Cell`s Pluripotency. Pluripotent cells are usually found within mammalian blastocysts and for brief moments in embryo after implantation. The embryonic cells are derivation from inner mass of the blastocysts, which is a renewable source of pluripotent stem cells that are striking high in basic science studies. Later may become an efficient source of cells for safe, effective, cell based therapies. Thus all these embryonic stem cell research are leading to generate a canvas for molecular signature of Pluripotency with the emergence of complex interaction of transcription factor networks, signaling pathways, and epigenetic processes involving modification in structure of DNA, histones and chromatin(Gearhart, Pashos et al. 2007). High content studies on the mouse and human embryonic stem cells (m/hESCs) using various genome wide technologies like transcriptomics and proteomics is continuously done and published as well for a novel purpose. But the integration of these data with the motive to obtain global map of molecular network in m/hESCs was missing. Thus, m/hESCs centered database called ESCAPE compiles data from many recent high throughput studies including chromatin immunoprecipitation, followed by deep sequencing, genome wide inhibitory RNA screens, gene expression microarray or RNA sequence after KD (Knock down) or over expression of critical factors, immunoprecipitation followed by mass spectrometry proteomics and phosphor proteomics. Retrieved Protein-DNA data from ESCAPE comprised 206521 documented interaction form ChIP-chip/seq studies connecting 61 transcription factors to their respective target genes. Besides Protein-DNA dataset in ESCAPE it also has datasets on protein-protein, LOF (Loss of Function)-GOF(Gain of function) interactions and many more(Shih and Parthasarathy 2012).

4.1 Node degree analysis

Node degree analysis of the network is the important measure to estimate the spatial distribution of different nodes within the network and their transmission range(Bettstetter 2002). The different nodes like protein, gene and small molecules interact to represent the

cellular system, and by analyzing individual components` interaction to each other of a complex organism it appears to be a daunting approximation to gain full understanding of the system. However in recent years of biological science, there have been using several advancements in technology to study the interaction of molecules. The range of techniques like microarray, co immunoprecipitation, two hybrid assay and different ChIP approaches has been used in Protein-Protein interaction (PPi) and gene regulatory networks. The number of research work that has been done has the motive not only to unravel the different biochemical phenomena; instead the larger objective is to elucidate essential principle and cardinal mechanisms of cellular system(Wang, Joshi et al. 2006).

However the degree (K) of a node is number of biochemical reactions it gets involved in. For example in encoding genes, transcription factor (AR) that is turned on, induce the expression of gene (FOXO3). The transcription factor and target genes are the nodes and the technique called ChIP-chip which is used to elucidate the interaction is usually technique in HTRIdb dataset. Since the edge joins the participating nodes, edge may vary upon the choice of picking any interaction between the Tf-Tg. Most of the time it is weather the technique employed to detect binding, regulations, species or cell type etc.

4.1.1 HTRIdb dataset

The HTRIdb dataset has the scale free with the hierarchical modurity in the degree distribution. The numbers of nodes forming the regulatory hubs within the network gives the shape of hierarchical network to the datasets. More than 15 nodes have the node degree (K) < 500. Even though they have unit frequency, there are thousands ((K) =5335) of other nodes with the node degree close to 1. These nodes join the different regulatory hubs together. The scarcely available nodes with the low node degree joined to the distinctly separable nodal hubs enable the network to acquire the small world property. Basically the transcription factors in the dataset have the node degree in the higher side. The transcription factors like ETS1, GATA, AR, YBX1, FOXP3 and many more has more out degree as well with minimal indegree.

The nature of nodes shows the phenomena of preferential attachment, basically the nodes which are already involved gets more nodes towards it and the nodes which has very low degree remains same. The topology of these network appear to be dynamical in nature, instead of being static, the evolution of the network can be characterized by hypothesis of the growth and preferential attachment. The growth hypothesis delivers the suggestion about the network that the networks continuously expand through the addition of new nodes and links

between the edges and the preferential attachment hypothesis focusses on the rate $\prod(K)$ with which the node k gets attached to new links is actually a subsequently increasing function of K(Jeong, Néda et al. 2003).

Similarly the Outdegree of transcription factors in HTRIdb dataset has the similar pattern of distribution like overall node degree distribution. The reason behind this might being the transcription factors making regulatory hubs in the network. The several modules that are formed by the transcription factors gives the out degree distribution the hierarchical configuration.

However the appearance or involvement of transcription factors in huge number can be correlated with the obvious suspicion every biological reaction in vivo carries. The noise in the data is suspected in the HTRIdb dataset.

The distribution of different techniques employed to detect the binding between the transcription factor and target genes in the dataset are shown in Bar chart below

