ORIGINAL RESEARCH PAPER

VOLUME-6 | ISSUE-6 | JUNE-2017 | ISSN No 2277 - 8179 | IF : 4.176 | IC Value : 78.46

INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH

Evaluation of the specific hubs in the transcriptional regulatory network (TRN) of Mycobacterium Tuberculosis

Science

Fozail Ahmad Principle Author, CIRBSc, Jamia Millia Islamia, New Delhi-110025, India

Kalpana Principle Author, Dept of CS, PGDAV (University of Delhi), Nehru Nagar, New Delhi-110021, India., CIRBSc, Jamia Millia Islamia, New Delhi-110025, India

Tanvi Dept of CS, PGDAV (DU),Nehru Nagar,New Delhi-110021,India Corresponding Author,

Ravins Dohare CIRBSc, Jamia Millia Islamia, New Delhi-110025, India

ABSTRACT

Interaction of molecules in any system is complex, such as protein-protein, gene-gene, gene-mRNA-protein interaction etc. Hence, we are seeing complex behavior of second most life claiming agent Mycobacterium tuberculosis (M. tb) due to rampageous boom across globe for its potentiality to withstand against most drug discovered. It concerns with the robustness of the entire TRN (Transcriptional Regulatory Network) of Mycobacterium tuberculosis.

Summary: This work has been divided into two approaches; firstly random and secondly directed mode of attacks against nodes in network. Findings have demonstrated that if network is attacked randomly then it would remain unaffected as its overall functionality remains intact. If network attacked in directed mode, it broke into small isolated fragments showing higher path length and hampered flow of information between nodes. Moreover, this implication put forth which can be met with novel discovery and development of drug to cure this acute disease.

KEYWORDS:

Random attack, Directed attack, Transcriptional Regulatory Network.

Introduction

A gene is a discrete unit of hereditary information, consisting of nucleotide sequence in the DNA. A number of genes which are factories of proceeding proteins of their own are arranged along the DNA length. Transcription by RNA polymerases is initiated together with a transcription factor or by activator and it can be suppressed by some repressor in various ways. Transcription factors, activators and repressor are proteins or protein complex. Thus, synthesis of proteins is activated or inhibited by other proteins, or gene regulation is activated or repressed by other protein which is regulated by other gene. Therefore, this process of gene-mRNA-protein interaction makes a network is referred to as Transcriptional Regulatory Network (TRN). TRN is controlling web of genes collaborating with every cellular element as an indispensable part of genetic machinery that the cell depends on to respond to environmental changes. Transcriptional changes are vital phenomenon, and construction and characterization of TRNs is nowadays a subject of high scientific interests. Transcriptional factors are basically responsible for mediating the regulation of gene expression that recognizes and binds to specific nucleotide sequences and affects transcription of nearby and distantly located genes.

1.1 Network Represents a Graph

Graph G=(V, E) where V is set of genes and E is the set of e_i=(x,y) such that gene x transcribes into mRNA and then is translated into a protein. Protein is a transcriptional factor which activates or represses a gene y. This graph G is representation of Transcriptional Regulatory Networks (fig. 1). Every graph has so many topological characteristics, which can be classified into two groups – local and global parameters, corresponding to the measurable elements. Local topological parameters describe individual network components while global parameters insights the whole network (Mason & Verwoerd, 2007). Various topological properties play important role in studying and characterizing the internal organization of a biological network, the functionalities of molecules among cellular processes and its various roles in different prospects.

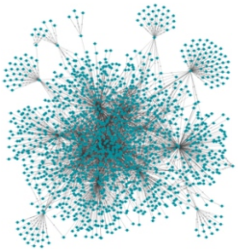

Figure 1: Graph of Transcriptional Regulatory Network of Mycobacterium tuberculosis.

Here is a short description of vital properties. The degree is an important concept used in network analysis. It quantifies the total connectivity of a particular node with other nodes present in a network. For directed network, degree is subdivided into two classes: Incoming edges (in-degree) and outgoing edges (out-degree), depending on the direction of interactions to its neighbour nodes. In a network there are some highly connected nodes (hubs) that play important role in the functioning of entire network at large. There are many nodes that are having only few connections and some isolated nodes with no connections to other nodes are also present in a network. In-degree and out-degree of a node is denoted by k_in and k_out respectively.

