Mathematical Modelling of Layered Cellular Network

Tejas Sanjay Wagh

A dissertation submitted for the partial fulfilment of BS-MS dual degree in Science

Indian Institute of Science Education and Research Mohali
November 2017

Certificate of Examination

This is to certify that the dissertation titled **Mathematical Modelling of Layered Cellular Network** submitted by **Tejas Sanjay Wagh** (Reg. No. MS12130) for the partial fulfillment of BS-MS dual degree program of the Institute, has been examined by the thesis committee duly appointed by the Institute. The committee finds the work done by the candidate satisfactory and recommends that the report be accepted.

Dr. Kuljeet Sanndhu (Supervisor) Dr. Shashi Bhushan Pandit Dr. Lingraj Sahu

Dated: December 4, 2017

Declaration

The work presented in this dissertation has been carried out by me under the guidance of Dr. Kuljeet Sandhu at the Indian Institute of Science Education and Research Mohali.

This work has not been submitted in part or in full for a degree, a diploma, or a fellowship to any other university or institute. Whenever contributions of others are involved, every effort is made to indicate this clearly, with due acknowledgement of collaborative research and discussions. This thesis is a bonafide record of original work done by me and all sources listed within have been detailed in the bibliography.
I, Tejas Wagh, also take the sole responsibility of any possible plagiarism this thesis might reflect to the readers.

Tejas Sanjay Wagh
(Candidate)

Dated: December 4, 2017

In my capacity as the supervisor of the candidates project work, I certify that the above statements by the candidate are true to the best of my knowledge.

Dr. Kuljeet Sandhu
(Supervisor)

Acknowledgment

Firstly, I would like to thank my project guide Dr. Kuljeet Sandhu for his expert guidance. I have been amazingly fortunate to have adviser like him who gave me an opportunity to explore the subject and at the same time giving me the valuable guidance. I will forever be grateful to you.

This journey of five and half years would not have been possible without my friends. I would like to thank Himanshu, Tarun, Amit, Rohan and Vinay for making these five years memorable. Thank you for inspiring me throughout the journey and for making these years unforgettable. Thank you everyone this. This would not have been possible without you.

I fall short of words to thank my parents for being very supportive and encouraging throughout my whole life. Their support and unconditional love have helped me and guided through life.

Abstract

We live in a very complex and dynamical world. From minuscule molecules to giant human social network we can clearly see each single unit in these system is interacting among themselves as well as units from outside the system. And these collective interaction among the units of these systems gives rise to the fascinating phenomenon that we see in our surrounding. To understand this phenomenon from mathematical perspective we have used the framework of Networks in my thesis as we are not just looking at individual units but also the different interactions among them. A network is the framework that comprises of set of nodes or individual units in the system and interactions among those units. Since the real world network are complex and different nodes may have different kind of interactions, we have to extend this idea of Network to Multilayer Network where in each layer nodes are having an unique intralayer interactions and while doing so they are also interacting from one layer to another in a totally separate way by interlayer interactions. For example, in a cell proteins are interacting among themselves but also interacting with DNA.

I have worked on the this same Protein-DNA system and to describe effect of perturbation in protein-protein interaction network I have used the model give in the paper "Propagation of large concentration changes in reversible protein-binding networks"[MI07] and how proteins regulate gene transcription I have used Hill function. The aim of my thesis is to come up with a multilayer network model to see how perturbations in protein-protein interaction network can affect transcription activity of gene.

Contents

Chapter 1

Networks

1.1 Introduction

To understand complexity and dynamics of cellular molecular network, I will be using the framework of multilayer networks. In this chapter I have tried to explain very basic idea of how multilayer networks can be understood mathematically.

1.2 Graphs,Adjacency Matrix and Adjacency list

Mathematically a network can be understood as a graph. A graph G formally defined as an ordered pair $G = (V, E)$ where V is set of vertices and E is set of edges which is subset of $V \times V$.[New09].

1.2.1 Adjacency Matrix

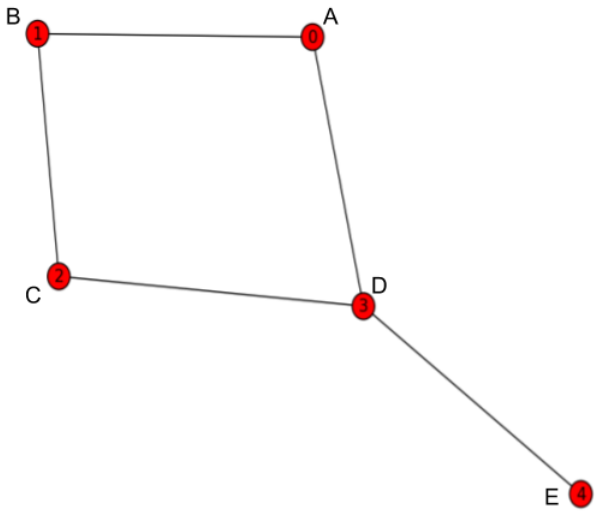
Adjacency matrix is one of the mathematical representation of graphs. Entries of Adjacency matrix are defined as [New09]:

$$\mathrm{A}_{ij} = \begin{cases} 1, & \text{if } i \text{ is connected to } j \\ 0, & \text{otherwise} \end{cases}$$

If the the graph is directed then:

$$A_{ij} = \begin{cases} 1, & \text{if there is an edge from } j \text{ to } i \\ 0, & \text{otherwise} \end{cases}$$

$$\begin{pmatrix} 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

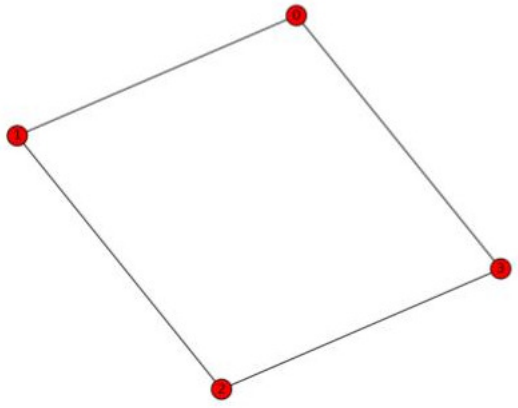


Adjacency matrix and Graph

The Adjacency matrix helps us define very basic structure of graph. If in a graph an edge connecting any two nodes have some weight associated with it then instead of 1 that weight can be an entry in the adjacency matrix. Such an adjacency matrix representing weights of the edges of the graph is called as weighted adjacency matrix. If the graph is undirected then the matrix is symmetric and if it is directed the it is not symmetric as in that case $A_{ij} \neq A_{ji}$.

1.2.2 Adjacency List

An adjacency list representation for a graph associates each vertex in the graph with the collection of its neighboring vertices or edges[New09].



Edge list for an undirected graph:

0	1
0	3
1	2
2	3

1.3 Multilayer Networks

Multilayer networks are basically extension of simple planer networks. Consider two social network websites Facebook and twitter. Users are nodes in these networks. These nodes can be same in both networks or one layer may have more nodes than the other. In each layer users share information in different way therefore the interactions or edges on each network are different called as intralayer interactions. These users can share content of one layer to other that is there are also some interlayer interactions. Together these two layers form a multilayer network[BBC$^+$14].

I will be using this concept of multilayer network to describe how dynamical changes in protein-protein interaction network can affect gene's transcription activity in DNA network(layer 2) because of protein-DNA interlayer connections.

A multilayer network is a pair $\mathcal{M} = (\mathcal{G}, \mathcal{C})$ where $\mathcal{G} = \{G_\alpha; \alpha \in \{1, 2, ..., M\}\}$ is a family of(directed or undirected, weighted or unweighted) graphs.

$G_\alpha = (X_\alpha, E_\alpha)$ (called layers of M) and

$\mathcal{C} = \{E_{\alpha\beta} \subseteq X_\alpha \times X_\beta \; ; \alpha, \beta \in \{1, 2, ..., M \; , \alpha \neq \beta\}$

is the set of interconnections between nodes of different layers G_α and G_β with $\alpha \neq \beta$.

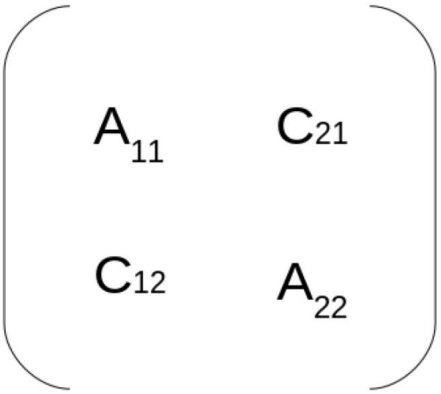


Figure 1.1: Adjacency matrix for multilayer network

The elements of $\mathcal{C}$ are called crossed layers , and the elements of each E_α are called interlayer connections of $\mathcal{M}$ in contrast with the elements of each $E_{\alpha\beta}$ that are called interlayer connections[BBC$^+$14].

The adjacency matrix representation of multilayer network is given in Fig(1.1)

where A_{11} is Adjacency matrix for layer 1,

A_{22} is adjacency matrix for layer 2,

C_{12} is Adjacency matrix for connections from layer 1 to layer 2.

C_{21} is Adjacency matrix for connections Connection from layer 2 to layer 1.

Chapter 2

Protein-Protein Interaction Networks

In Transcription network interactions among proteins is very important dynamical aspect. In the model that I have developed each individual node in protein-protein interaction network represent individual proteins and their weights on the nodes represent the concentration associated with the nodes. Before we move on to gene regulatory network, first let's capture the dynamics of Protein-Protein interaction(PPI) networks. Here in my thesis I have analyzed two models of PPI networks:
1) Propagation of perturbation using communicating vessel model [SC13]. And
2) Propagation of large concentration changes in reversible protein-binding networks [MI07]

2.1 Propagation of perturbation using communicating vessel model[SC13]

The basic idea behind the model was that intensive physical variables (e.g. temperature) tend to perform an equalization-like dynamics behaving like communicating vessels. For example in a thermodynamical network system,it explains how temperatures of nodes will get redistributed to an equilibrium if any of the nodes' are heated. Therefore this model can also be thought of as a heat equation equivalent. It can also be viewed as network for water containers connected through pipes. If one containers water content is increased, all others show increase in level of water till all come to equilibrium. In the communicating vessels model network nodes represent the vessels and edges represent their connecting pipes. The algorithm of the model is as follows:

in each time step, every node transfers a proportion of its available energy through every available edge, proportional to 1) the duration of the time-step; 2) the weight of the edge ; and 3.) the difference of the weight of the nodes on the two ends of the edge at that time instant (corresponding to pipe pressure). A very important dynamical aspect of these systems is that a constant amount of available energy could be dissipated in environment during diffusion.

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -\sum_{i=1}^{l}\left(\frac{S - S_i}{2}w_i\right) - D_o$$

where S is the energy of the current node,
l is the number of edges of the current node, w_i is the weight of the i^{th} edge, S_i is the current energy of the node on the other end of the i^{th} edge and
D_0 is a parametere that defines energy dissipated in that time step. An important practical restriction is that: -1 $\leq \Delta t \sum_{i=0}^{l} \mathrm{w}_i \leq 1$.

Given below is the undirected graph with six nodes with equal edge weight of 0.2 and concentration defined on each node at t=0 are: 4.5, 3.0, 6.6, 2.6, 8.0, 5.0.