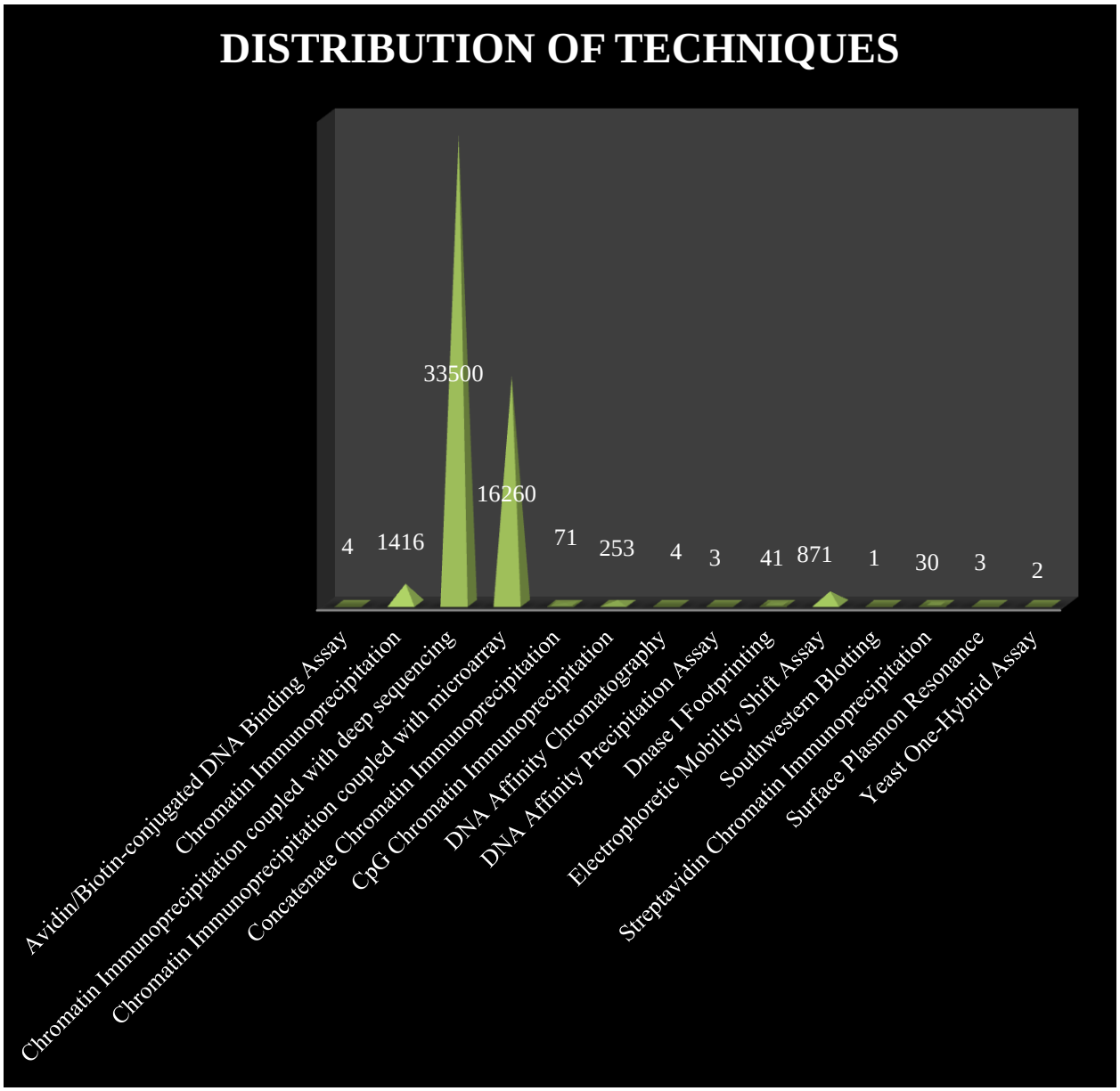


Fig 5.1: The distribution of different techniques employed to detect the binding of Tf-Tg in the HTRIdb dataset.

During the ChIP-seq experiments, the majority of unbound DNA fragments are meant for wash in immune precipitation procedure. Thus ChIP processed library has the fragments pulled down from the genomic loci with high chance of protein-DNA interaction or histone modifications. But sometime the non-useful fragments remain in the library because of random protein-DNA or antibody-DNA contacts that are not position specific. The sequence read from these fragments are widely spread in genome and later gets considered as noise in addition to the real enrichments by ChIP experiments. The noise rate stated in the model estimates the accuracy of data quality(Xu, Handoko et al. 2010).

The important aspect of gene regulation is the interaction of genomic *cis* acting elements with transcription factors. The binding sites for transcription factors are found by promoter studies on known genes. These studies typically are sequence analysis for consensus binding sites, electro mobility shift assays, promoter-reporter analyses and chromatin immunoprecipitation (ChIP) experiments. These process are worked basically on already studied genes and specifically towards the binding sites located in their proximal 5` flanking sequences. The less biased approach for discovery of transcription targets is through administrating the activity of transcription factors followed by gene expression analysis, such as RNA differential display or expression microarrays(Barski and Frenkel 2004). But the real problem is on the interpretation of such experiments. It is difficult to conclude weather a gene under study is direct or indirect target of transcription factor of interest. Secondly, in experiments in which there is overexpression of transcription factors, the response of some genes may be forced by exaggerated concentration of the transcription factors, hence resulting in physiologically insignificant results. Next flaw behind such study is that they do not provide information on the location of cis-acting regulatory elements. And these expression studies are simply futile in disclosing genes, which get bound to transcription factor of interest without influencing the respective mRNA levels under different experimental conditions, due to a) compensatory mechanism or b) absence of co-activators that may be present under different conditions(Barski and Frenkel 2004).

Most of the interaction between transcription factor and target genes in HTRIdb dataset is detected by the Chromatin immunoprecipitation coupled with deep sequencing (ChIP-seq) and Chromatin immunoprecipitation coupled with microarray (ChIP-chip) technique. ChIP-seq is used for 64 % of interaction and ChIP-chip for 31 %, which is the predominating margin where there are 12 other techniques used. Similarity another façade of data reveal

comparatively very few transcription factors are involved in these interactions. The transcription factors- target genes interaction which have been isolated by these ChIP techniques can be visualized as the one to many interaction, as normally 1 transcription factor gets involved with many target genes.

Out of total 284 transcription factors involved in regulation with 18,302 target genes enabling 51,871 interactions between, only nine (9) transcription factors are involved in 33,500 interactions is detected from the chromatin immunoprecipitation coupled with deep sequencing (ChIP-seq similarly, eight (8) transcription factors are involved in 16,260 interactions which are revealed by the Chromatin immunoprecipitation coupled with microarray (ChIP-chip).). Those transcription factors are scripted in the discussion to know about those highly active transcription factor namely ESR1, ETS1, GATA1, GATA2, GATA3, FOXA1, YBX1, TFAP2C, PRDM14 in case of ChIP-seq detection and AR, E2F4, ESR1, HIF1A, MYC, TP53, TP73, FOXP3 were involved in interactions, later were detected by ChIP-chip.

Huge number of target genes of HTRIdb dataset has the distribution following the scale free topology. That does have inferences after observing the indegree distribution graph, like there is uniform distribution of highly connected node, i.e. descending in indegree count for nodes is uniform. For example CDKN2A has the highest indegree of 65, this is target gene with highest indegree, there are 65 transcription factors getting involved with this target gene to regulate and result the desired regulation. However the involvement of target gene falls after that to 19, 17, and 16, and so on, for the target genes VEGFA, CDKN1A and BCL2 respectively. After the VEGFA, the transcription factor load on the genes gradually decreases. However there are 124 target genes involved with only one transcription factor for transcription. Since it is expected for the transcription factor, it has many target gene to act on in case of HTRIdb dataset, thus the In-degree of target gene can be adjusted to be at least 1. This correlation for minimum transcription factor acting on target gene can be extracted by the fact that dataset has only 283 Transcription factor acting with 18028 Target genes. The relationship between transcription factors to the target gene is of `one to many`. So each one from the set of `many` target genes gets at least one transcription factor with the help of estimation conferred with HTRIdb data.

