Project report
Network inference using structural priors
Systems Biology & Machine Learning

Abhishek Pathak
Entry no. 2015CS10424

1 Introduction

Gene regulatory networks, or transcription networks, are a way of modeling the interactions that occur between genes in the cell. A directed edge from gene A to gene B indicates that gene A controls the rate of production/expression of gene B by binding to its regulatory regions. In a very simplified view of such causal relationships, correlated gene expression can imply that there is a regulatory interaction between the two. Such networks are often inferred via wet experiments, with knocking out genes and observing changes in expression. However, computationally predicting networks from expression data of genes in time series form can be cheaper and less time-consuming, and help guide the direction of further wet experiments.

A problem exists of the problem at hand being an inverse problem, which is underconstrained, and allows multiple solutions for the same set of data i.e. there is degeneracy. This arises by assuming that every possible network over the genes (nodes) is equally likely, whereas this is not the case - biological networks are observed to have certain properties with regard to degree distributions, occurrence of motifs and modular structures. Giving networks with such properties a higher prior probability can help constrain the search space to infer biologically feasible networks.

2 Review of computational biology

In the beginning of the project, a wide review of existing literature was carried out in systems biology, and computational biology in general.

- A related problem was studied - that of using machine learning to infer the functionality and importance of genes in a known interaction network [2].
- A review of deep learning applications in computational biology describes the use of deep learning and convolutional neural networks in predicting mutation effects, multiple traits in the field of regulatory genomics [3]. However, this approach has not yielded much success in systems biology/network inference, as deep learning requires lots and lots of data to learn rich, complicated models, whereas lack of training data is one reason why network inference is hard.
- Work in [9] describes a way of validating computationally inferred networks by studying their structural controllability properties. This could potentially form the basis for future lines of investigation into pruning the space of inferred networks.

3 Analysis of existing methods

The SAprior method [1] was studied in detail. This method uses simulated annealing to infer the network parameters - specifically, the in-degree distribution parameters, which is assumed to be scale-free. At each step of the simulated annealing, networks are sampled using current and proposed parameters, and the new network is accepted or rejected based on the simulated annealing condition.

The expression data was synthetic in nature, generated using GeneNetWeaver software from an underlying network constructed with certain degree constraints. The underlying gold networks were 100-gene (100-node) networks. The advantage of using synthetic data is that the underlying ground truth network is known, and the data is generated according to the dynamics as observed in real networks.

- DREAM3
- DREAM4
- Exponential In-degree Power-law out-degree (EIPO)
- EIPO with modules
- Power-law In-degree Power-law Out-degree (PIPO)

The DREAM3 and DREAM4 networks were based on networks in the third and fourth run of the DREAM network inference challenges. The SAprior results were replicated for this data, and found to match with the results obtained earlier. Some of the plots for PR curve, ROC curve and barplots for the score metric (as defined in the DREAM challenge) [6] are shown below.