Degree distribution is global pattern of network to give knowledge about the degree. It is that probability of any node having degree k for k=1, 2,....,n.

A path from node v_1 to node v_n is a alternatively sequence of nodes and edges v_1 e_1 v_2 e_2…v_n such that e_1=(v_1,v_2), e_2=(v_2,v_3), … are edges of the graph. There might be lot of path between two nodes. The number of edges in a path is called path length of any path. A path between two nodes having smallest path length is called shortest path and numerical value (number edges in this shortest path) is called shortest path length. Various algorithms are exist to determine this shortest path length like Dijkstra Algorith, Johonson's algorithm etc.

The average of all shortest path length from a node to other connected node is called average path length of a node. The average of average path length of all nodes in network is called average path length of network or diameter of network.

Robustness is the ability of living systems to maintain functions in the phase of perturbations and uncertainty. Robustness can provide insight of cellular complexity and functioning. Biological systems maintain functional stability in the case of diverse perturbations arising from environmental changes, stochastic events (intra cellular noise) and genetic variations. It has been recognized that robustness is an inherent property of all biological systems. Cells face perturbations in the form of external perturbations through variable environments (modified inputs) and internal perturbations such as mutations that affect or remove components and interactions between two elements. Measuring the robustness of biological networks (systems) mean to determine the output or the Behavior of networks in response to the perturbations (Stelling, Sauer, Szallasi, Doyle, & Doyle, 2004).

The effect of removal of node will explore the importance of hubs or communities. Nodes (gene, protein, biomolecules etc.) or communities (group of genes etc.) are possible to remove in networks because of genes deletion or proteins degradation due to pharmacological treatment. It is impossible that the pharmacological treatment deduce side effect of drugs by analyzing topological properties of a network. Further, most of biological and other complex network follows scale-free behavior (R. Albert & Barabási, 2002). This scale-free network follows power law degree distribution followsP(k)~k^(-γ), interpreted as in homogeneously degree of node distributed through the network. It implies that low number of node having high degree nodes, rest are low degree node in the scale-free network. Barabasi and Albert gives an algorithm based on a preferential attachment principle. That gives a network which will follow power law degree distribution (Barabasi & Albert, 1999). These networks are surprisingly displaying degree of tolerance against random attack. The communication of node in network will have unaltered by random removing of nodes in these networks. Contrary, nodes will sparse as well as breaking in large number of small community whenever attack on high degree nodes in these networks takes place(null Albert, Jeong, & Barabasi, 2000).

In the current work, we are simulating the effect of gene deletion as in a case of mutation caused by natural circumstances and other molecular intervention. This implies loss of coding genetic material and corresponding encoded protein. In this paper, we are studying the transcription regulatory network of today's most threatening bacteria Mycobacterium tuberculosis that causes tuberculosis. As known today is one of the major life claiming agent worldwide that causes approximately 1.8 million deaths and infects 8 million people each year. Tuberculosis caused by Mycobacterium is very unusual among different pathogens and possesses the ability to survive in various environments in the host. A hallmark in tuberculosis infection is the shifting of pathogen between replicative and non-replicative (dormant) states over the resisting environment which is generated by host immune system. During the process of infection pathogen is advance enough therefore it is induced. Then it slows growth and eventually confines in the state of dormancy which enables pathogen to survive in an immunocompetent atmosphere for many years. This information confirms that the expression of different sets of genes at various stages of infection. It is also the resisting environment of host which is crucial to the survival and pathogenicity of Mycobacterium (Sanz et al., 2011). Simulation is based on directed and random deletion of gene in TRN of M. tb. The attack and tolerance of gene and protein (Transcription factors) of TRN of M. tb have vital role in discovery of drug and binding molecules. Those binding molecules may highly impact to causing the disease of TB.

2. Methods

2.1 Tools and Data

The construction of TRN of *M. tb* contributed of so many researcher. Primarily, Balazsi et. al. complied this network in 2008 then extended by Yamir et. al. through adding nodes and edges in the network in 2011. Network interaction data for construction of transcriptional regulatory network of Mycobacterium tuberculosis had been taken from the "complex systems and networks Lab (Cosnet)" Institute for Bio-Computation and Physics of Complex System, University of Zaragoza. Network consists of N=1624 nodes and E=3212 edges (Balázsi, Heath, Shi, & Gennaro, 2008) (Sanz et al., 2011). We have used cytoscape software to draw this network (Shannon et al., 2003).