4.1.2 Sign-less TFactS dataset

The total of 342 transcription factors and 2450 target genes in the dataset, have the 6823 experimentally verified interaction in the different organisms Rattus norvegicus,

Homosapiens, Mus musculus. The data retrieved from the catalogue has three set of data, one has the features about the transcription factor and target gene interaction in in different species and in combination of them as well. The microarray data of TFactS without any information on the transcriptional activation or inhibition was analyzed on the basis on the node distribution, clustering coefficient and sub networks for this dataset.

The node degree distribution of Sign less data has the strong preference towards forming regulatory hubs within the network. Since the data set has the observation of the transcription factors like Myc which has the nodal degree of 579, however it has the character of acting as the target gene as well, since it has indegree of 25 and Outdegree is 552. In human genome , Myc is expected to regulate the 15% of all the genes(Gearhart, Pashos et al. 2007), by binding to the enhance box sequences (E-boxes) and recruiting histone acetyltransferases (HAT). This enables Myc to function for regulating global chromatin structure by performing regulating histone acetylation in human genome both in gene rich regions and also in sites with no known genes(Cotterman, Jin et al. 2008). The likes of nodes like SP1, CTNNB1, E2F1, TP53 has disproportionate share of higher Outdegree with minimal indegree like Myc. That means those transcription factors also act as the target genes during the regulation.

The dataset has the 10 more nodes able to form regulatory hubs in the network similarly there are 107 nodes with the node degree of 1. The dataset had distribution of regulation between different species. The out degree of the transcription factors in sign less TFactS has the Myc transcriptional factor which is highly involving to the many target genes.

4.1.3 Sign-sensitive TFactS dataset

The purpose of this dataset seems to provide the general information about the transcription factor and their respective target gene interaction, however when the increasing necessity of information on level of regulation between Tf-Tg, the catalogue of TFactS offered sign-sensitive dataset as well. The up and down regulation of the interaction is mentioned as an additional information to the sign less dataset. In sign sensitive dataset, there are 114 target genes and 1635 transcription factor. The node degree distribution (Fig: 15) of TFactS dataset has the scale free nature of distribution with hierarchical modularity in the network. However it can be observed that the transcription factors have the dominant presence on the overall nodal distribution of Sign-sensitive TFactS dataset. The graph obtained by the Outdegree distribution (Fig: 16) highlights the

MYC, SP1 and other transcription factors has the lower count of out degree. The highest for CDKN1A is Outdegree $(Kout)$ of 20. They both have the modular scale free hierarchical architecture. However the Scale free coefficient observed in the indegree distribution of target gene is observed. Therefore the network of the whole TFactS dataset is scale free in topology, with nodes always aiming for growth through possible preferential attachment. The RBPJ and FOXO3 nodes have the more indegree than for average target genes.

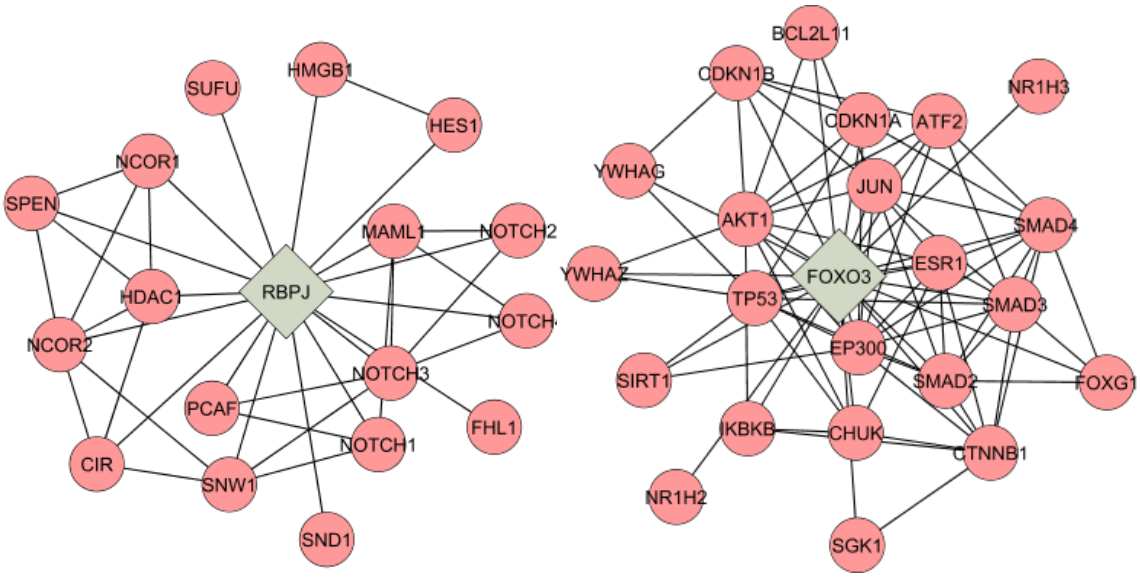


Fig 32: The Cytoscape visualization of the two genes, queried in human through MiMI plugins. The Square shaped genes (RBPJ and FOXO3) are interacting with the proteins in pink circle.

The RBPJ gene is found to be involved in many biological processes with hundreds of transcription factors acting on it. It is involved in notch signaling pathway, negative regulation of gene expression and Regulation of DNA dependent transcription with the p-value of $5.2699e^{-16}, 1.4243e^{-13}\ and\ .7147e^{-19}$. Similarly the FOXO3 gene has major role in the regulation of developmental process and cell differentiation with the p-value showing as $1.22e^{-15}\ and\ 8.393e^{-15}$, for the respective processes.

Comparison of HTRIdb and TFactS datasets.

The genes from the HTRIdb dataset were compared to the TFactS, by the data analysis feature on webpage of TFactS. The purpose behind the comparison is to know about the gene repertoire of TFactS dataset. Since it is already known about TFactS, it has the catalogue for Tf-Tg interaction. The observed genes from the HTRIdb dataset will be

analyzed in the Data analysis in the TFactS. The data observed has the 0.5 P-value, E. value and Q value threshold.

SN	Technique of HTRIdb	No of genes from HTRIdb	No of genes that are present in TFactS.
1.	Streptavidin Chromatin Immunoprecipitation	30	5
2.	Chromatin Immunoprecipitation	1416	451
3.	Chromatin immunoprecipitation coupled with deep sequencing	33500	1878
4	Chromatin immunoprecipitation coupled with microarray	16260	1328
5	Electrophoretic mobility shift assay	871	282

Table 13: The table observed as the result of analysis in TFactS data analysis of HTRIdb genes.