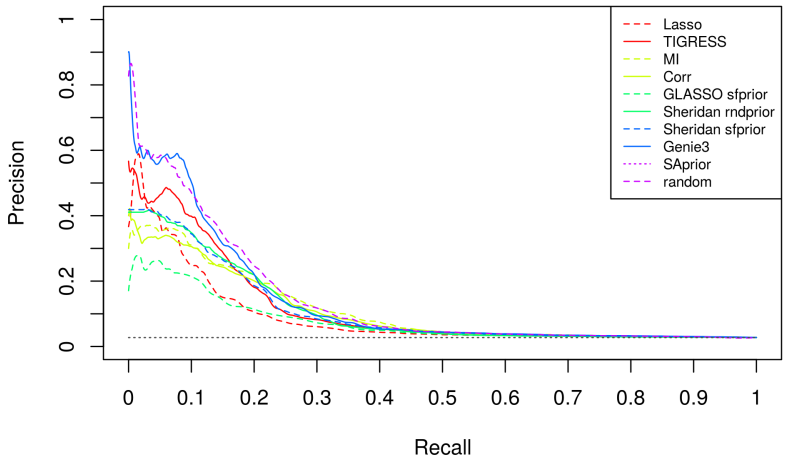

Figure 1: PR curve for various methods on DREAM3 data

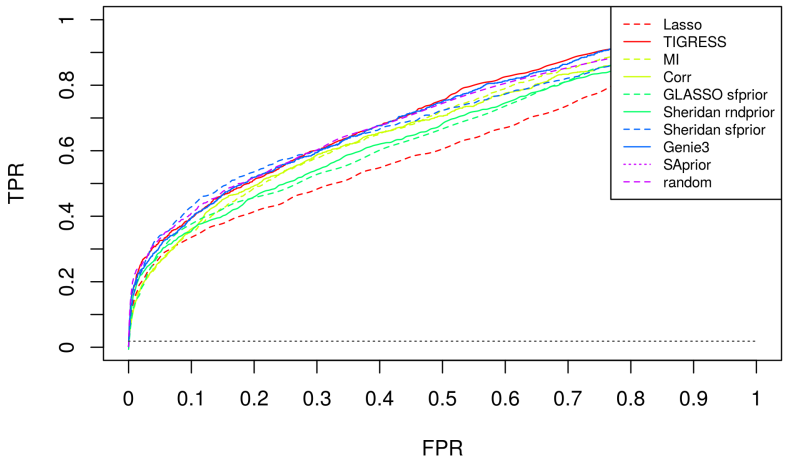

Figure 2: ROC curve for various methods on EIPO data

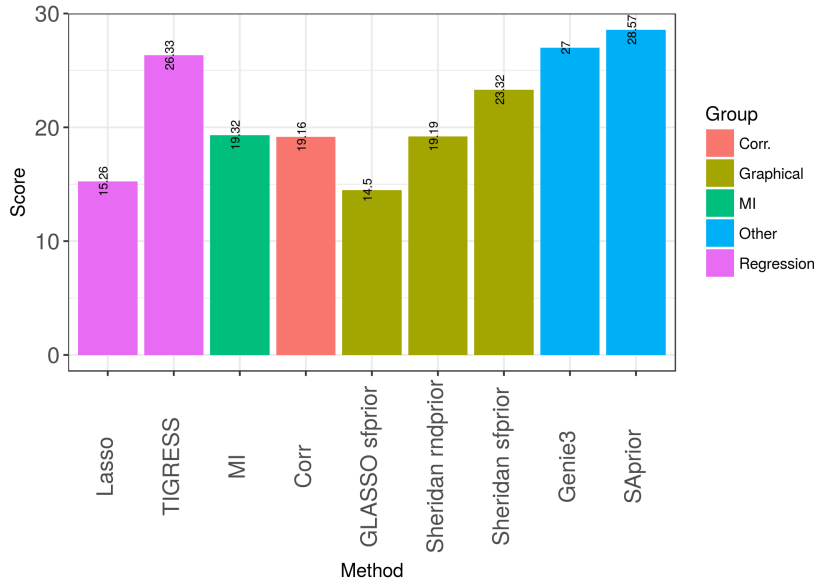

Figure 3: Barplots for score metric averaged over all types of data

In particular, the ordering of the SAprior method with respect to the other methods in terms of AUPR, AUROC and score metric was found to match with the earlier obtained results.

3.1 Techniques to analyze inferred networks

The techniques were based on error metrics, keeping degree distribution of the network into account.

- The links in the inferred network were categorized based on the out-degree of the node from which they emanate, and the in-degree of the node at which they terminate. On each of these categories, the FPR, TPR and FNR were recorded. (2D)
- The links were categorized solely on the basis of the out-degree of their source node. (1D)

- The links were categorized solely on the basis of the in-degree of their source node. (1D)

We had earlier categorized nodes on the basis of their in and out-degrees (taken together), but this came with the pitfall the for each node, incoming and outgoing links had to be handled separately, which was not meaningful, given that we are really interested in the links present/inferred. Also, for the 1D cases, the categorization can be done either on the basis of nodes or links, as it comes out to be the same thing.

For the 2D case, heatmaps of FPR, TPR and FNR were generated as follows. Blue regions indicate the categories for which their were no links.