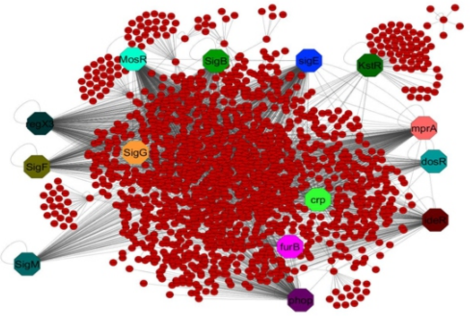

Figure 5: A layout of Transcriptional Regulatory Network of M. tb drawn using Cytoscape

Network Attack

In this work we have used MATLAB programming to analyze the network in order to check network robustness in two different modes of knock-out of genes from network. First, Raw data of TRN of M.Tb was converted to adjacency matrix. Adjacency matrix is defined as if edges exist between two nodes it takes as 1 otherwise 0, i.e., 〖A=[a〗 _ij], where a_ij = 1 if i and j nodes are connected, otherwise 〖 a〗 _ij=0. Shortest path length from i^th node j^th is calculated using Johonson's algorithm which is inbuilt in MATLAB function and matrix of shortest path '1_ij' was created and if there is no path between i^th & j^th nodes then l_ij will ∞. But here in our work we have taken infinity as a large value i.e. l_ij =100. Average path length (apl) of each node is calculated, i.e., a_i= average of shortest path from i to all other nodes in the network. Next, we calculate diameter of the network $d = \sum_{1}^{n} \frac{a_i}{n}$. $d_d(n),\ d_r(n)$ are calculated by removal of n highest degree (in-degree, out-degree) and n random nodes respectively. Here we calculate $d_d(n)$, $d_r(n)$, for $n = 1,2,3,\ldots\ldots 50$.

a. Diameter deviation from random network during attack

There are three parameters $D_o, \overline{D_R}$ and σ required to determine Z-score; $Z = \frac{D_o - \overline{D_R}}{\sigma}$, where D_o is diameter of TRN of M. tb; $\overline{D_R}$ is the average of diameters of 100 random networks; where we are generating random network of same characteristics using Erdos and Rainey algorithm.

Sigma (σ) is the standard deviation of diameters of 100 random networks. This Z-score physically implies that how much diameter of TRN of M.*Tb* is deviated from diameter of random networks. Now, in this article we are observing this Z-score of the n number of directed attack. Therefore, Z-score is function of n which is defined as:

$Z(n) = \frac{D_o(n) - \overline{D_R(n)}}{\sigma(n)}$, where Do (n) is diameter of TRN of M. tb for n number of highest degree of nodes removed for undirected graph. Analogously, (D_R (n) and σ(n) are the mean and standard deviation of diameter of all random networks for n number of removal of highest degree nodes respectively. *The* result of directed attack performed on TRN of *M. tb* is compared to the result of directed attacked performed on random network with same characteristics.

3. Result and Discussion

3.1 Hubs and Degree distribution

Network having 1624 genes and 3212 interactions, the majority of nodes occurring represent less connections with neighboring genes (degree), while there are few genes (nodes) that are highly connected (hubs). Hubs plays vital role in functionality of the network. Nodes represent genes and link (edges) represents their regulatory interactions in TRN. The nodes of a graph can be characterized by the number of connected edges that they have. This property is called the degree of the nodes in the network. In fig. 6 is depicting top 20 highest degree nodes when the graph considered as undirected. In directed networks we distinguish the in-degree and out-degree. The number of directed edges that point toward the node which is depicted that gene is regulated by others transcriptional proteins where those proteins was translate followed by transcribed by other genes. In-degree of genes are not so high, the highest in-degree of a node is 12. Fig. 7 is showing the top 20 highest in-degree nodes. Out-degree is the number of directed edges that start at the node which depicted that a gene regulates number of proteins whose are Transcriptional factors. These transcriptional factors are going to bind in same of genes. In fig. 8 is showing top twenty highest out-degree genes.