The Inference obtained from this analysis is that the TFactS data catalogue is not updated. Thus it cannot become a very good source of data analysis source. Beside this issue of TFactS not getting updating, we can conclude that HTRIdb dataset has the good repertoire of genes retrieved through different techniques.

Comparing regulatory hubs

The node degree analysis is done in data sets' network revealed that there are discrepancies in the number of regulatory hubs forming. Even though it is expected since they are both different datasets to work on. But a different result appeared. We sorted the nodes based on the node degree (K) above 50 in both networks. In case of TFactS, Out of total 1748 nodes, 44 nodes with node degree (K) above 50 were observed. The node degree ranges from 57 to 579. The transcription factor MYC had the highest node degree. The HTRIdb database has only 24 highly connected nodes, i.e. node degree (k) <50. Even though the size between them is not comparable, HTRIdb is large enough than TFactS. It has radical range of a node having degree (K) 51 to 9759. Among them 15 nodes have node degree (K) above 1000 nodes in the total of 18308 nodes. As a whole in network they are involved with more than 91.98 percentages of the nodes. This shows network of few densely involved nodes and other one

with many but smaller regulatory hubs in large number. Here comes an approach of preferential attachment. The preferential attachment is the cause behind these biological networks to form regulatory hubs, instead of getting sparsely distributed. One node tend to go the one which already in group/connection/ as hub.

Similarly the issue of attack to these networks is very benile but when the targeted attack is attempted the lethal effect is observed. Hence we can conclude that the regulatory hubs formed in the biological networks have some centrally important nodes/biomolecules. If the central component of node is removed the effect is really massive, in case of biological process.

In network science the centrality of the node can be measured by different parameters like node degree, clustering coefficient and shortest path length and connectivity value. Thus the assistance of network science is taken to elucidate the various biological processes.

4.1.4 YeaSTRACT dataset

The documented list of transcription factor and target genes was retrieved through the YeaSTRACT dataset. The matrix form of the data was first changed in two columns of transcription factor and target gene. The initial impression was 255 transcription factors were observed 287 target genes. The similar density of transcription factor and target gene was obtained for the first time in all these datasets. However the total interaction observed was 7876.

The node degree distribution of the YeaSTRACT had a very unusual pattern for any biological network to behave. The numbers of sub graphs appears with the network but minimal clustering coefficient value per node. It is observed that the transcription factor and target genes only do interact one to one or one to many. I believe the absence of reversible biochemical reaction in them. After observing the recurring motifs that are obtained from MCODE analysis it is visible that most of them are single input motifs and the most of the regulation can be operating via simple regulation between regulators and the sequence.

The node degree distribution corresponds to the earlier model presented by the two Holland mathematic, Erdos and Renyi, the random network model can be predicted with the observation of the Node degree (K) distribution, Outdegree ($Kout$) distribution of transcription factor and indegree (Kin) of target gene.

4.1.5 ESCAPE dataset

The humongous size of the data was dealt with lot of effort, regarding efficient computer with good memory to simulate the data in Cytoscape and generate the network parameter result. The huge size corresponds to the node distribution for some nodes. In this case MYC along with other transcription factors like SMC1A, NANOG, SUZ12, MED1, MED12 are forming huge regulatory hubs within the network. The character of the network is determined by these transcription factors. However there are only 61 transcription factors compared to 23047 target genes.

The node degree distribution shares a beautiful graph of a hierarchical model of network where there is few highly involving central point of the network. However the Outdegree $(Kout)$ distribution has the totally different nodes with different Outdegree $(Kout)$ values. They suggest the formation of random network within. Since the average of the Outdegree $(Kout)$ value is equal to the most of the node` Outdegree. The scale free distribution of the target genes in the ESCAPE dataset has the distribution of all the 23047 genes throughout the network. Since compared to the transcription factor number the target genes are in more number the sharing transcription factor activity on all the genes has caused in the uniform distribution of Outdegree value. The highest indegree was for DIDO1 with $(Kout) = 49$.

Normalization study

Basically the purpose behind making Normalization study is to observe a particular gene/transcription factor, its redundancy in any set and compare them in between.. Normalization actually delimits the observation frequency according to the relative calculation of its frequency to make in same level. The observation will basically focus on the frequency of MYC in the three different datasets; HTRIdb, TFactS and ESCAPE.

The calculation is made to know the frequency of selected node as s transcription factor, s a target gene and as a whole in undirected network. The calculation is done is percentage:

$$\frac{Observations\ in\ dataset}{whole\ dataset} * 100 \quad \text{.....(ii)}$$

Data set	In undirected network	Transcription factor	Target gene
MYC-TFactS	42.43%	40.45%	1.8%
MYC-ESCAPE	3.3%	3.2%	0.065%
MYC-HTRIdb	2.97%	2.8%	0.092%
AR-TFactS	0.505%	0.439%	0.065%
AR-HTRIdb	36.10%	0.021%	36.06%
SP1-TFactS	3.18%	O.036%	3.12%
SP1-HTRIdb	1.14%	0.004%	1.09%
JUN-TFactS	1.11%	0.146%	0.95%
JUN-HTRIdb	0.207	0.043%	0.16%

Table 14: The observation of MYC, JUN and AR presence in the different datasets.

Thus it is observed that MYC is found in the highest frequency in TFactS dataset compared other in other two. However, the remaining 3 set of observation was not found in ESCAPE dataset. The AR transcription factor is more than in TFactS dataset. The SP1 is in higher proportion at TFactS than HTRIdb. And lastly the JUN was relatively found in relative number in both dataset, yet TFactS has a bit higher per percentage, precisely observing.

4.2 Clustering coefficient distribution analysis

The clustering coefficient distribution of the nodes in the network was carried out in the undirected network of different datasets. The clustering coefficient distribution is expected to provide the complexity parameter of the network. In an ideal condition, the biological network forming a cluster most probably regulate by feed forward loop between two regulators and a gene sequence. The interaction between the transcription factor (regulators) and gene sequence (DNA) forms a pattern of transferring the information in the central dogma of biology. The clustering coefficient maps out the path between them and correlate with the essence of the direction of that way. If the paths between these macromolecules of the biological process are intertwined to form triangles in between frequently - Then it is believed that the transfer of information/energy is proportionate for the end product in biology. The most clustered network or the sub graphs within in is expected to have more valuable clique than other in the same network.