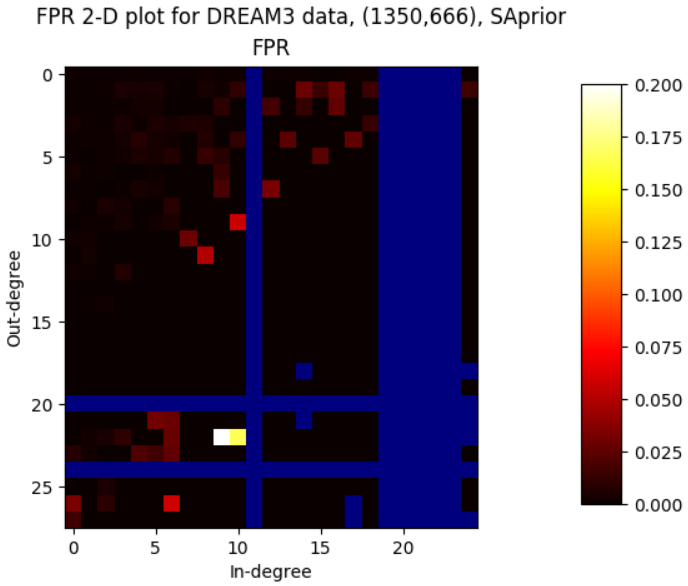

Figure 4: 2D FPR heatmap

In the caption, 1350 is the number of links in the ground truth network, and 666 is the number of links in the inferred networks.