Degree of nodes characterizes individual node whereas one can define a degree distribution to quantify the diversity of the whole network. Degree distribution represents the number of node having degree r degree, where =1,2,3,… . In fig. 9 showing degree of node is on x-axis and number of nodes having such degree is on y-axis. This plot is reflect that the large number of nodes having low degree and as degree increase the number of nodes decrease drastically. Similarly, the in-degree and out-degree distribution are plotted in fig. 11 and fig. 13. Further, we have fitted degree distribution in a curve P(k)=*Ak*(-γ). In fig. 10, 12 and 14 are showing log-log plot of degree, in-degree and out-degree distribution and their curve fitting respectively. Degree distribution and out-degree distribution are follows power law but in-degree distribution is not follows power law. In-degree distribution is almost uniformly distributed.

Hubs (high degree genes) have been characterized by having high degree nodes in a network. They affect every node either directly or indirectly in some way. An effective number of hubs are supposed to exist as is the case of M. tb with around 14 hubs throughout the network. Theses fourteen hubs are having degree more than fifty. Such a large number of hubs in a network possess potential to withstand the external perturbation and are lesser prone to any damage caused by malfunctioning of the system than the One with less number of hubs. Hubs, in most of the cases, are found to be regulating elements. In transcriptional regulatory network (TRN) of M. tb, occurrence of all 14 hubs represents its robustness against any kind of perturbation and each hub is possessing unique regulatory feature demonstrate that how précised and well-tuned transcriptional processes takes place in this pathogen. In fig. 15 highlighted those hubs genes in TRN whose have already drawn in fig. 6 as bar diagram.

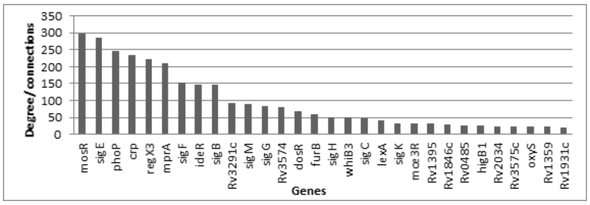

Figure : Name of top thirty highly connected genes in transcrip tional process of M. *tb.*

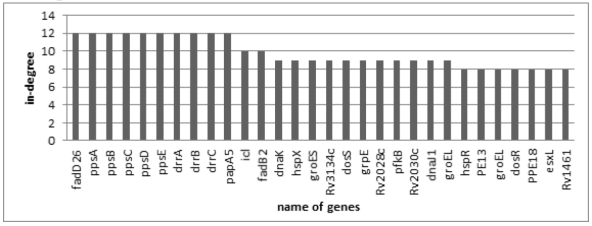

Figure : Name of thirty highest incoming connected edges genes in directed TRN of M. *tb*

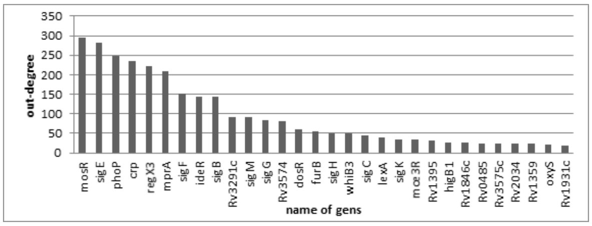

Figure : Name of thirty highest outgoing connected edges genes in directed TRN of M. *tb*

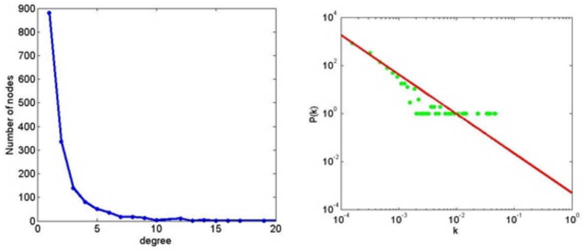

Figure 9: Degree distribution of undirected TRN of M. *tb*

Figure 10: log-log plot of degree distribution and fitted in ln *P*(*k*)=ln (*A*= with A=5.105×10^{-4} and *γ*=1.642.