In case of HTRIdb dataset, the clustering coefficient to is respective degree was plotted to observe the graph. The graph resulted in the hierarchical conformation. The negative slope of the graph, the clustering coefficient decreasing with the increase of degree count, leads to the solid assumption of referring it as the scale free hierarchical network.

With the 30 target genes forming 30 triangle's in the network, the architecture of the network coincides with the hierarchical topology. The nodes clustering coefficient of the node having degree ranging 1 to 579 formed a hierarchical network, representing the different modules within. Modules are the sub graphs with the specific biological functioning.

On the other hand in this TFactS Sign-sensitive dataset, the highly clustered portion of the network is formed by transcription factors only. The Scale free topology of the graph is complemented with the hierarchical inputs.

The graph of YeaSTRACT has the clustering coefficient not dependent on the degree of the transcription factor or target gene of the network. It simply forms no triangle, i.e. equal to two nodes having a common acquaintance. The graph formed is of random graph topology.

In this gigantic network, the 608 nodes appear with the clustering coefficient value of 1. But one thing is to be making sure that size of any network do not assures its increment in the overall clustering coefficient. The clustering coefficient of overall network is 0.540 (from table). The data forms the hierarchical network.

4.3 Sub network analysis

The Local level of analysis studies is the units of a network, pattern of regulation which is called the Network motifs. The regulatory network is organized as the repeated appearance of highly significant motifs. The easily characterizable feature of network motifs' view in the entire known transcriptional network of the organism is the defining feature in analyzing computational elements of biological network(Shen-Orr, Milo et al. 2002).

4.3.1 HTRIdb

HTRIdb dataset has the 30 sub networks within contain 13 motifs with simple regulation (SR) pattern, 6 dense overlapping regulon (DOR), 6 Feed forward loops (FFL), 5 Single input modules (SIM). The diversity of different pattern can be observed with 100 target genes and 63 transcription factors. The transcription factor and target gene interact with each other for the regulation of transcription in human transcription regulation system by forming motifs throughout.

The overrepresentation analysis is performed to find out the genes of motifs, that are involved in particular process. Out of 30 motifs 10 motifs are briefly described here to observe the different processes that happen in the motifs. These motifs are recurrent throughout the body.

- 1st cluster:

The genes like cdc6, rad1, e2f4, fanCD, e2f6, pim1, wee1, foxn3, rad51 are basically involved with cell cycle phase and interphase mitotic cell cycle with the p-value of $3.0848e^{-7}$,$2.8868e^{-6}$ respectively. The MCODE score of the cluster is 1.875.

- 2nd cluster

The likes of gene FOS, RARA, STAT1, SLIT2 are the one which are involved in the cellular response to hormone stimulus and endogenous stimulus with the p-value$1.0190e^{-6}$ and $1.2386e^{-6}$ respectively. Gene GATA3 is involved in the same process but with the p-value of$1.3963e^{-6}$. The MCODE score is 1.75.

- 3rd cluster:

The genes like GATA2 and SP1 are involved in the positive regulation of gene specific transcription with the p-value showing$2.7141e^{-4}$. The MCODE score of the cluster is 1.667

- 4th cluster

The ETS1, VEGFA, TRF4, TFA4, TFA2a, IRF2, PAX5, RORA,USF1and MXD4 are involved in the regulation of transcription from RNA polymerase II promoter. The p-value is $1.9793e^{-3}$. The MCODE value of the cluster is 1.647.

- 5th cluster

The PLA2 and G4A genes are involved in many biological processes like response to methyl mercury, positive regulation of vesicle fusion, platelets activating factor biosynthetic process and regulation of post glandin biosynthetic process. The p-value for these processes is $5.9515e^{-4}$. The MCODE score is 1.5.

- 6th cluster:

The two genes APOP4 and PPARA are involved in numerous processes; however the positive regulation of lipid catabolic process is more dominant. The MCODE score is 1.5

- 7th cluster

The genes like RARG and ESR1 are involved with prostate gland epithelium morphogenesis with the p-value of$5.6354e^{-4}$. Along with another gene called HOXA10 has involvement

with the skeletal system development. The p value of the process is $4.0449e^{-4}$. The MCODE score is 1.5.

- 8th cluster

The eighth cluster has the different genes like ATF4, ASNS, RBPJ, MYC and PTCN involved with the positive regulation of cellular process. The p-vale of the process is $2.847e^{-4}$. The negative regulation of cell migration is controlled by RBPJ, PTEN with the p-value of $3.021e^{-4}$. The MCODE score is 1.5.

- 9th cluster

The cluster have genes FOXO1, IFBP, PARP1, STAT3 have the cellular response to the peptide hormone stimulus. And genes like SATB1, YY1, CREM, EZH2, BFCA2, FOXO1, PRM2, FOX3, RUNX3 and STAT3 has the regulation of transcription, DNA dependent. The MCODE score is 1.467.

- 10th cluster

The genes like CDKN1A, BAX and TP53 were found to be involved in induction of apoptosis by intracellular signals. The p-value is $4.5308e^{-8}$. Another set of genes within same motif are found to activate caspase activity by cytochrome C, the genes are BAX and TP53. The p-value is $8.2081e^{-7}$. The MCODE score is 1.333.

4.3.2 Sign sensitive TFactS:

The MCODE analysis of the dataset provided 8 sub networks for the dataset. The different combination of the motifs is observed throughout the network. The sub network had different pattern of combination of sub networks. The 4 Dense overlapping Regulon (DOR), 2 single input modules (SIM) patterned modules, single Feed forward loop and Simple regulations motifs were found.

The gene which was involved in different sub network formation in dataset was very few. Only 5 target genes STAT5A, STAT5B, SMAD3, CREB1 and RBPJ were involved in several regulations in sub network. The transcription factor acted both as transcription regulators and site to act on, for other regulators in the regulation process. All of the interaction within Sub graphs resulted in the activation only 6 are repressed.

The overrepresentation analysis of target genes in the BiNGO led some biological significance of these genes in Human. The overrepresentation analysis was done in default parameters of BiNGO: It was observed that the genes were involved in positive regulation of

transcription and positive regulation of gene expression with the p-value of $2.587e^{-6}$ and$3.126e^{-6}$. The genes CREB1 and SMAD3 were involved in process of positive regulation of transforming growth factors $\beta 3$ production with the p-value of $5.863e^{-8}$ and many more.