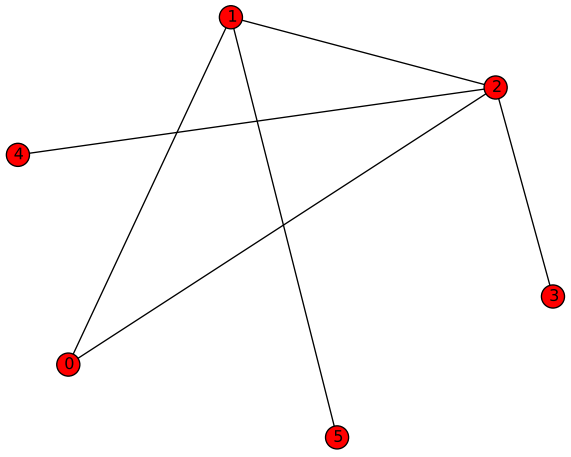


Figure 2.1: Network to implement Communicating vessel model

After implementing communicating vessel model on this network and perturbing 2^{nd} node by +10 (here we have assumed $D_0 = 0$) we observe following:

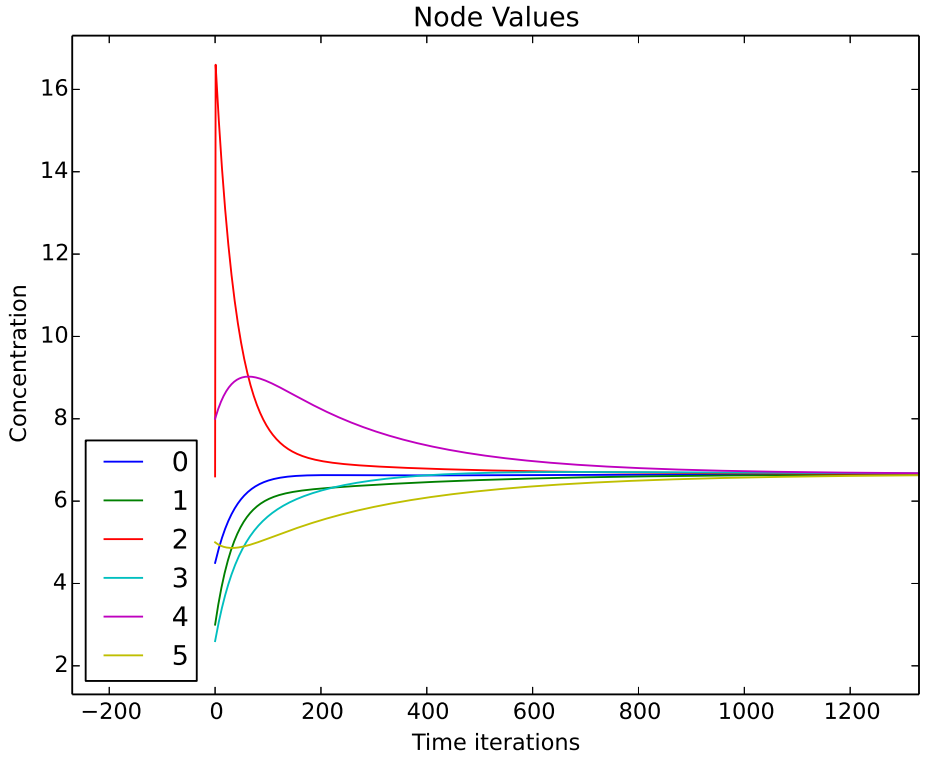


Figure 2.2: Dynamics of communicating vessel model

I implemented this model using Python programming language. t=1 concentration of node 2 is perturbed from from 6.6 to 16.6. For later time instances, concentration of node 2 decreases exponentially, whereas nodes which are directly connected to node 2 which are nodes 0,1,3,4 show rapid increase in concentration over successive time iteration before whole system comes to an equilibrium due to overall increase of concentration of system. Node 5 which is not directly connected to perturbed node 2 show relatively slow rate of increase in concentration. Around t=1559 the whole system attains an equilibrium concentration where concentration on each node is 6.655 units.

While communicating vessel model provides an insight into the dynamics of concentration change on a network it fails to capture many other real world molecular biology parameters.

2.2 Propagation of large concentration changes in reversible protein-binding networks[MI07].

This model is derived using law of mass action. What I find very interesting about this model is that,unlike turbine model, if a node's concentration is increased, free concentration of adjacent nodes of it does not necessarily show increase in concentration. Some free concentrations go up and some go down. Consider free concentration of two proteins i and j to be $[F_i]$ and $[F_j]$ respectively. The main equation of equilibrium in reversible protein binding reaction is [Sne06]:

$$[F_i][F_j] = k_{ij}[D_{ij}]$$

which expresses the free concentration for two proteins i and j in terms of their heterodimer D_{ij}. k_{ij} is dissociation constant. In this model k_{ij} is defined as $max(T_i, T_j)/20$ because PPI data sets lack information on dissociation constant of each individual interaction[MI07]. Each free concentration F_i, is in turn related to its total concentration T_i by:

$$[T_i] = [F_i] + \sum_{dimers} [D_{ij}] = [F_i] + \sum_{dimers} \frac{[F_i][F_j]}{k_{ij}}$$

where the sum run overs all links in the protein-protein network. Thus the sum takes into account all dimers in the network. The above equation can be solved iteratively:

$$[F_i] = \frac{[T_i]}{1 + \sum_{dimers} \frac{[F_j]}{k_{ij}}} \tag{2.1}$$

starting with F_i=T_i for all i.

When we implemented model with the help of a Cytoscape module called Perturbation Analyzer on a small modelled network we found the change in free concentration of proteins after perturbing node 2 as shown in Fig(2.3).

Here we have perturbed total concentration of node 2 by 2 fold. We can clearly see the here from the Figure(2.3), free concentration on the nodes 2,4,5 and 6 has increased while on remaining nodes it has gone down. After perturbation free concentration on nodes 1,2,3,4,5,6 and 7 changed by 0.922693795509693, 5.210974874370512, 0.5094650501304583, 1.433574490863499, 1.0218143432286024, 1.0342504224417308, 0.5698561556035878 fold respectively.

In my thesis I will be using this Protein perturbation model to describe PPI dynamics and to calculate bound concentration.

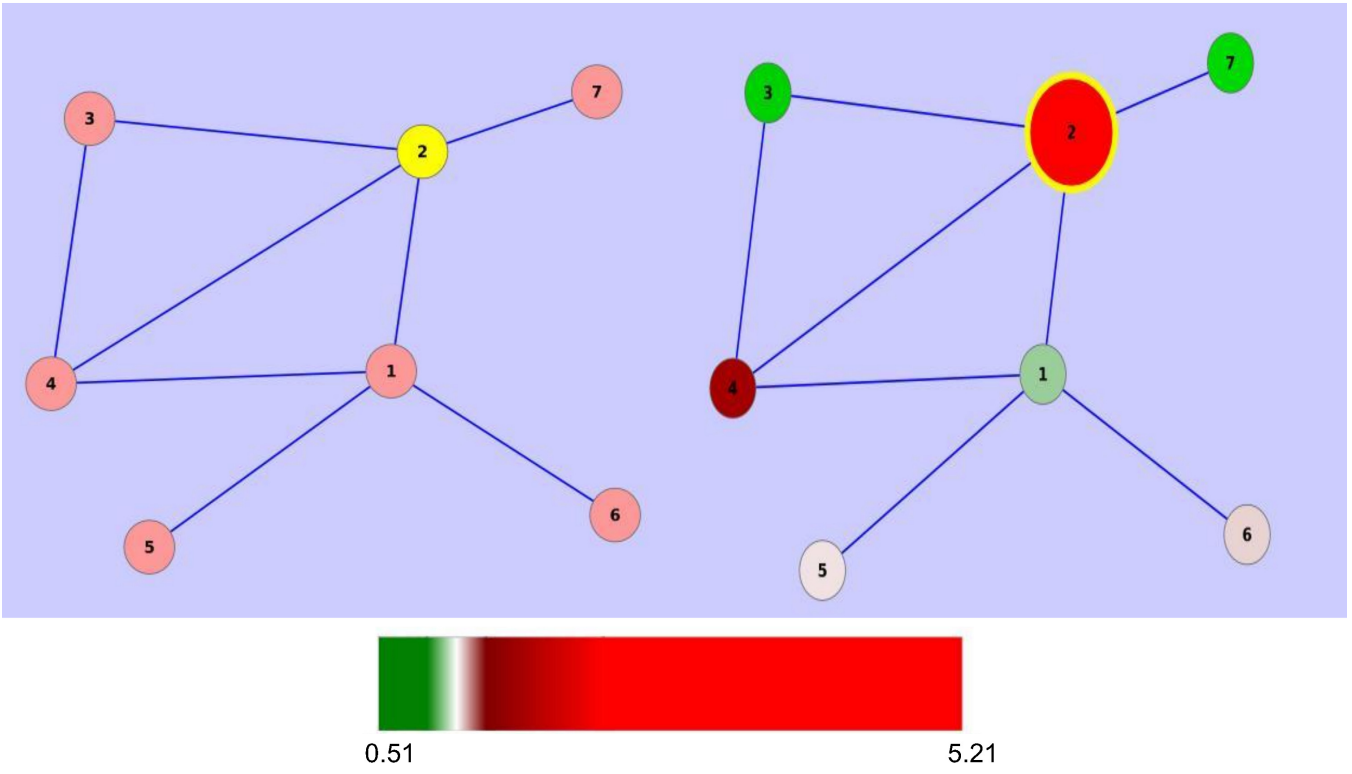


Figure 2.3: Change in Free Concentration

Chapter 3

Transcription Network

3.1 A Brief Idea

Because of the complexity of the environment that cells live in they are exposed to variety of signals including physical parameters such as temperature and osmotic pressure, biological signaling molecules from other cells, harmful chemical etc. In response to these signal they produce appropriate proteins that act upon the internal or external environment. The cell uses special proteins call transcription factors to represent these environmental states. Transcription factors transit rapidly from active and inactive molecular state. Each active transcription factor can bind the DNA to regulate the rate at which a specific gene are read and translated into messenger RNA(mRNA). These mRNA molecules later play a very important role in producing proteins which act upon environment.The rate at which gene is transcribed is controlled by the promoter, a regulatory region of DNA that precedes the gene. Transcription factor can act as a activator that increase the transcription rate of gene or a repressor that reduce the transcription rate[Alo06]. Fig.3.1 shows the transcription network and its elements.

3.2 Input Function

The effect of the environmental signals on gene's transcription activity by transcription factors is captured by the Input Functions[Alo06]. Lets consider first the production rate of protein Y controlled by a single transcription factor X. When X regulates Y, represented by $X \rightarrow Y$. the number of molecules of Y produced per unit time is a

function of bound concentration of X in its active form, X^* :

$$\text{Rate of production of } Y = f(X^*)$$

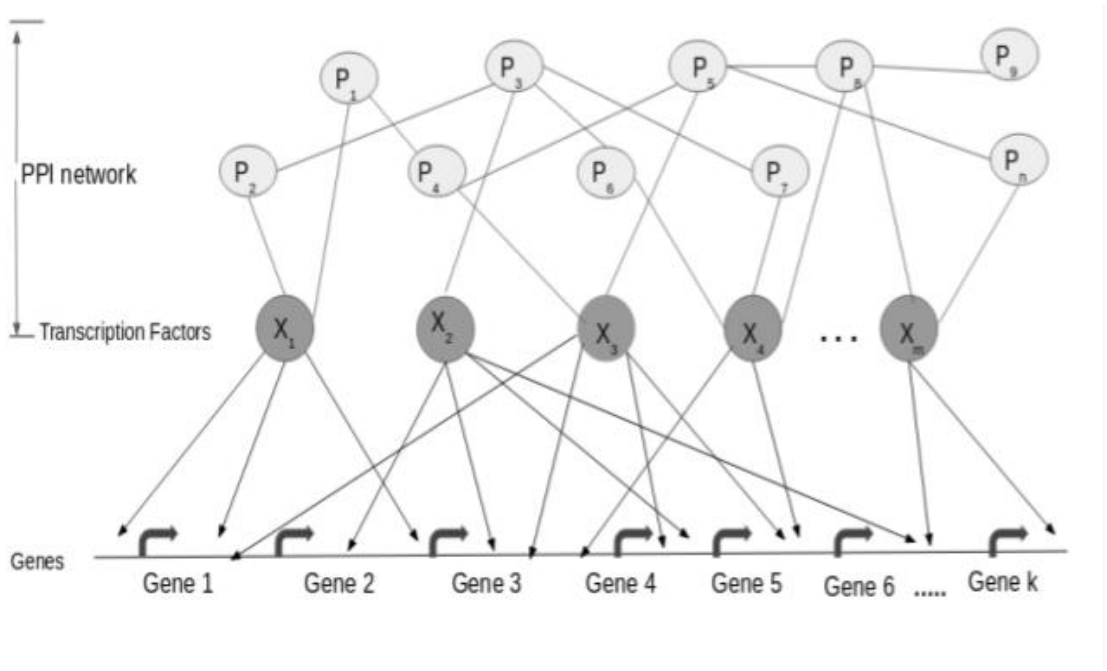


Figure 3.1: Transcription Network

Typically the input function $f(X^*)$ is a monotonic, S-shaped function. When X is an activator, it is a increasing function. The function that I will be using in my model describes many real gene input functions is called the Hill function.