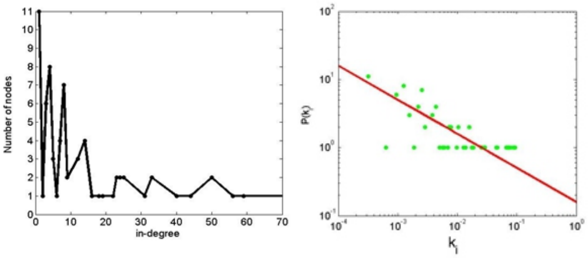

Figure 11: In-degree distribution of directed TRN of M. *tb*

Figure 12: log-log plot of in-degree distribution and fitted in ln P(k)=ln (A×k(-*γ*)), with A=0.1587 and *γ*=1.642.

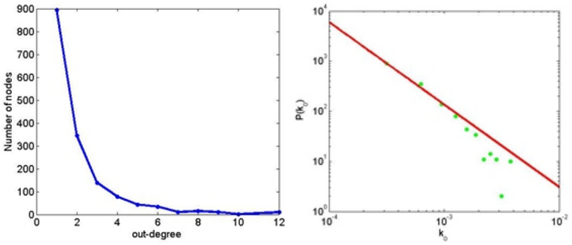

Figure 13: Out-degree distribution of directed TRN of M. *tb.*

Figure 14: log-log plot of out-degree distribution and fitted in ln P(k)=ln (A×$k^{-γ}$) , with A=1.589×10^{-3}) and γ=1.644

Figure 15 : highlighted hubs (genes) in TRN of M. *tb.*

3.2 Average path length and its distribution

The average path length of a node in the network is defined as the average of shortest path lengths from a node to other connected and unconnected nodes where shortest path length of unconnected pairs of nodes consider as 100. Average path length is one of the three most robust measures of network topology, along with its clustering coefficient and its degree distribution. Fig. 16 is a snapshot of average path length distribution which is drawing the average path length of genes in increasing order. In fig. 16 depicting 350 genes out of 1624 gene have about 80 average path lengths and rest are drastically high average path length. Therefore, those remaining genes are showing highly disconnectedness. It is implies that there are some gene having vital role in transforming the information as well as regulatory process in M. tb. Out of 350 genes name of 30 smallest average path lengths can be seen in fig. 18. This scenario is little bit different than the directed graph. When we consider the graph is directed then about 15 genes have low average path length. It implies that these 15 genes are highly involved in mediated information and significantly involve in the transcriptional regulation process. A snapshot of average path length distribution in directed TRN can be seen in fig. 17. In fig. 19 depicted those genes whose are significantly connected one gene to other gene in transcriptional process when we are preferring directions.

Here in this network of mycobacterium tuberculosis, the 30 most connected genes (hubs) and their average path length are measured. The path length have been crucial in any network as is the case of M. tb where mosR (most connected hub) has the shortest path length and therefore highly vulnerable to perturbation. In the fig. 20, thirty genes have been represented in ascending order of average path length based on their links with other genes in the network. This shows that the information between these nodes will flow rapidly in a very short period of time. The average path length does not only depend on the maximum links of nodes but it also gets standardized by the distance between the neighbouring nodes that how closely two connected nodes are. Therefore, assumption based on the links of the nodes in the network will set one more criterion prior to average path length for the sake of examining the robustness of any network.

In this network 'sigE' has lowest average path length followed by 'mosR'. Removal of nodes that have large average path length tends to be robust in contrary to the directed attack or removal of nodes that have low average path length. Removal of such nodes will lead to breakdown of network into isolated clusters which will hinder in global coordination in information flow in the network.

Basically, 30 nodes having higher connections were taken and subjected to the speculation of average path length. The nodes with maximum degrees in the network show minimal path length because higher connectivity that allows each node to exchange information at a rapid rate.

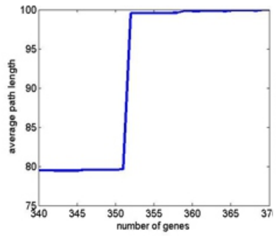

Figure : Snapshot of distribution of average path length in undirected TRN of M. tb.

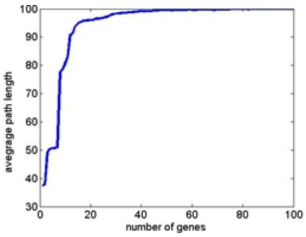

Figure : Snapshot of distribution of average path length in undirected TRN of M. tb.