The overrepresentation of up regulated genes of humans results; those genes are mostly involved in DNA dependent regulation of transcription, regulation of RNA metabolic processes and regulation of transcription from RNA II polymerase promoter with the p-value of $2.0585e^{-57}$,$1.3314e^{-56}$, $3.0561e^{-52}$ respectively and many more.

The down regulated genes are also involved in Regulation of transcription, DNA dependent, regulation of metabolic process and regulation of transcription with the p-value of $1.73732e^{-22}$, $3.7292e^{-22}$ and $1.7558e^{-21}$ respectively and many more.

The over representation analysis of unregulated genes of Mus musculus has the regulation on the DNA dependent transcription and regulation of RNA metabolic processes with the p-value of $6.9569e^{-72}$ and$1.7709e^{-71}$. The genes involved are E2F1, PPARA; E2F2, MYOD1, MEF2A, CDX2, E2F4, HNF1A, STAT5A and PPARG.

The over representation analysis of down regulated genes of Mus musculus resulted in: genes like E2F1, PPARA, E2F4, THRB, CREB, YY1, SOX2, TP53, SOX6 and HMGA in the regulation of DNA dependent transcription and Regulation of RNA metabolic processes. The p-value of the occurrence was observed as $2.7413e^{-31}$ and$3.9659e^{-31}$.

The over representation analysis of Rattus norvegicus up regulated genes like E2F1, SREBF, HNFIB, HNFIA, RXRA, ESR1, SMAB, NFYA and many more are involved in processes of positive regulation of DNA dependent transcription and positive regulation of RNA metabolic process. The p-vale of these two biological process occurrence by these were $3.9269e^{-29}$ and$4.9879e^{-29}$.

The overrepresentation analysis of down regulated genes SREBF1, SREBF2 are found involved in positive regulation of transcription via sterol regulatory element binding with p-value $1.7622e^{-7}$ and same gene set including SMAD3 get themselves down regulated at the positive regulation of gene specific transcription from RNA polymerase II promoter. The p-value is later down regulation is $5.9904e^{-7}$.

4.3.3 YeaSTRACT

The saccharomyces cerevisae dataset observed from YeaSTRACT database was retrieved in order to know the different pattern of Transcription factor and target genes. Total of 13

different sub graphs were observed. The different pattern of sub graphs in the dataset were like 5 Single input module (SIM), 6 Dense overlap regulon (DOR) and 2 Feed forward loop (FFL). The separation of the sub graphs were labelled according to its MCODE score.

The overrepresentation analysis of the different sub graphs gives the biological significance these sub graphs, within the network. The recurring pattern of these sub graphs/motifs are the reason behind giving the importance to a particular clique/sub graph/motif in the network.

- 1st cluster

The first cluster has 6 transcription factor and 14 target genes. The small molecule catabolic process is regulated by the genes of this cluster. The genes are CHA1, ADH5, GLK1 and the p-value is $2.4705e^{-4}$. The process of cellular aminoacid catabolism and amine catabolic process has these genes namely CHA1, ADH5. The p-value of the process was$8.2573e^{-5}$. The MCODE score is 3.35.

- 2nd cluster:

No overrepresentation was observed in this cluster.

- 3rd cluster

No overrepresentation was observed in this cluster.

- 4th cluster

In this cluster, the only one gene was over expressed. The TSC10 gene was found to be involved with 3-keto sphinganine metabolic process. The p-value observed is$9.6634e^{-4}$. The MCODE score is 1.8.

- 5th cluster

No overrepresentation was observed in this cluster.

- 6th cluster

No overrepresentation was observed in this cluster.

- 7th cluster

No overrepresentation was observed in this cluster.

- 8th cluster

No overrepresentation was observed in this cluster.

- 9th cluster

The ninth feedback motif has genes like ARL1, getting involved in Golgi to plasma membrane protein transport. The p-value is $2.8971e^{-2}$. Similarly another gene Sec66 is found to be involved in post translational protein targeting to membrane translocation. The p-value is $4.3436e^{-3}$. The MCODE score is 1.2.

- 10^{th} cluster

No overrepresentation was observed in this cluster.

- 11^{th} cluster

In the sub graph, MET8 is overexpressed, with the role in Siroheme biosynthetic process, Siroheme metabolic process and sulphate assimilation. The p-value for the processes are$6.4428e^{-4}$,$6.4428e^{-4}$ and $3.5410e^{-3}$ respectively. The MCODE score was 1.125.

- 12^{th} cluster

No overrepresentation was observed in this cluster.

- 13^{th} cluster

No overrepresentation was observed in this cluster.

5 Conclusion

The aim of the thesis is to analyze, compare and isolate some biologically relevant understandings from the different datasets retrieved from public resource database. The dataset were analyzed on the baseline of graph theory. Most of the data analysis was performed through Cytoscape, an open source software platform. Different Cytoscape applications like Network analyzer, BiNGO, MCODE and MiMI were the yardstick for the network analysis. The topological parameters were selected and worked accordingly for the further progression in the analysis; by focusing on parameters like Node degree, clustering coefficient and Sub networks of the network formed from datasets.

5.1 Node degree distribution

The node degree enacted as tool to compare between the datasets by elucidating the distribution of transcription factor and target genes. The undirected network node degree distribution, Outdegree of transcription factor and indegree of target genes were compared in between the datasets. Three scale free networks had the prominent presence in the distribution of Outdegree and Indegree of transcription factor and target genes.

Outdegree of transcription factors has characteristics functional regulatory hubs, whereas the indegree of target genes is purely Scale free in case of HTRIdb and Sign-les TFactS. In case of Sign sensitive TFactS, the Outdegree was observed having scale free distribution and the Indegree had exponential distribution. Similarly, In case of ESCAPE dataset, Outdegree was found having random graph, small world effect in distribution and indegree exposing beautiful scale free network. Whereas, the YeaSTRACT dataset had both distributions of target gene and transcription factor as an example of random network model. A conclusion is derived from the different distribution pattern of these datasets in this report. If in a dataset, if the number of transcription factor is way higher than target genes (Tf>>Tg), it has been observed that Outdegree of transcription factor in this case gets a pure scale free distribution and Outdegree a modular hierarchical network with various regulatory hubs. In the same way, if the target gene is way higher than transcription factor (Tg>>Tf) Indegree of target gene appears in the pure scale free distribution and another can be exponential or hierarchical distribution.