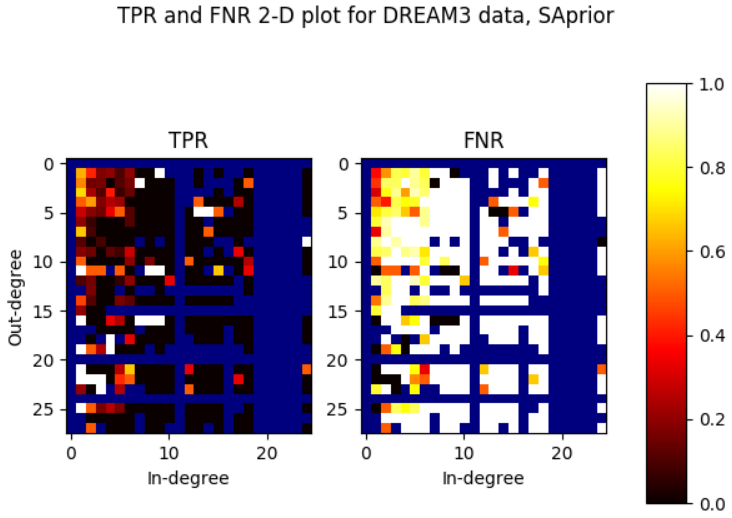

Figure 5: 2D TPR and FNR heatmaps

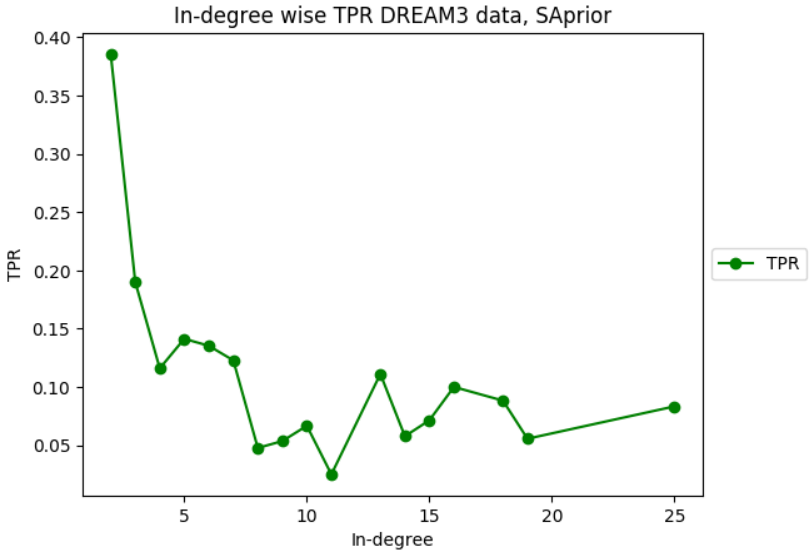

Figure 6: 1D TPR plot, in-degree wise

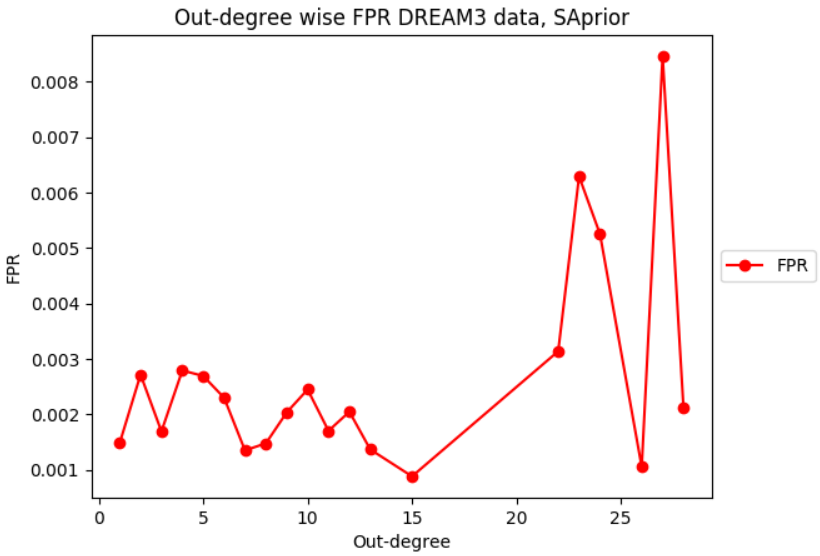

Figure 7: 1D FPR plot, out-degree wise

3.2 Comparison with other inference methods

We would also like to apply these analysis techniques on the results afforded by other inference methods, and compare the analysis of the results obtained from both SAprior and the other methods. In addition to applying the analysis techniques to the results of other methods, we also developed a variant of the analysis techniques to directly compare SAprior and the other method.

- For the 1-D case, we plotted the error metrics on the two methods in an overlaid manner, and observed where the two methods were making similar kinds of 'guesses' and 'mistakes'.
- For the 2-D case, we took a difference of the two heatmaps (excluding the categories which had no links in the data).

We compared many methods against SAprior, and studied in detail the comparison between SAprior and GENIE3 [4]. Some of the comparative plots are outlined below.