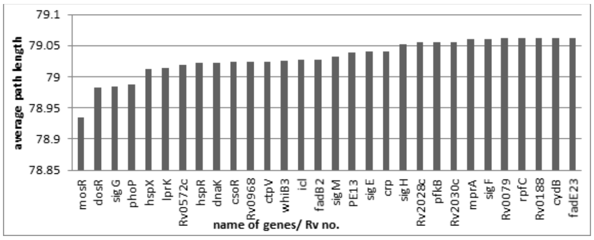

Figure 18 : Name of 30 smallest average path length genes in undirected TRN of *M. tb.*

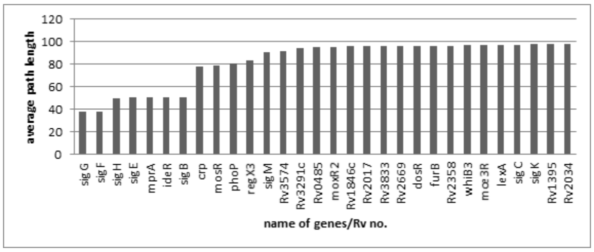

Figure 19 : Name of 30 smallest average path length genes in directed TRN of M. tb.

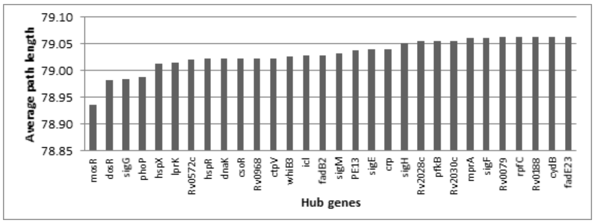

Figure 20 : Depiction of average path length of 30 TRN nodes. These nodes are having maximum links in the network and are represented with their ascending average path lengths.

4 Robustness of Transcriptional Regulatory Network (TRN) of *M. tb*

Fig. 21 represents the diameter of undirected TRN of *M. tb* versus number of removing nodes with corresponding edges from the network. Initially, we have seen diameter of network is about 7 when we have taken 100 shortest path lengths for unconnected pairs of nodes. This implies that the network is highly connected when the direction of edges ignored. Now, we are seeing the diameter of network after removing numbers of nodes from the network. Here, we are removing nodes in two manners, one is random remove and other is directed attack, i.e., remove top highly connected node. Whereas in fig. 22 represents random, highest in-degree and out-degree removal of nodes in TRN of *M. tb*.

In case of random attack (removal) as well as directed in-degree node attack against nodes, the diameter of TRN of M. tb increases very slowly. This is due to the wonderful architecture of this network having maximum nodes with connection to other nodes in the network. Therefore, network remains unaffected over their removal or knock-out leaving diameter uninfluenced in size (navigability). Under a directed mode of attack, removal of node make network greatly crucial in context of sensitivity and vulnerability against interference, a phase-transition is observed breaking the network into tiny, non-communicating islands of nodes that influences network diameter to increase in many folds. After the 20th knock-out, the diameter levels off: the directed attack in undirected network and directed out-degree attack in directed network has removed all the hubs; the increase in diameter becomes constant. The network gets scattered into isolated sub-clusters and the majority of routes connecting the nodes got lost.

This network is the 'Scale free' network which is very robust against any accidental failures even if 80% of randomly selected nodes are exposed to malfunction or accidental failure, the remaining 20% still form a compact cluster with a path connecting any two nodes. This is because random failure affects mainly the numerous small degree nodes, the absence of which doesn't disrupt the network's architecture. Hubs nodes are uncommon and so they are rarely hit in random mode of attack and thus orchestrate entire network cooperatively. This architecture of TRN of M. tb and other real networks depends basically on hubs for overall functioning of network; directed attack on hubs will collapse network architecture at large.