5.2 Clustering coefficient Distribution

The clustering coefficient distribution in the different dataset was observed as each one had the negative slope on the graph that means the clustering coefficient goes on decreasing with increasing number of degree on node. Thus, some conclusions can be inferred from this

observation is; the clustering coefficient is not dependent on the size of the network. Instead clustering coefficient has the requirement of only three nodes joined together for the peak value or proper transmission of information between different components of biological processes. The clustering coefficient for YeaSTRACT was found due to the reason because no regulators acted as the target gene during the various processes. There is no transcription factor with the character of target genes.

5.3 Sub-network analysis

Different functional motifs within the network were analyzed. The analysis was based on the ranking provided by the MCODE score, higher the score higher is prevalence in the network. Every dataset had a set of sub networks with specific pattern of formation. The sub network provided the biological processes it is more involved in together with all the joint molecules of motif. It is assumed everyone works together, but in real picture at least one of the members of the motif is involved in a biological process. The p-vale determines the probability of obtaining a desired result, the p-value should be, $p - value \leq o.o5$, for the selection of biological processes in which the gene is involved most. Since the BiNGO, provides the overrepresentation analysis in wide spectrum of p-value-Thus selecting the probable best provides the understanding on the biological processes. By analyzing different sub networks of a whole network, one can gain insights on the different prominent biological processes. .

6 Future perspectives

The transcriptional regulatory databases were primarily maintained by the different techniques like ChIP, ChIP-chip, ChIP-seq; but the emergence of noisy data in the database may not resource the community with the prolific understandings on the biological processes. Rather, other reliable techniques can be emphasized for the better and updated formulation of transcription regulation related datasets. The analysis of network by thus, formed sound dataset will convincingly propagate the essence of studying the biological networks more.

7 Bibliography:

. from http://www.TFactS.org/ <http://www.tfacts.org/>.

Almaas, E., A. Vázquez, et al. (2007). "Scale-free networks in biology." Biological networks **3**(1).

Alon, U. (2007). "Network motifs: theory and experimental approaches." Nature Reviews Genetics **8**(6): 450-461.

Babu, M. M., N. M. Luscombe, et al. (2004). "Structure and evolution of transcriptional regulatory networks." Current Opinion in Structural Biology **14**(3): 283-291.

Bader, G. D. and C. W. Hogue (2003). "An automated method for finding molecular complexes in large protein interaction networks." BMC bioinformatics **4**(1): 2.

Barabasi, A.-L. and Z. N. Oltvai (2004). "Network biology: understanding the cell's functional organization." Nat Rev Genet **5**(2): 101-113.

Barski, A. and B. Frenkel (2004). "ChIP Display: novel method for identification of genomic targets of transcription factors." Nucleic acids research **32**(12): e104-e104.

Bettstetter, C. (2002). On the minimum node degree and connectivity of a wireless multihop network. Proceedings of the 3rd ACM international symposium on Mobile ad hoc networking & computing, ACM.

Bollobás, B. and E. J. Cockayne (1979). "Graph-theoretic parameters concerning domination, independence, and irredundance." Journal of Graph Theory **3**(3): 241-249.

Bovolenta, L., M. Acencio, et al. (2012). "HTRIdb: an open-access database for experimentally verified human transcriptional regulation interactions." BMC Genomics **13**(1): 405.

> BACKGROUND:The modeling of interactions among transcription factors (TFs) and their respective target genes (TGs) into transcriptional regulatory networks is important for the complete understanding of regulation of biological processes. In the case of experimentally verified human TF-TG interactions, there is no database at present that explicitly provides such information even though many databases containing human TF-TG interaction data have been available. In an effort to provide researchers with a repository of experimentally verified human TF-TG interactions from which such interactions can be directly extracted, we present here the Human Transcriptional Regulation Interactions database (HTRIdb).DESCRIPTION:The HTRIdb is an open-access database that can be searched via a user-friendly web interface and the retrieved TF-TG interactions data and the associated protein-protein interactions can be downloaded or interactively visualized as a network through the web version of the popular Cytoscape visualization tool, the Cytoscape Web. Moreover, users can improve the database quality by uploading their own interactions and indicating inconsistencies in the data. So far, HTRIdb has been populated with 284 TFs that regulate 18302 genes, totaling 51871 TF-TG interactions. HTRIdb is freely available at http://www.lbbc.ibb.unesp.br/htri.CONCLUSIONS:HTRIdb is a powerful user-friendly tool from which human experimentally validated TF-TG interactions can be easily extracted and used to construct transcriptional regulation interaction networks enabling researchers to decipher the regulation of biological processes.

Brenowitz, M., D. F. Senear, et al. (1989). "DNase I Footprint Analysis of Protein-DNA Binding." Current protocols in molecular biology: 12.14. 11-12.14. 16.

Buck, M. J. and J. D. Lieb (2004). "ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments." Genomics **83**(3): 349-360.

Chen, K. and N. Rajewsky (2007). "The evolution of gene regulation by transcription factors and microRNAs." Nat Rev Genet **8**(2): 93-103.

Cheng, C., K.-K. Yan, et al. (2011). "Construction and Analysis of an Integrated Regulatory Network Derived from High-Throughput Sequencing Data." PLoS Comput Biol **7**(11): e1002190.

<title>Author Summary</title> The precise control of gene expression lies at the heart of many biological processes. In eukaryotes, the regulation is performed at multiple levels, mediated by different regulators such as transcription factors and miRNAs, each distinguished by different spatial and temporal characteristics. These regulators are further integrated to form a complex regulatory network responsible for the orchestration. The construction and analysis of such networks is essential for understanding the general design principles. Recent advances in high-throughput techniques like ChIP-Seq and RNA-Seq provide an opportunity by offering a huge amount of binding and expression data. We present a general framework to combine these types of data into an integrated network and perform various topological analyses, including its hierarchical organization and motif enrichment. We find that the integrated network possesses an intrinsic hierarchical organization and is enriched in several network motifs that include both transcription factors and miRNAs. We further demonstrate that the framework can be easily applied to other species like human and mouse. As more and more genome-wide ChIP-Seq and RNA-Seq data are going to be generated in the near future, our methods of data integration have various potential applications.

Chockalingam, P. S., L. A. Jurado, et al. (2000). DNA affinity chromatography. Affinity Chromatography, Springer**:** 141-153.