3.2.1 Hill Function

$$f(X^{*^n}) = \frac{\beta X^{*^n}}{K^n + X^{*^n}}$$

The Hill function given above has three parameters, . The first parameter, K is termed as activation coefficient. This parameter has units of concentration. It defines concentration of active X needed to significantly active expression[Alo06]. The second parameter in input function is the maximal expression level of the promoter β. Maximal expression is reached at high activator concentration. The last parameter is

the Hill coefficient governs the steepness of Hill function. The larger the n value, the more step like the input function. Like many functions in biology, the Hill function also reaches saturation at high levels of X^*. This saturation of Hill function at higher values of X^*.

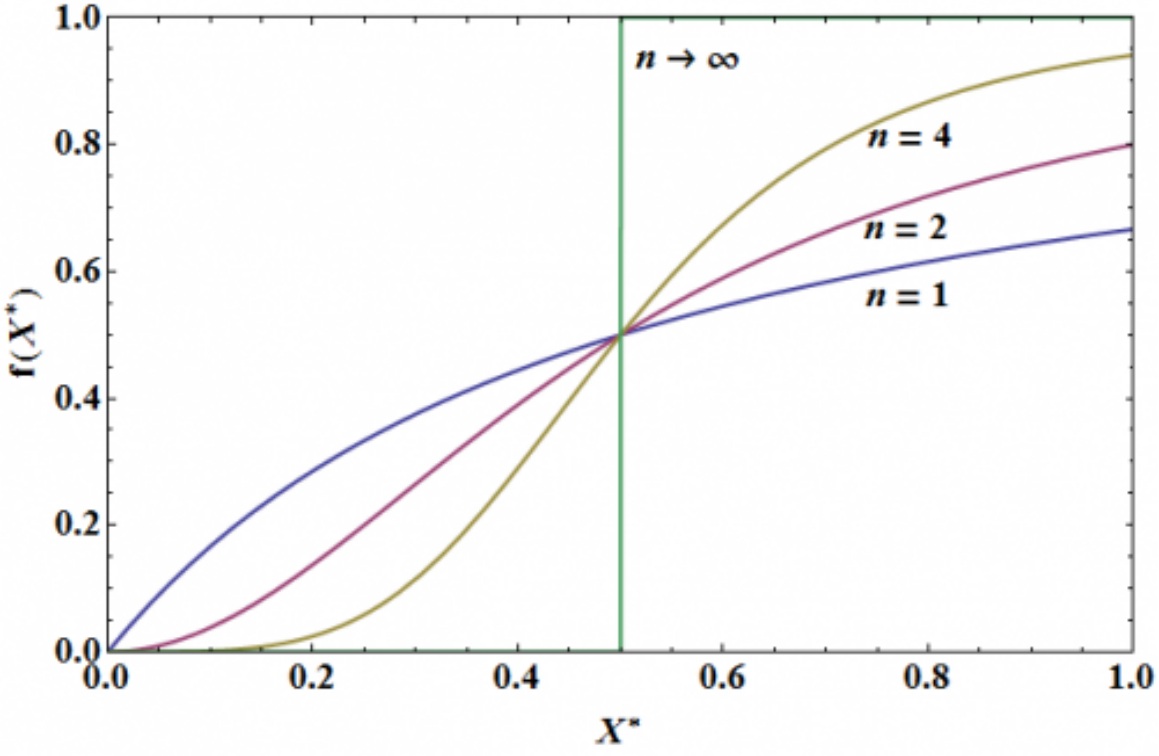


Figure 3.2: Hill Function

3.2.2 Logic Input Function

The idea behind this input function is transition from low value to high when the variable crosses characteristic threshold value K. Therefore logic input functions are step like approximation for the smoother Hill function. In this approximation the gene is either in OFF state, $f(X^*) = 0$ or in ON state, $f(X^*) = \beta$. The threshold of approximation is K[Alo06]. For activator the logic function can be described by using a step function θ that makes a step when X^* exceeds K. :

$$f(X^*) = \beta\theta(X^* > K)$$

There are many instances where a gene is regulated by more than one transcription factor. Let's consider the gene regulated by two activators. Many genes require binding of both activators proteins to the promoter in order to show significant expression[Alo06]. This is similar to AND gate:

$$F(X_A^*, X_B^*) = \beta\theta(X_A^* > K_{X_A})\theta(X_B^* > K_{X_B}) \sim \mathrm{X}_A^* \text{ AND } X_B^*$$

3.3 Dynamics of Transcription Network

In my thesis I have made an attempt to develop a model in which a gene requires two activators. To do so, the function that I have used follows the logic that, it will have the basic characteristic of AND gate and individual functions are Hill functions. Therefore input function here will be:

$$F(X_A^*, X_B^*) = f(X_A^*)f(X_B^*)$$

where f is Hill function and X_A^* and X_B^* are bound concentration of activator.

The network that I have used here is a subnetwork of PPI netwok and Protein-DNA interaction network of mouse embryonic stem cell(Fig.3.3) The target genes are yellow nodes while rest is PPI network.

3.4 Analysis

In my thesis, I have used the data sets available for mouse embryonic stem cells. PPI network data and Protein-DNA interaction data is readily available [XAS$^+$14][XBD$^+$13]. In the analysis I have perturbed protein nodes by layer. Layers are defined by protein distance from target genes. Proteins which are one hand shake distance away from gene are in layer 1, proteins at distance two are in layer 2 and so on. I have made analysis till distance 4. As I am dealing with the subnetwork here,there are only 7 target genes which I will be considering(Yellow nodes in Fig.3.3). This whole PPI network mode and its effect on gene regulation is implemented using Python programming language.

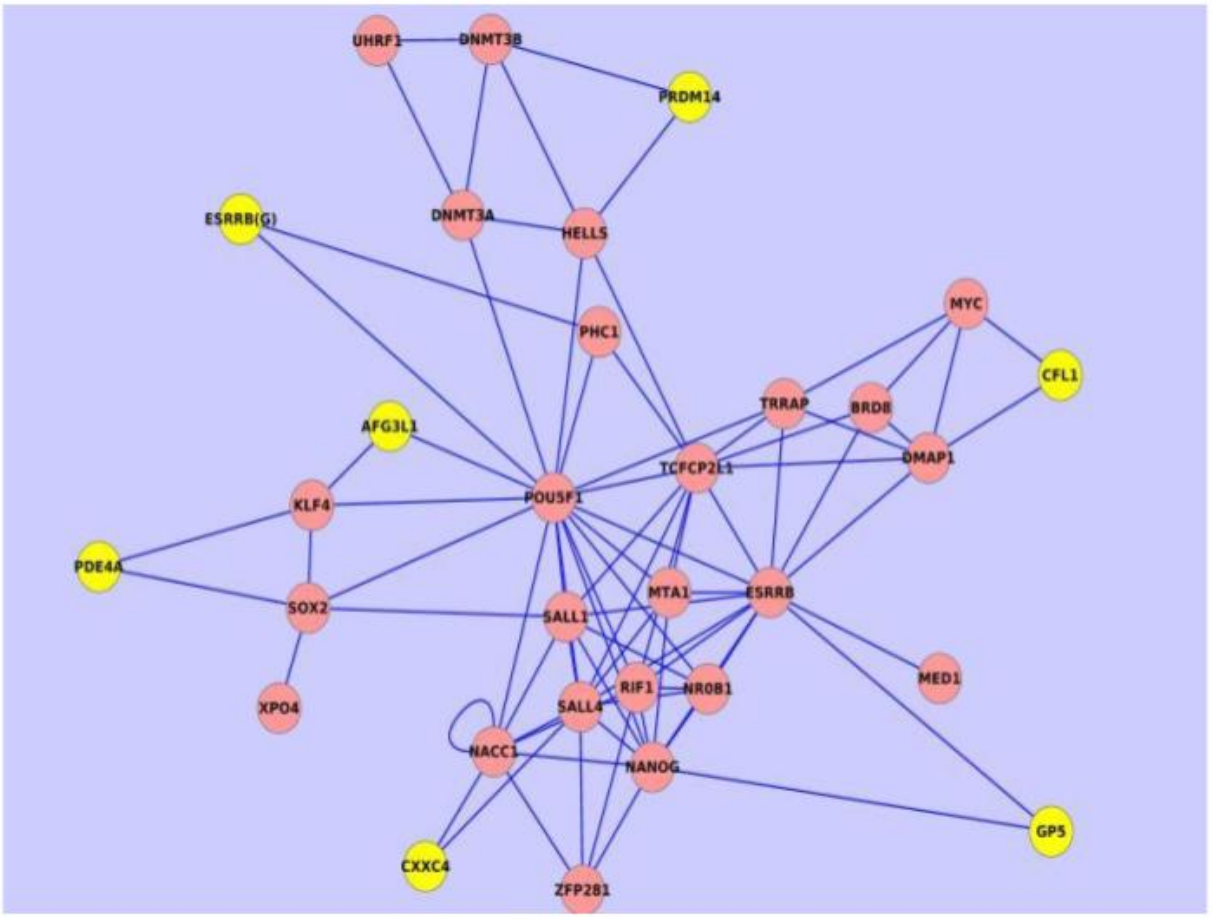


Figure 3.3: Subnetwork View

3.5 Results

3.5.1 Gene: ESRRB

POU5F1 and PHC1 are the transcription factors which are directly interacting with target gene and as we can see from Figur(3.4), perturbation in those nodes shows significant positive transcription activity. In perturbation in layer two in the node TCFCP2L1 show significant positive transcription activity. Whereas MTA1 perturbation is showing down regulation of gene but with relatively slower rate. In layer 4 MYC is showing positive transcription activity but ACTL6A is getting down regulating gene at much more faster rate.