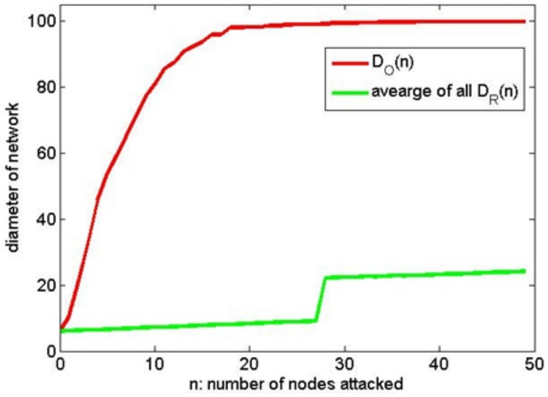

Figure 21 : A plot of network diameter for the first 50 knock-outs by two modes of attack. This plot represents degradation of network diameter when nodes are knocked out in either a random pattern or in directed attack against hubs in undirected TRN of *M. tb.*

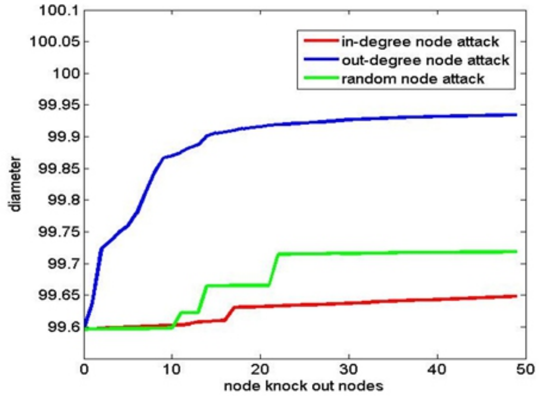

Figure 22 : A plot of network diameter for the first 50 knock-outs by two modes of attack. This plot represents degradation of network diameter when nodes are knocked out in either a random pattern or in directed attack against hubs directed TRN of M. tb.

In every biological system, as it has been explained previously, almost all protein/gene interaction processes are substantially robust contrary to any specific malfunctions due to mutations and external perturbations. The same is the case of subjected Mycobacterium tuberculosis TR-networking processes. Since this organism has been the second most life claiming agent, its genes and proteins have always been a focal point with regard to the development and discovery of therapeutic drug. As we have put forth the implications of system biology into novel drug discovery, we may also consider and evaluate it for the result obtained in this research work, not only for this pathogen but for all of them, that diameter could have been set as a criterion for standardizing the process of finding a drug target.

5 Z-Score

We have studied direct and random attack in TRN of M. tb. Now, we are seeing only directed in undirected TRN of M. tb as well as in theoretical random networks with same number of node and edges of TRN of M. tb. In fig. 23, showing diameter of undirected TRN of M. tb where we have 500 shortest path length of disconnected pairs of nodes in network. Diameter is reaching at constant close to 500 after 20 attacks of nodes in network; it implies that almost all pair of nodes is disconnected after 20 attack of nodes. In fig. 24 is showing that the directed attack in various random networks with same number of node and edges of TRN of M. tb, it is depicting diameter is so high after 15 directed attack of node, but it not reaching very close to 500. So, the whole network is partially connected after 15 attack of node in network.

Therefore, in fig. 25 is number of node knock out versus Z-score of network. We are observing Z-score after each direct attacking of node in the TRN of M. tb. Up to removing 7 highest degree nodes then the diameter of TRN of M. tb is highly deviated from random network. After attacking of 10 nodes, behaviour of diameter in both original as well as theoretical random network is same. So, highest degree nodes are highly significant in many senses. These hubs (genes) are vital role in various aspect of study.

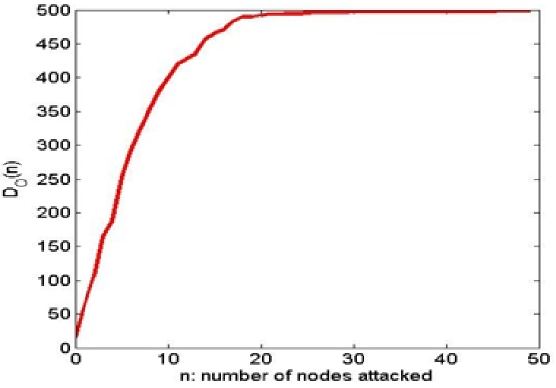

Figure 23 : Diameter of undirected TRN of M. tb for directed attack against node versus number of directed attack nodes.