Cline, M. S., M. Smoot, et al. (2007). "Integration of biological networks and gene expression data using Cytoscape." Nature protocols **2**(10): 2366-2382.

Collas, P. (2010). "The current state of chromatin immunoprecipitation." Molecular biotechnology **45**(1): 87-100.

Cotterman, R., V. X. Jin, et al. (2008). "N-Myc regulates a widespread euchromatic program in the human genome partially independent of its role as a classical transcription factor." Cancer research **68**(23): 9654-9662.

Gearhart, J., E. E. Pashos, et al. (2007). "Pluripotency redux—advances in stem-cell research." New England Journal of Medicine **357**(15): 1469-1472.

Geertz, M., D. Shore, et al. (2012). "Massively parallel measurements of molecular interaction kinetics on a microfluidic platform." Proceedings of the National Academy of Sciences **109**(41): 16540-16545.

Homola, J. (2003). "Present and future of surface plasmon resonance biosensors." Analytical and bioanalytical chemistry **377**(3): 528-539.

Ioshikhes, I. P. and M. Q. Zhang (2000). "Large-scale human promoter mapping using CpG islands." Nature genetics **26**(1): 61-63.

Janky, R. s., J. v. Helden, et al. (2009). "Investigating transcriptional regulation: From analysis of complex networks to discovery of cis-regulatory elements." Methods **48**(3): 277-286.

Jayavelu, N. D. and N. Bar (2014). "Dynamics of Regulatory Networks in Gastrin-Treated Adenocarcinoma Cells." PloS one **9**(1): e78349.

Jeong, H., Z. Néda, et al. (2003). "Measuring preferential attachment in evolving networks." EPL (Europhysics Letters) **61**(4): 567.

Lederman, I. and D. Morikis (2014). "Network Analysis of Intra-Molecular Interactions of the HIV-1 gp120 V3 Loop." Undergraduate Research Journal: 25.

Lewin, B., J. Krebs, et al. (2011). Lewin's genes X, Jones & Bartlett Learning.

Li, C. and W. Kim "Discovering larger network motifs: Network Motif clustering."

Maere, S., K. Heymans, et al. (2005). "BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks." Bioinformatics **21**(16): 3448-3449.

Mikkelsen, T. S., M. Ku, et al. (2007). "Genome-wide maps of chromatin state in pluripotent and lineage-committed cells." Nature **448**(7153): 553-560.

Ng, P., C.-L. Wei, et al. (2005). "Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation." Nature methods **2**(2): 105-111.

Orphanides, G., T. Lagrange, et al. (1996). "The general transcription factors of RNA polymerase II." Genes & development **10**(21): 2657-2683.

Ouwerkerk, P. B. F. and A. H. Meijer (2001). Yeast One-Hybrid Screening for DNA-Protein Interactions. Current Protocols in Molecular Biology, John Wiley & Sons, Inc.

Prill, R. J., P. A. Iglesias, et al. (2005). "Dynamic Properties of Network Motifs Contribute to Biological Network Organization." PLoS Biol **3**(11): e343.

The authors model how network motifs respond to small-scale perturbations and find a strong correlation between motif stability and abundance in a network, suggesting that dynamic properties of network motifs may play a role in overall network structure.

REINBERG, D. and L. Zawel (1993). "Initiation of transcription by RNA polymerase II: a multi-step process." PROG NUCLEIC ACID RES&MOLECULAR BIO **44**: 67.

Ren, B., F. Robert, et al. (2000). "Genome-Wide Location and Function of DNA Binding Proteins." Science **290**(5500): 2306-2309.

Understanding how DNA binding proteins control global gene expression and chromosomal maintenance requires knowledge of the chromosomal locations at which these proteins function in vivo. We developed a microarray method that reveals the genome-wide location of DNA-bound proteins and used this method to monitor binding of gene-specific transcription activators in yeast. A combination of location and expression profiles was used to identify genes whose expression is directly controlled by Gal4 and Ste12 as cells respond

to changes in carbon source and mating pheromone, respectively. The results identify pathways that are coordinately regulated by each of the two activators and reveal previously unknown functions for Gal4 and Ste12. Genome-wide location analysis will facilitate investigation of gene regulatory networks, gene function, and genome maintenance.

Shen-Orr, S. S., R. Milo, et al. (2002). "Network motifs in the transcriptional regulation network of Escherichia coli." Nature genetics **31**(1): 64-68.

Shih, Y.-K. and S. Parthasarathy (2012). "A single source k-shortest paths algorithm to infer regulatory pathways in a gene network." Bioinformatics **28**(12): i49-i58.

Siu, F. K. Y., L. T. O. Lee, et al. (2008). "Southwestern blotting in investigating transcriptional regulation." Nat. Protocols **3**(1): 51-58.

Spencer, V. A., J.-M. Sun, et al. (2003). "Chromatin immunoprecipitation: a tool for studying histone acetylation and transcription factor binding." Methods **31**(1): 67-75.

Strogatz, S. H. (2001). "Exploring complex networks." Nature **410**(6825): 268-276.

Valouev, A., D. S. Johnson, et al. (2008). "Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data." Nature methods **5**(9): 829-834.

Vidal, M., M. E. Cusick, et al. (2011). "Interactome networks and human disease." Cell **144**(6): 986-998.

Walhout, A. J. (2006). "Unraveling transcription regulatory networks by protein–DNA and protein–protein interaction mapping." Genome research **16**(12): 1445-1454.

Wang, Y., T. Joshi, et al. (2006). "Inferring gene regulatory networks from multiple microarray datasets." Bioinformatics **22**(19): 2413-2420.

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

Wei, C.-L., Q. Wu, et al. (2006). "A global map of p53 transcription-factor binding sites in the human genome." Cell **124**(1): 207-219.

Weinmann, A. S., P. S. Yan, et al. (2002). "Isolating human transcription factor targets by coupling chromatin immunoprecipitation and CpG island microarray analysis." Genes & development **16**(2): 235-244.

Wuchty, S., E. Ravasz, et al. (2006). The architecture of biological networks. Complex systems science in biomedicine, Springer**:** 165-181.

Xu, H., L. Handoko, et al. (2010). "A signal–noise model for significance analysis of ChIP-seq with negative control." Bioinformatics **26**(9): 1199-1204.

Zhang, S., G. Jin, et al. (2007). "Discovering functions and revealing mechanisms at molecular level from biological networks." Proteomics **7**(16): 2856-2869.