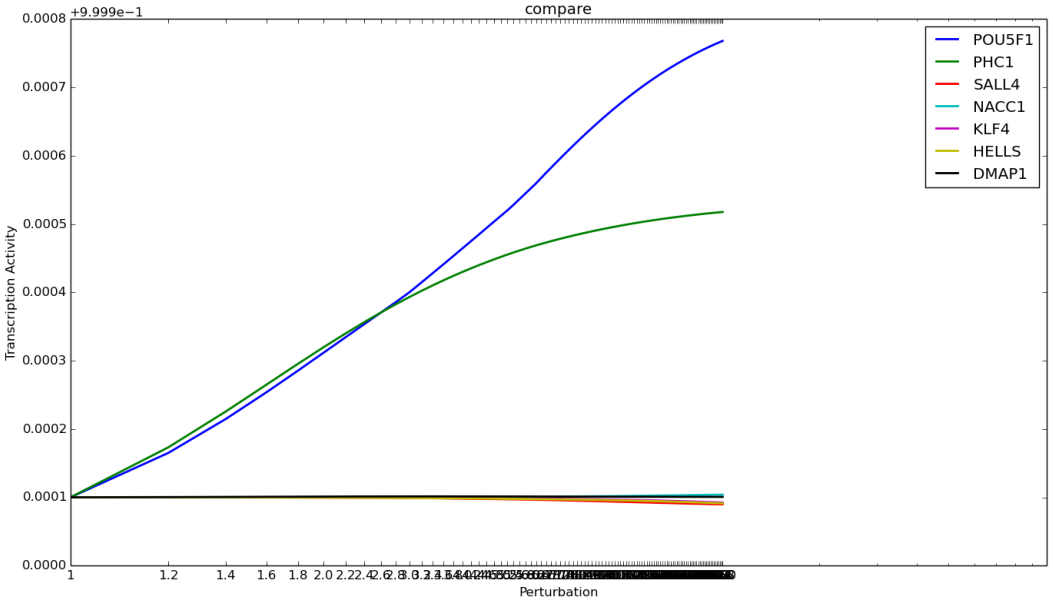


Figure 3.4: layer 1 perturbation

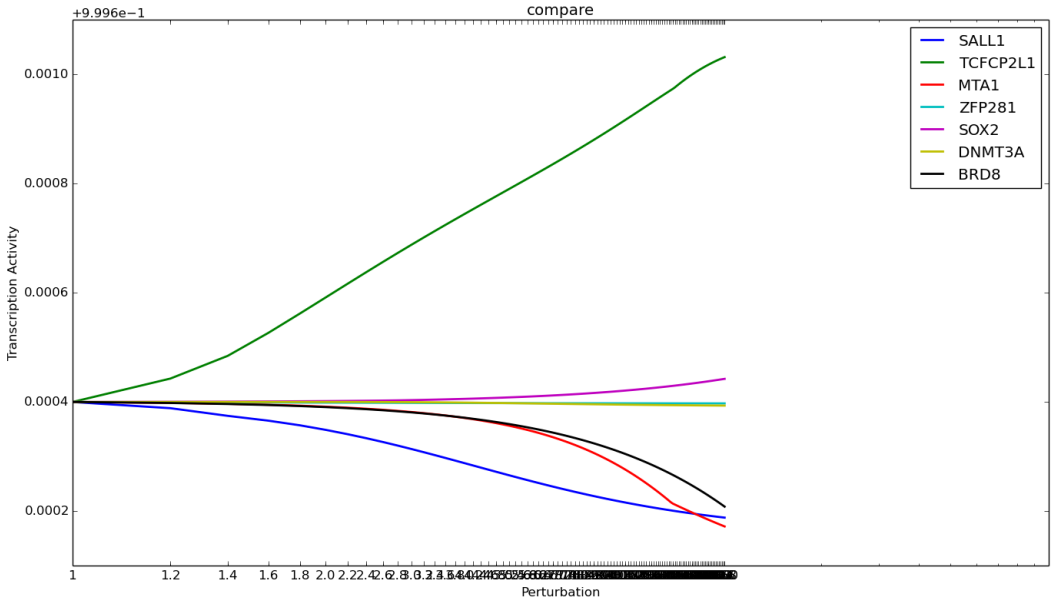


Figure 3.5: layer 2 perturbation

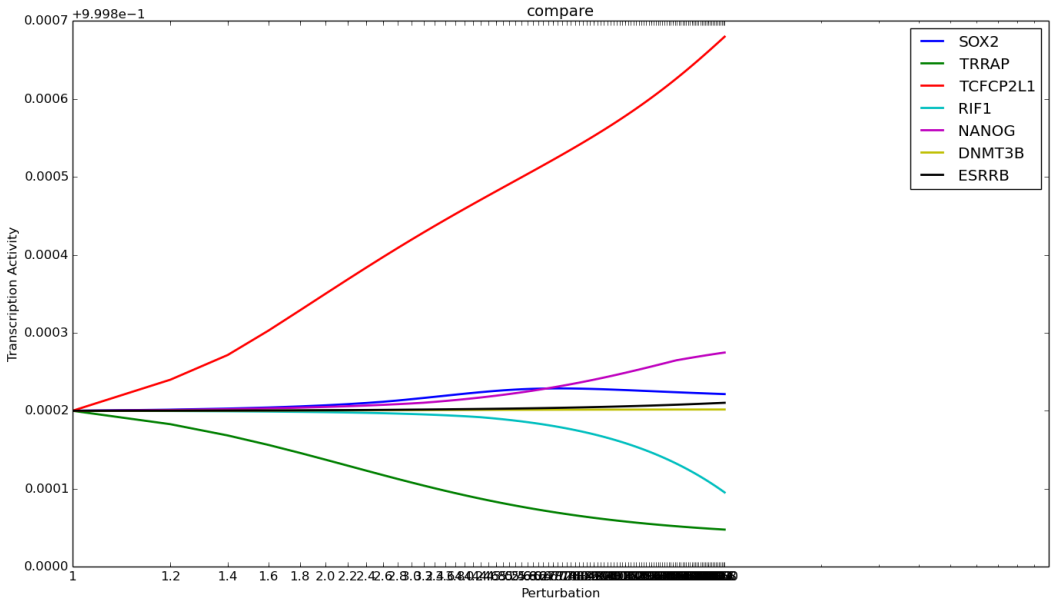


Figure 3.6: layer 3 perturbation

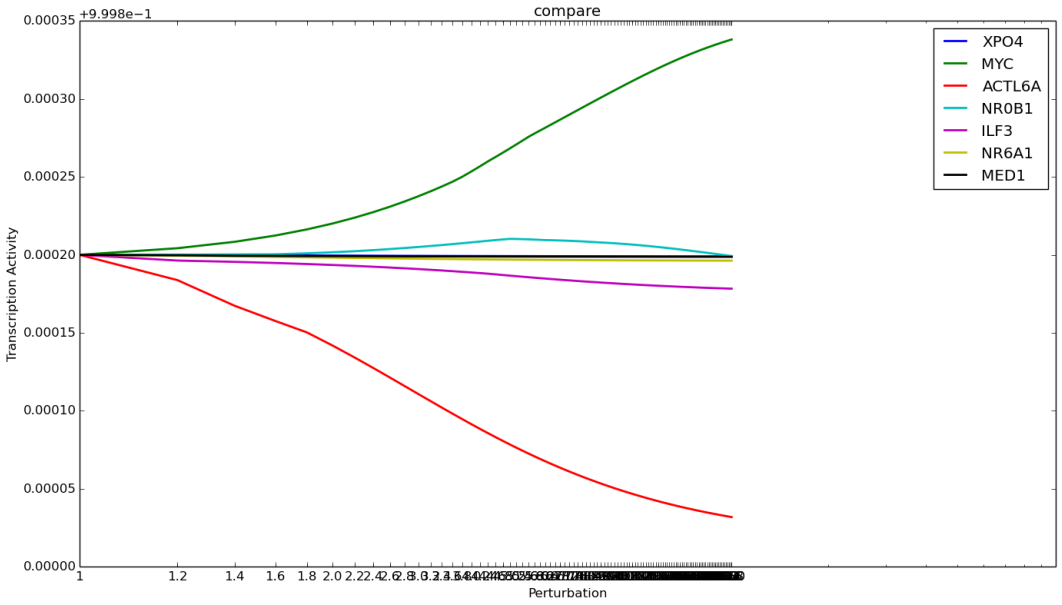


Figure 3.7: layer 4 perturbation

3.5.2 Gene: GP5

For gene GP5, when perturbation is made in layer 1, only transcription factors which are directly interacting with gene ESRRB and NANOG showed significant positive transcription activity. In layer 2, MTA1 is showing positive transcription activity but with very slower rate whereas SALL1 is deregulating with significantly higher rate. In layer 3 and 4, TCFCP2L1 and MYC is showing negative transcription activity significantly.