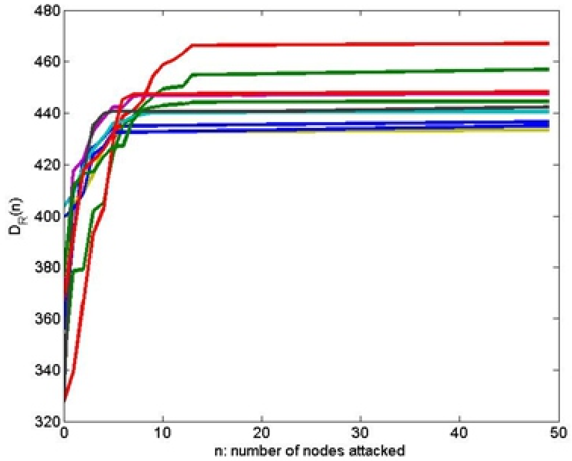

Figure 24 : Diameter of various random networks for directed attack against node versus number of directed attack nodes.

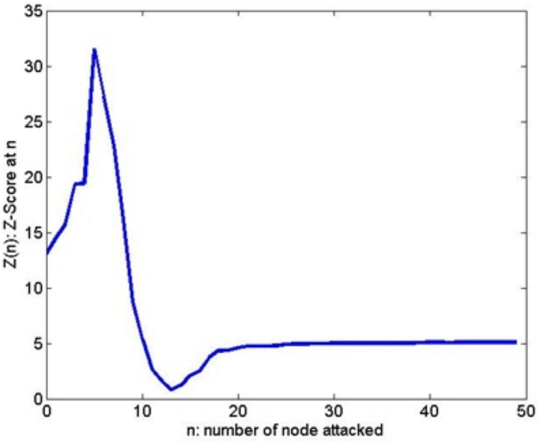

Figure 25 : Z-score of diameter for undirected TRN of *M. tb* for directed attack against node versus number of directed attack nodes.

6. Conclusion

Over the period of molecular evolution, organisms have evolved substantially robust system to the damage caused due to mutations and external effects and has empowered their structural and functional unit so as to remain resilient in the varying environment with time. The molecular base in Mycobacterium tuberculosis has been established comparatively much better than other tubercle pathogen. Throughout the regulatory processes in protein-protein and protein-gene interaction networks, pathogen have been found to be continuously engaged in rewiring of the regulatory elements since our results have demonstrated that robustness is greatly developed in transcriptional regulatory network. To extent of robustness in M. tuberculosis, random attack is always proved to be least prone to the overall functionality and topological properties of the network. The directed mode of attack demonstrates failure of the system by breaking and damaging network exclusively. It also represents the abolishment of information flow in between the biochemical processes by just showing the increase in overall diameter of the entire network. By the result obtained, we may also consider and evaluate that diameter and robustness could have been set as a criterion for standardizing the process of finding a drug target.

Acknowledgment

The authors are highly obliged to the Centre for Interdisciplinary Research in Basic Sciences, Jamia Millia Islamia, New Delhi for providing research grant under Innovative Research Activities.

References

1. Albert, null, Jeong, null, & Barabasi, null. (2000). Error and attack tolerance of complex networks. Nature, 406(6794), 378–382. https://doi.org/10.1038/35019019
2. Albert, R., & Barabási, A.-L. (2002). Statistical mechanics of complex networks. Reviews of Modern Physics, 74(1), 47–97. https://doi.org/10.1103/RevModPhys.74.47
3. Balázsi, G., Heath, A. P., Shi, L., & Gennaro, M. L. (2008). The temporal response of the Mycobacterium tuberculosis gene regulatory network during growth arrest. Molecular Systems Biology, 4. https://doi.org/10.1038/msb.2008.63
4. Barabasi, null, & Albert, null. (1999). Emergence of scaling in random networks. Science (New York, N.Y.), 286(5439), 509–512.
5. Mason, O., & Verwoerd, M. (2007). Graph theory and networks in Biology. IET Systems Biology, 1(2), 89–119.
6. Sanz, J., Navarro, J., Arbués, A., Martín, C., Marijuán, P. C., & Moreno, Y. (2011). The Transcriptional Regulatory Network of Mycobacterium tuberculosis. PLoS ONE, 6(7), e22178. https://doi.org/10.1371/journal.pone.0022178
7. Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., … Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Research, 13(11), 2498–2504. https://doi.org/10.1101/gr.1239303
8. Stelling, J., Sauer, U., Szallasi, Z., Doyle, F. J. 3rd, & Doyle, J. (2004). Robustness of cellular functions. Cell, 118(6), 675–685. https://doi.org/10.1016/j.cell.2004.09.008