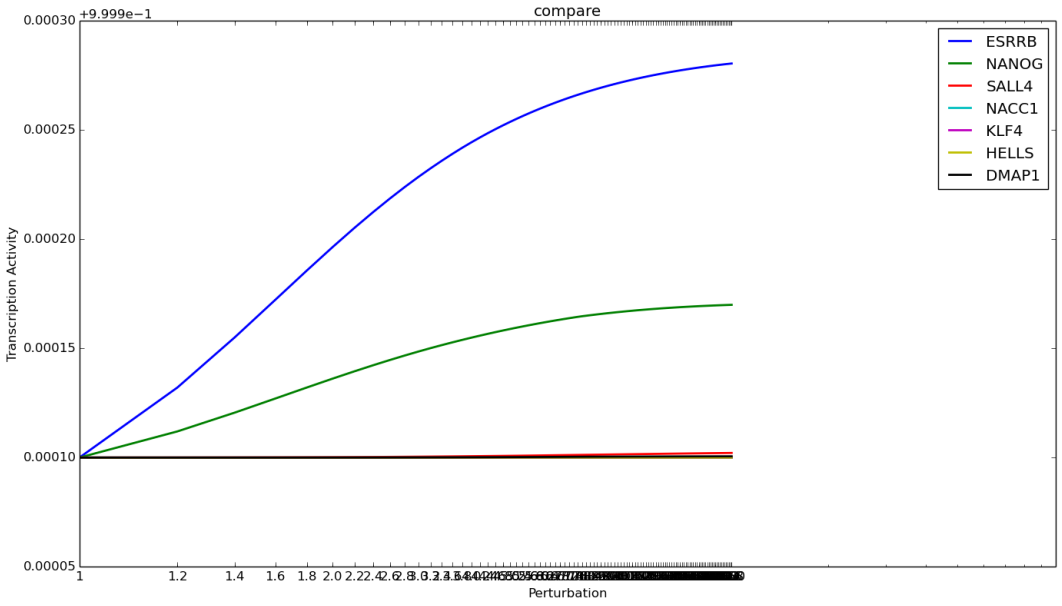


Figure 3.8: layer 1 perturbation

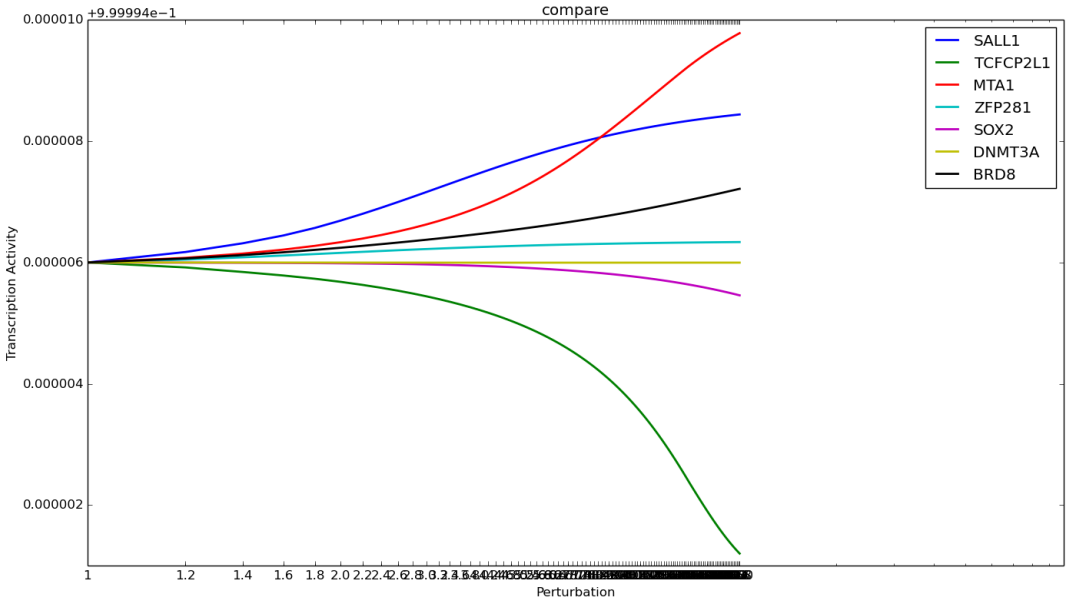


Figure 3.9: layer 2 perturbation

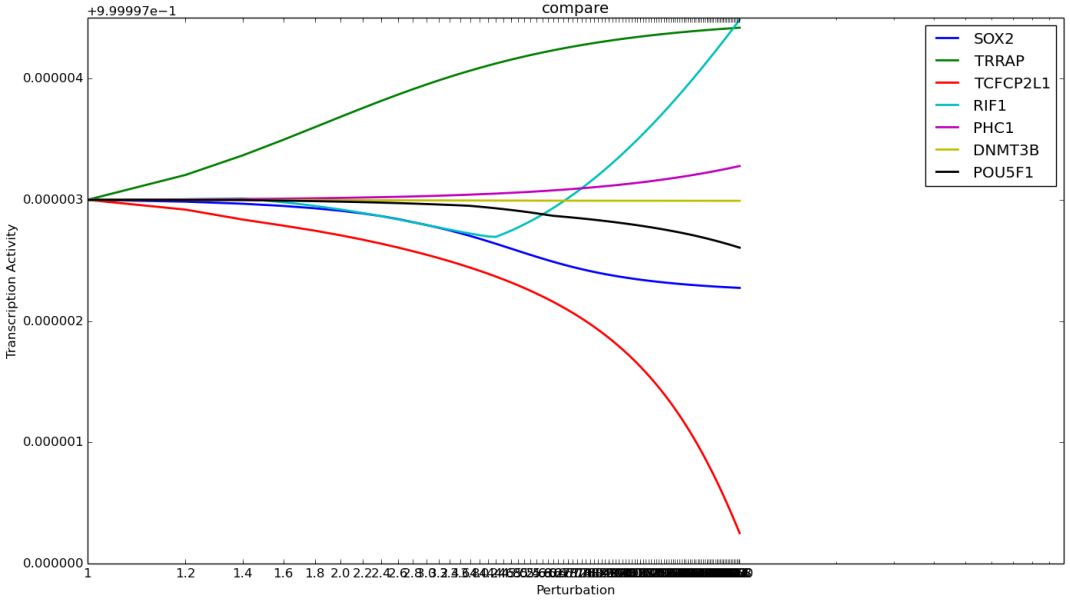


Figure 3.10: layer 3 perturbation

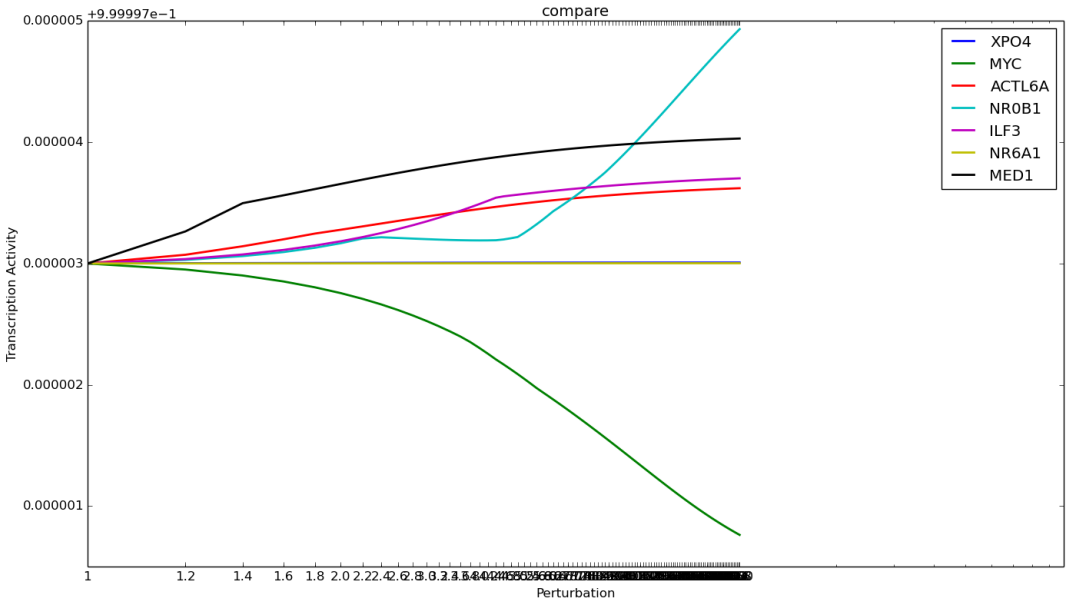


Figure 3.11: layer 4 perturbation

3.5.3 Gene: PDE4A

For gene PDA4A apart from the transcription factors KLF4 and SOX2 which are directly interacting with the gene, KLF4 is also showing positive transcription activity with relatively lower rate. In layer 2 TCFCP2L1 is showing positive transcription activity and SALL1 and MTA1 down regulation. Perturbation layer 4 shows significant down regulation by XPO4 and ACTL6A.

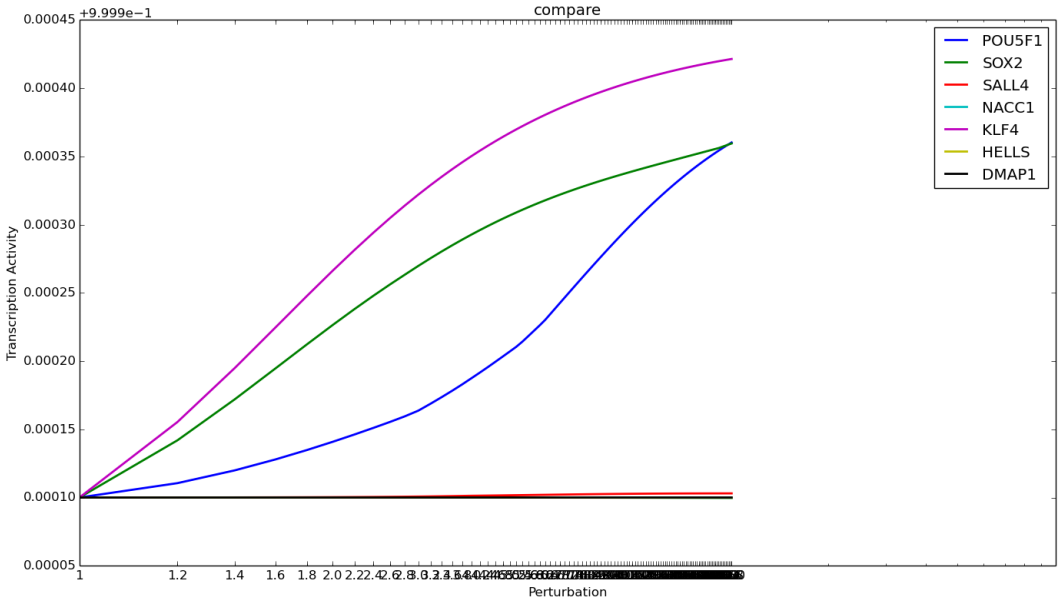


Figure 3.12: layer 1 perturbation

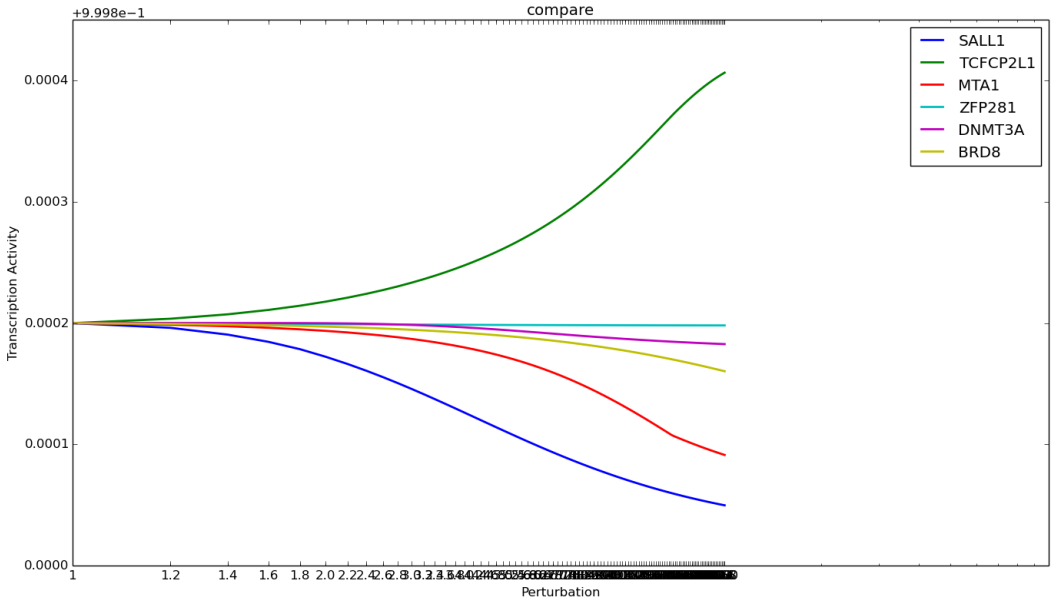


Figure 3.13: layer 2 perturbation

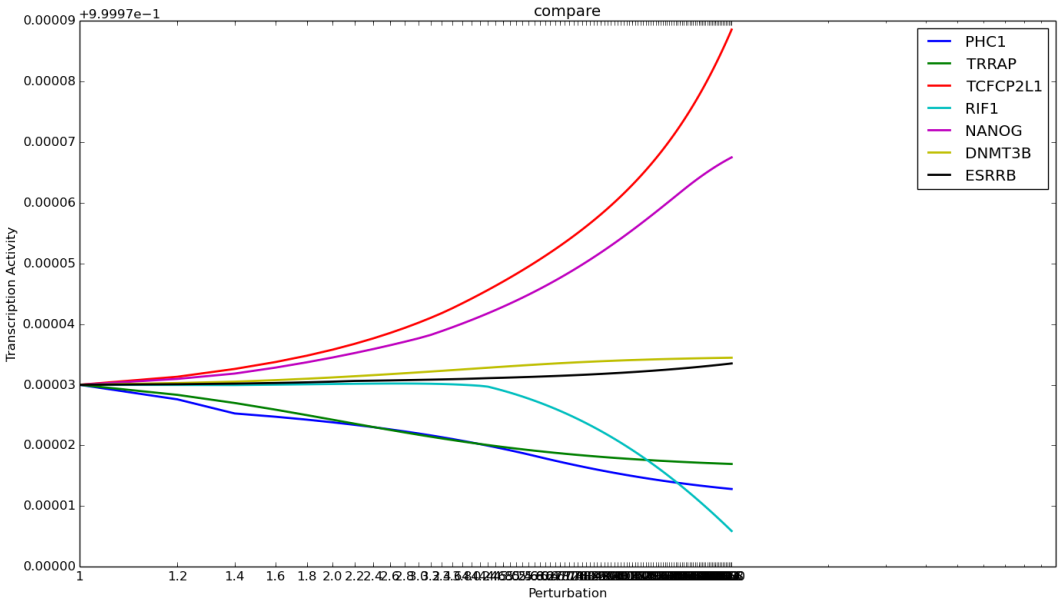


Figure 3.14: layer 3 perturbation

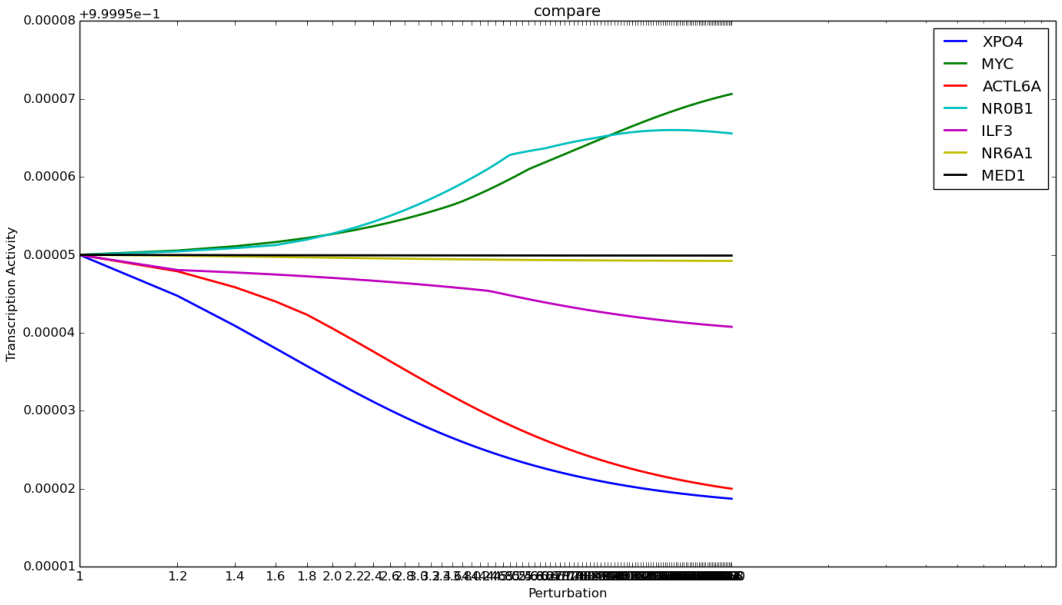


Figure 3.15: layer 4 perturbation

3.5.4 Gene: CXXC4

Here in layer 1, SALL4 and NACC1 are showing significant transcription activity whereas perturbation in other nodes in layer 1 is redundant. In layer 2 TCFCP2L1 is showing significant Down regulation but with slower rate. In layer 4, there is positive transcription activity due to ACTL6A.

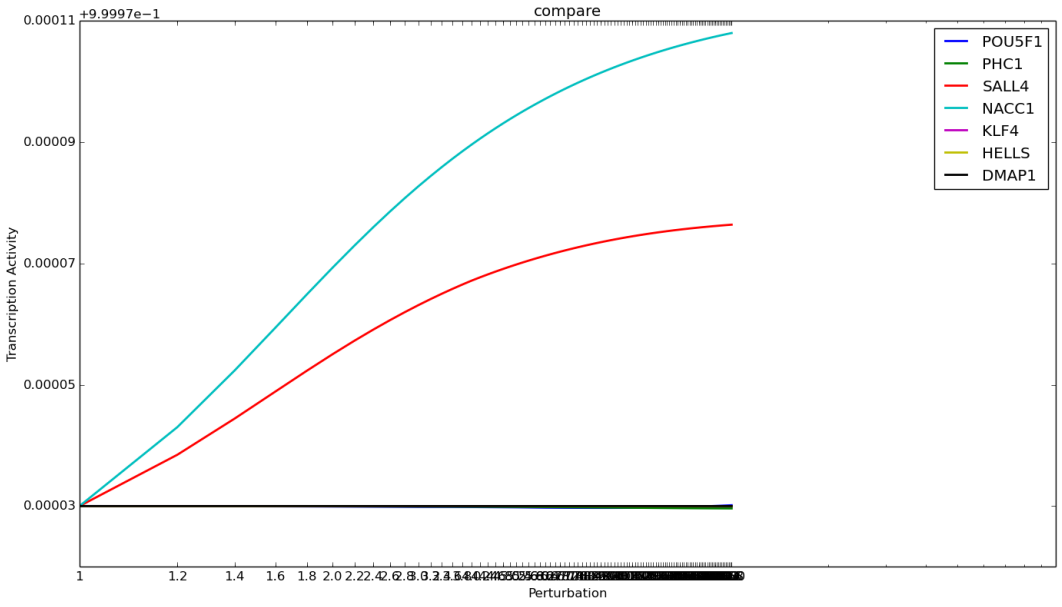


Figure 3.16: layer 1 perturbation

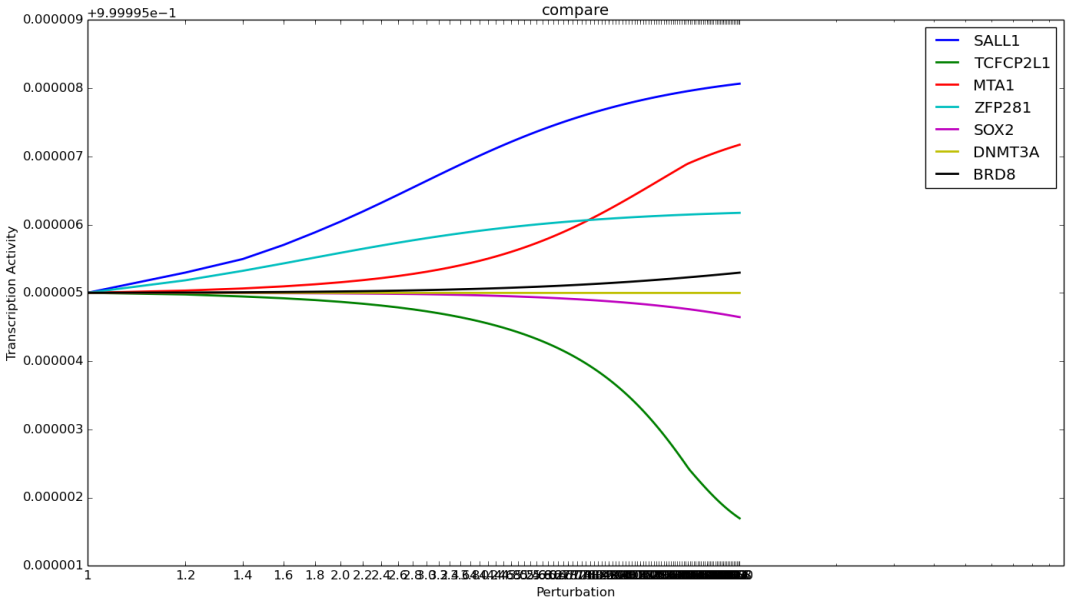


Figure 3.17: layer 2 perturbation

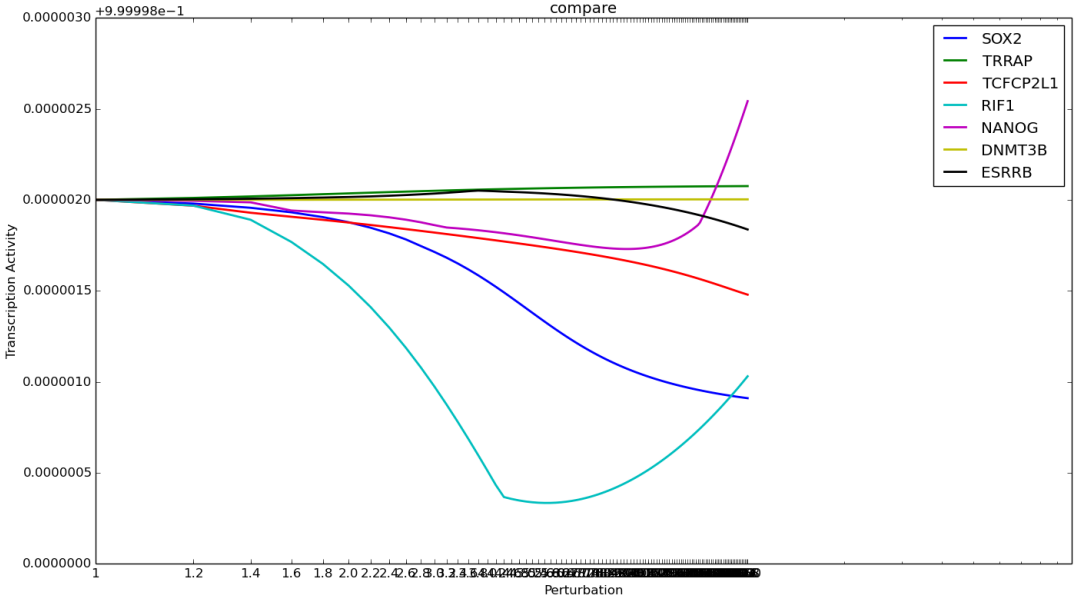


Figure 3.18: layer 3 perturbation

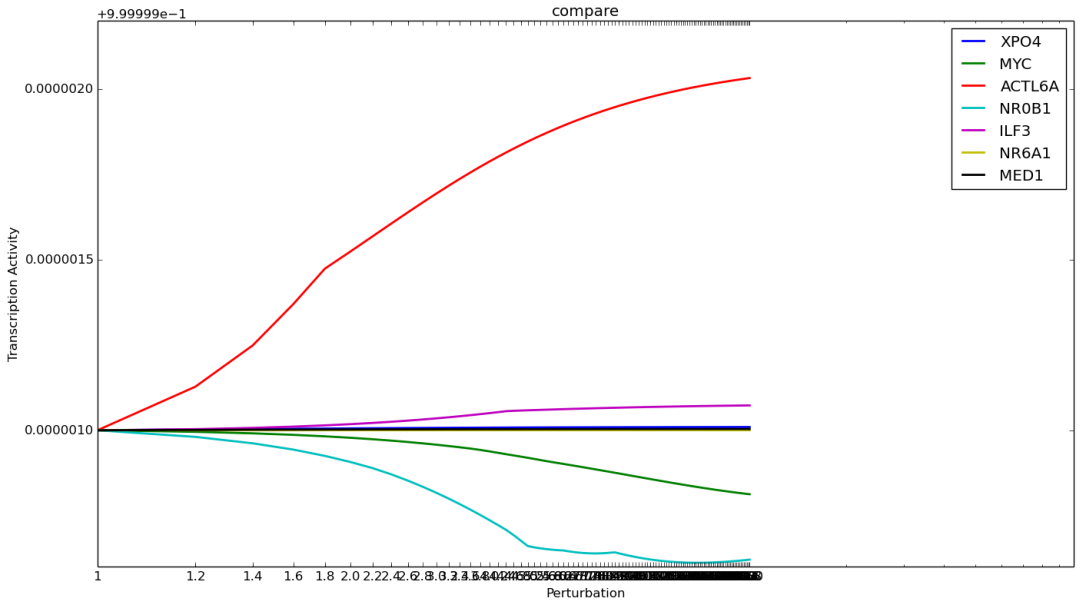


Figure 3.19: layer 4 perturbation

3.5.5 Gene: AFG3L1

For gene AFG3L1, in layer 1, the transcription factors which are binding directly to gene show positive transcription activity when perturbed. In layer 2, there is both positive and negative regulation by SOX2 and SALL1 respectively. In layer 4, there significant down regulation by ACTL6A and XPO4.

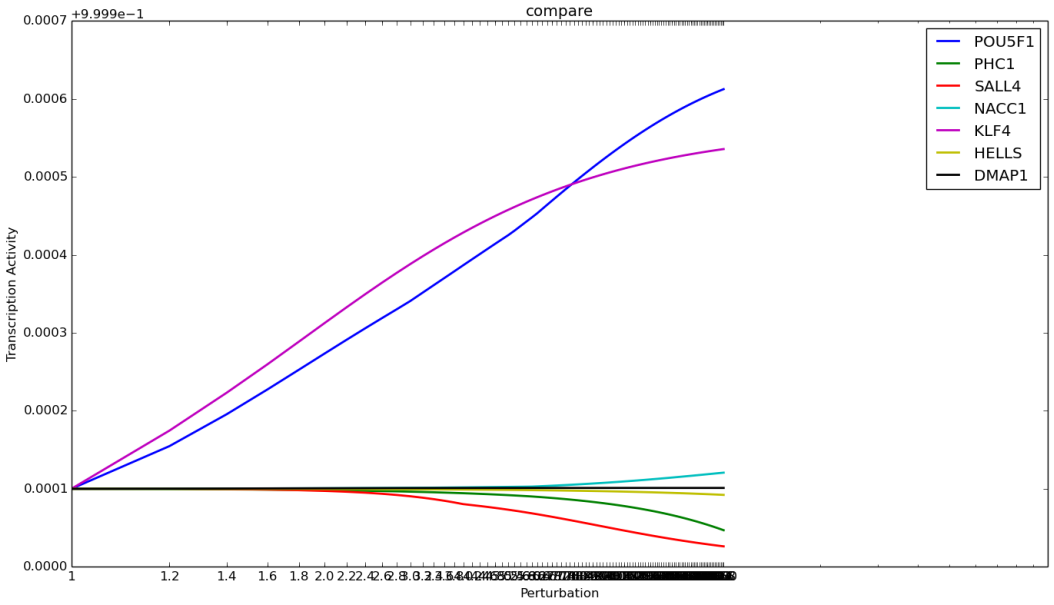


Figure 3.20: layer 1 perturbation

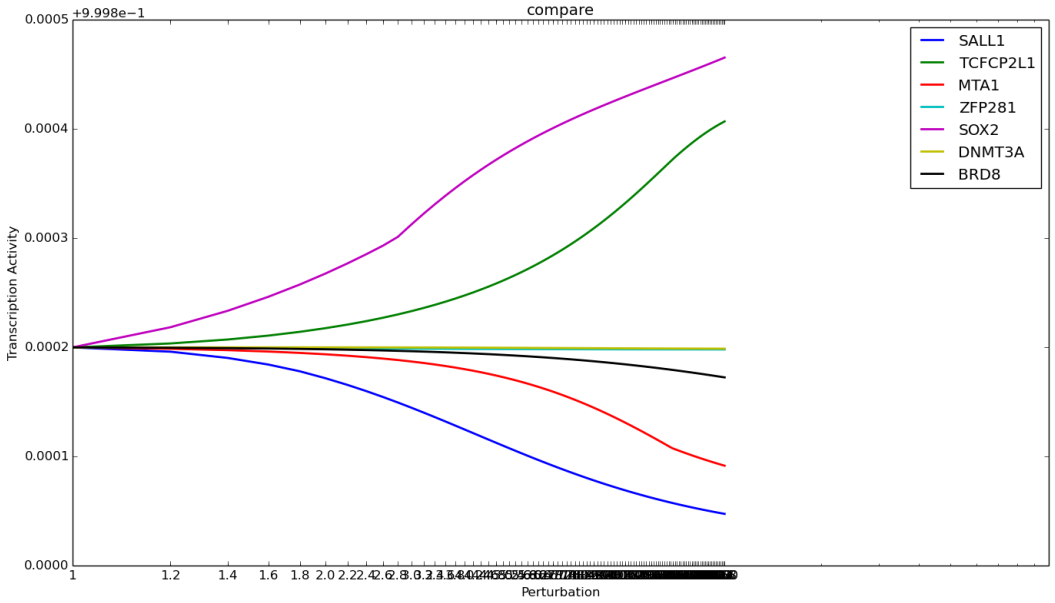


Figure 3.21: layer 2 perturbation

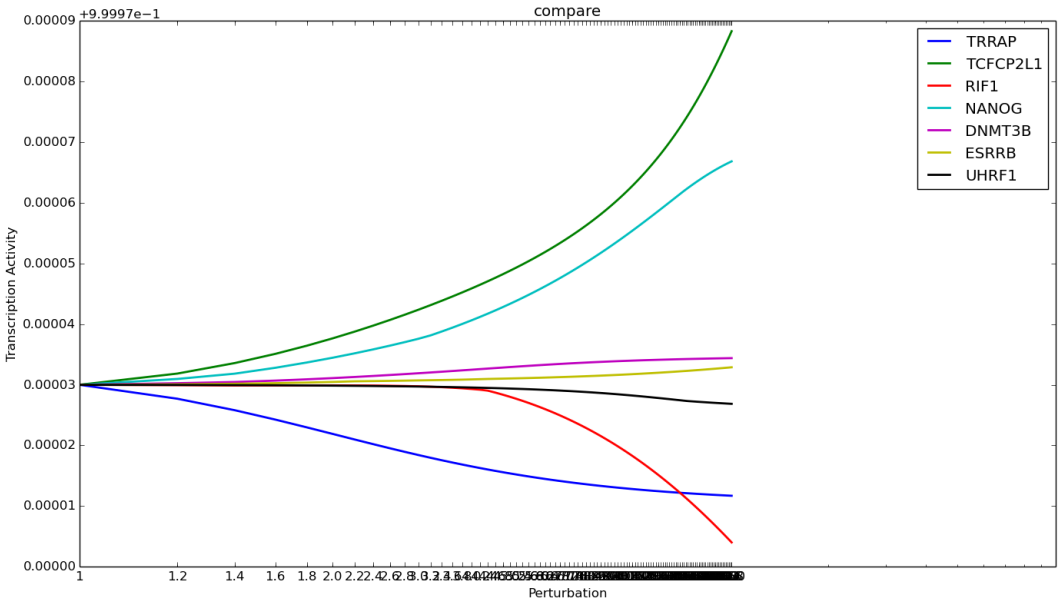


Figure 3.22: layer 3 perturbation

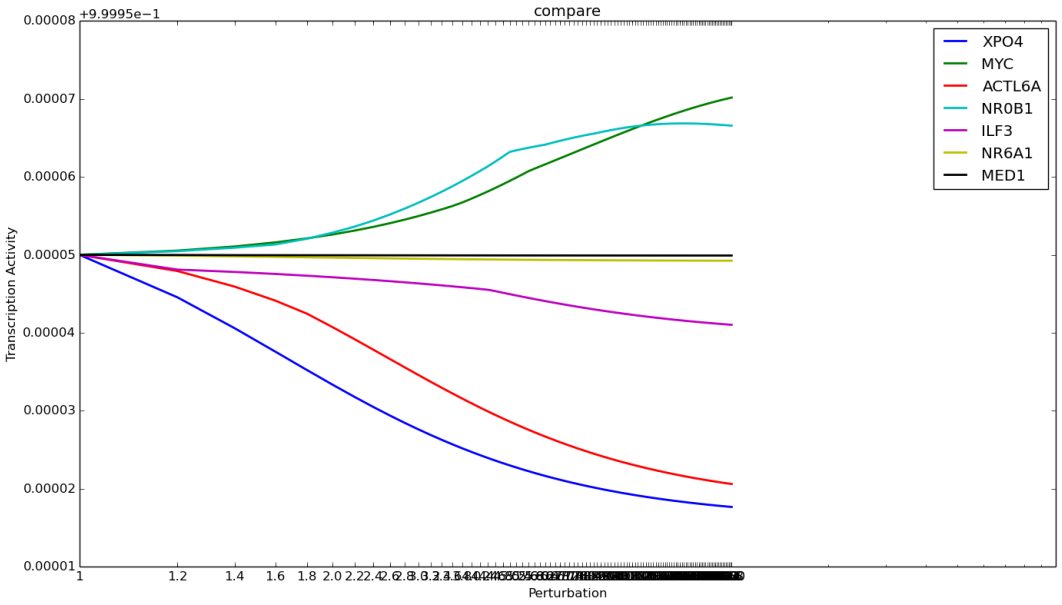


Figure 3.23: layer 4 perturbation

3.5.6 Gene: PRDM14

For gene,PRDM binding factors DNMT3B and HELLS are significant transcription activity. In layer 2, MTA1 shows down regulation at a slow rate. Perturbation in NR6A1 and ACTL6A show rapid down regulation of gene.

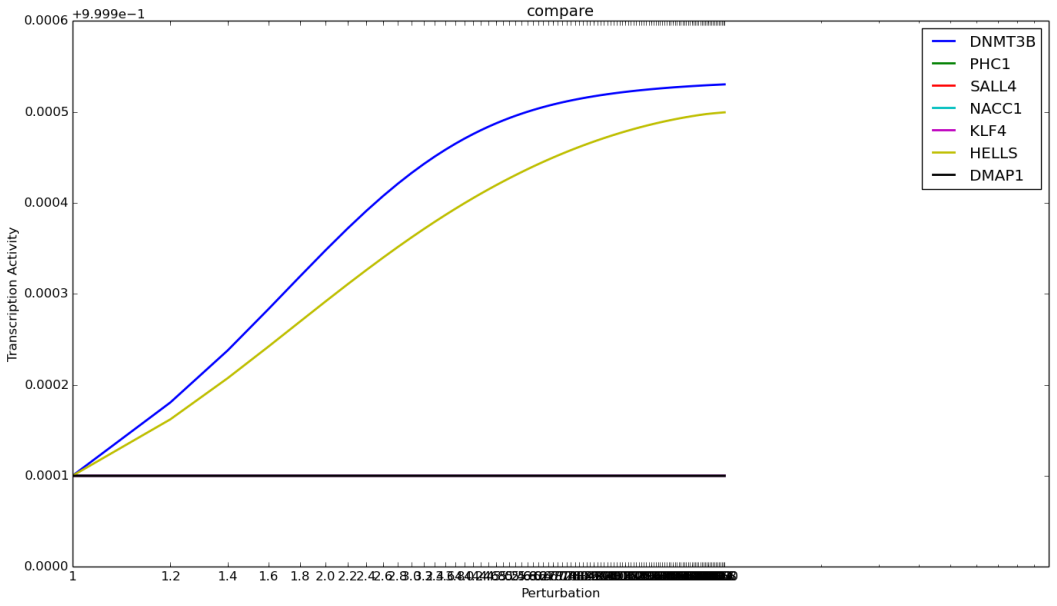


Figure 3.24: layer 1 perturbation

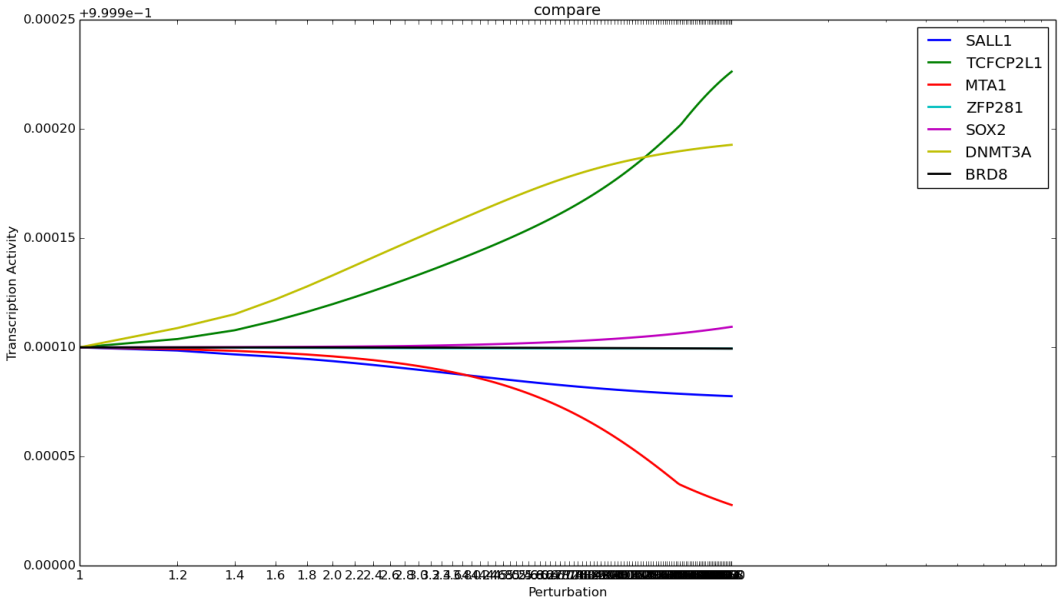


Figure 3.25: layer 2 perturbation

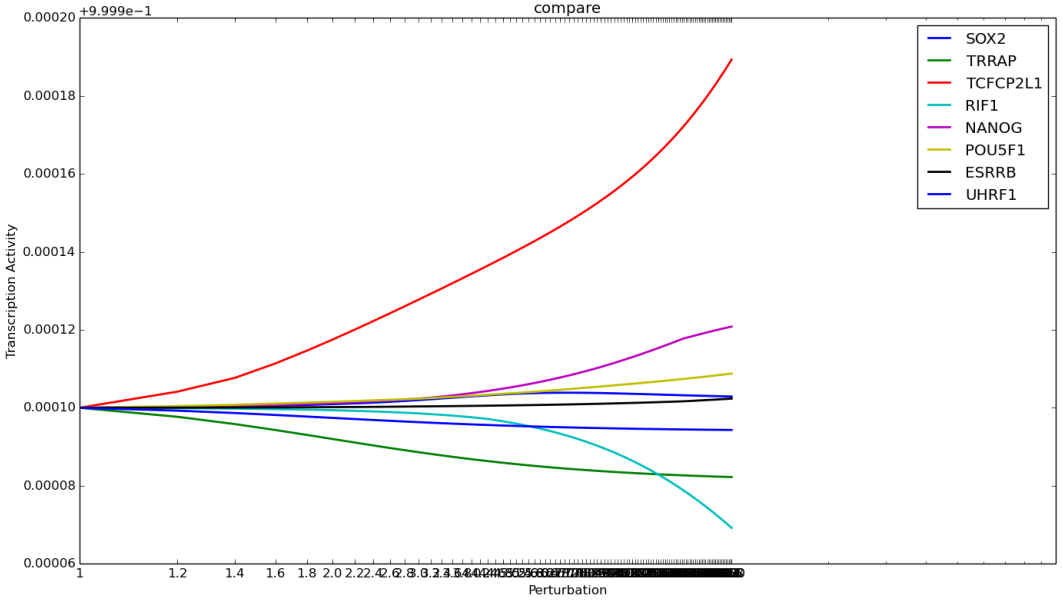


Figure 3.26: layer 3 perturbation

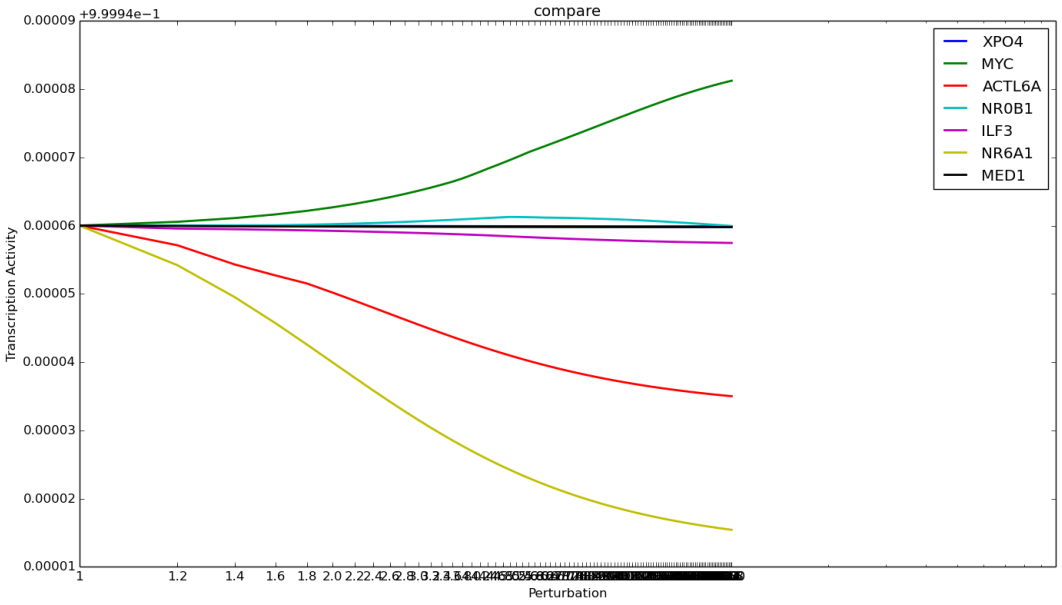


Figure 3.27: layer 4 perturbation

3.5.7 Gene: CFL1

For gene CFL1 there is significant transcription activity only one of the binding factor's perturbation: DNMT3B. In layer 2, TCFCP2L1 is showing significant down regulation of gene. Perturbations in layer 3 are redundant. In layer4, ACTL6A is showing very positive transcription activity.

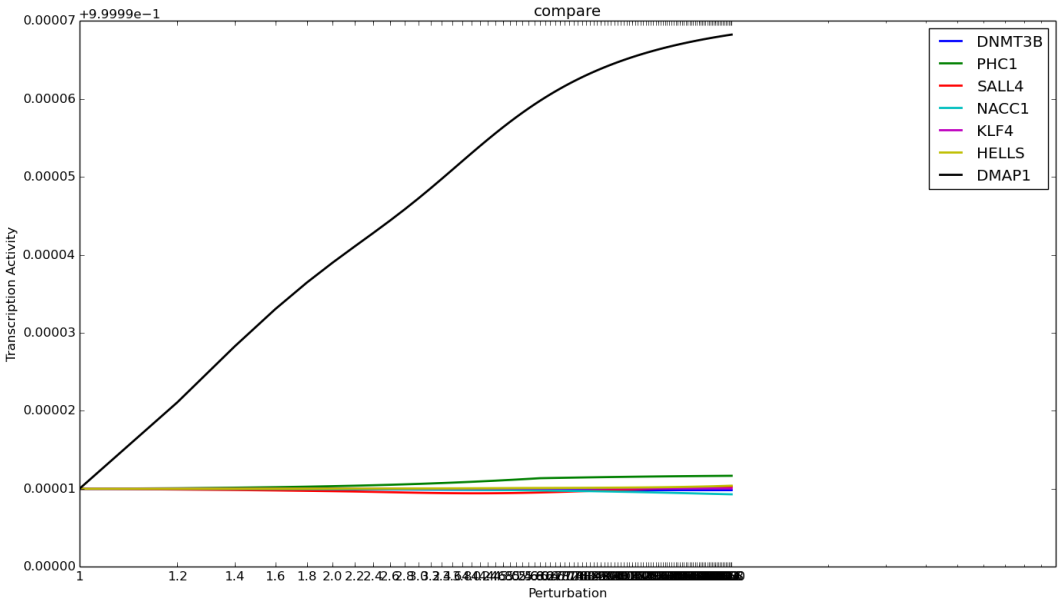


Figure 3.28: layer 1 perturbation

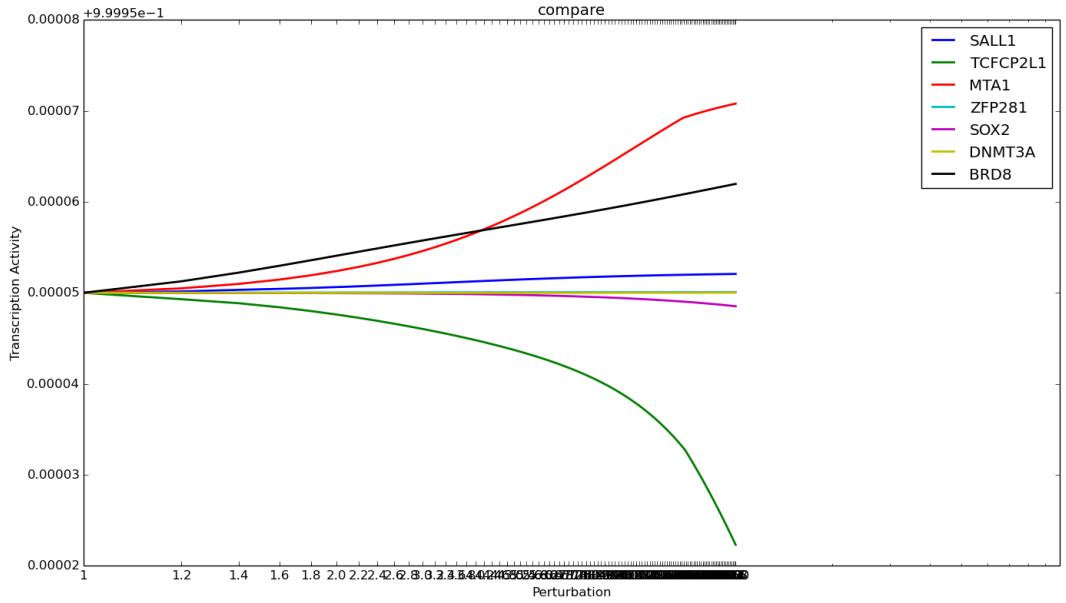


Figure 3.29: layer 2 perturbation

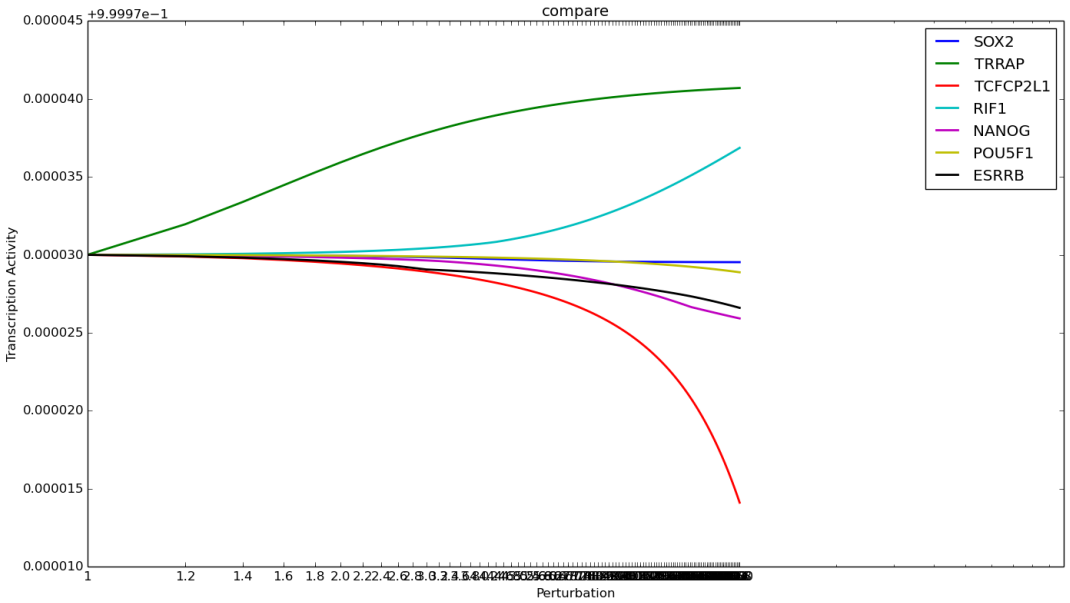


Figure 3.30: layer 3 perturbation

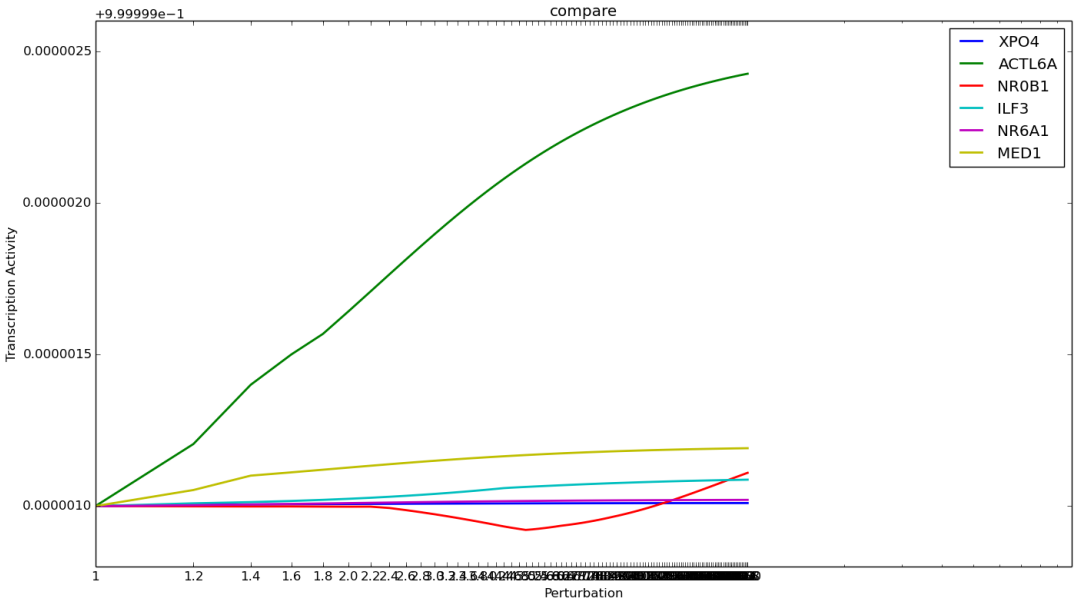


Figure 3.31: layer 4 perturbation

3.6 Conclusion

From this model we can see the perturbation in layer even at a layer 4 can have significant effect on Transcription activity of the gene. But the more important question is will this model help us understand further the unexplained cellular/disease phenotypes, which do not correlate with corresponding genotypes?

So far I have worked on a very small subset of network. But if we can extend this model to a the whole network then it would be a great step forward toward solving this problem. To make this model more fitting to real world gene regulatory network we also need to consider interactions among gene regulatory network and try to reduce back propagation.

Bibliography

[Alo06] Uri Alon, *An introduction to system biology: Design principles of biological circuits*, Chapman Hall/CRC, 2006.

[BBC+14] S. Boccaletti, G. Bianconi, R. Criado, C.I. del Genio, J. Gmez-Gardees, M. Romance, I. Sendia-Nadal, Z. Wang, and M. Zanin, *The structure and dynamics of multilayer networks*, Physics Reports **544** (2014), no. 1, 1 – 122, The structure and dynamics of multilayer networks.

[MI07] Sergei Maslov and I. Ispolatov, *Propagation of large concentration changes in reversible protein-binding networks*, Proceedings of the National Academy of Sciences **104** (2007), no. 34, 13655–13660.

[New09] Mark Newman, *Networks: An introduction*, OUP Oxford, May 2009.

[SC13] Kristf Z. Szalay and Peter Csermely, *Perturbation centrality and turbine: A novel centrality measure obtained using a versatile network dynamics tool*, PLOS ONE **8** (2013), no. 10, 1–11.

[Sne06] Kim Sneppen, *Models of life: Dynamics and regulation in biological systems*, Chapman Hall/CRC, 2006.

[XAS+14] Huilei Xu, Yen-Sin Ang, Ana Sevilla, Ihor R. Lemischka, and Avi Ma'ayan, *Construction and validation of a regulatory network for pluripotency and self-renewal of mouse embryonic stem cells*, PLOS Computational Biology **10** (2014), no. 8, 1–14.

[XBD+13] Huilei Xu, Caroline Baroukh, Ruth Dannenfelser, Edward Y. Chen, Christopher M. Tan, Yan Kou, Yujin E. Kim, Ihor R. Lemischka, and Avi Ma'ayan, *Escape: database for integrating high-content published data collected from human and mouse embryonic stem cells*, Database **2013** (2013), bat045.