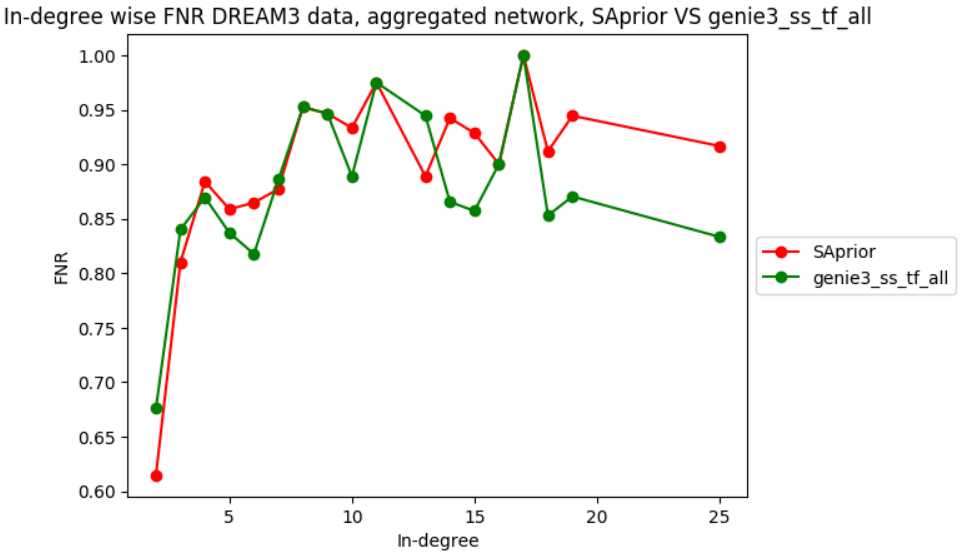

Figure 8: Comparing on DREAM3 data, in-degree FNR

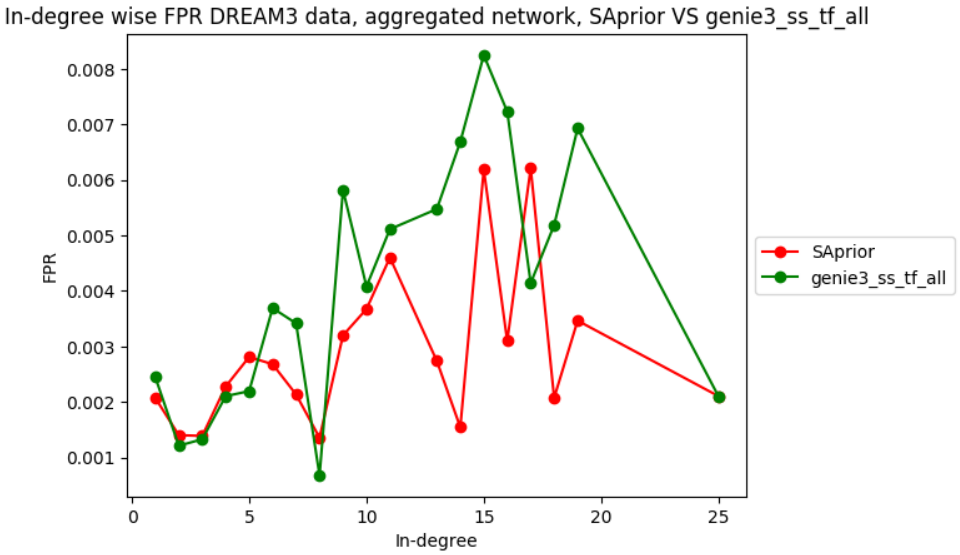

Figure 9: Comparing on DREAM3 data, in-degree FPR

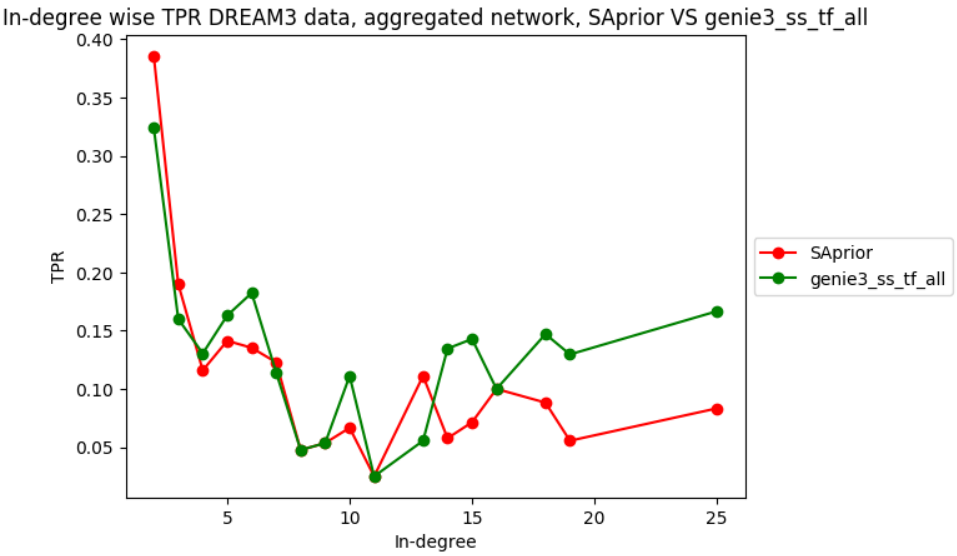

Figure 10: Comparing on DREAM3 data, in-degree TPR

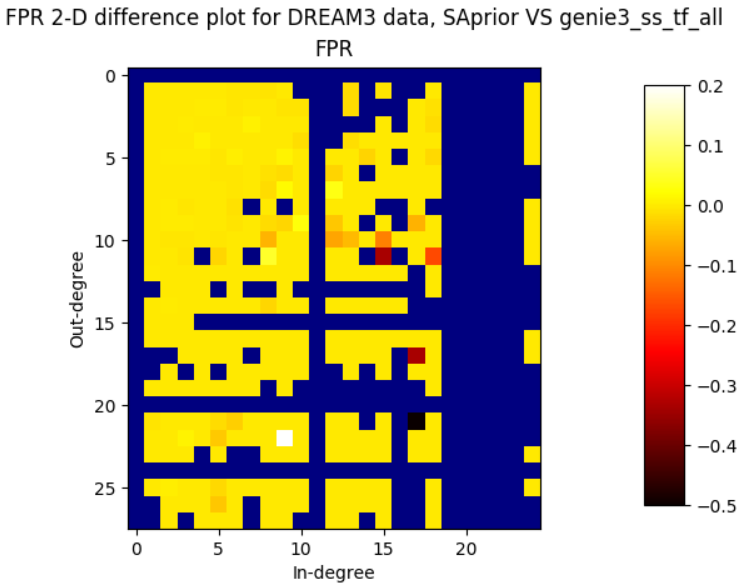

Figure 11: Comparing on DREAM3 data, FPR difference plot

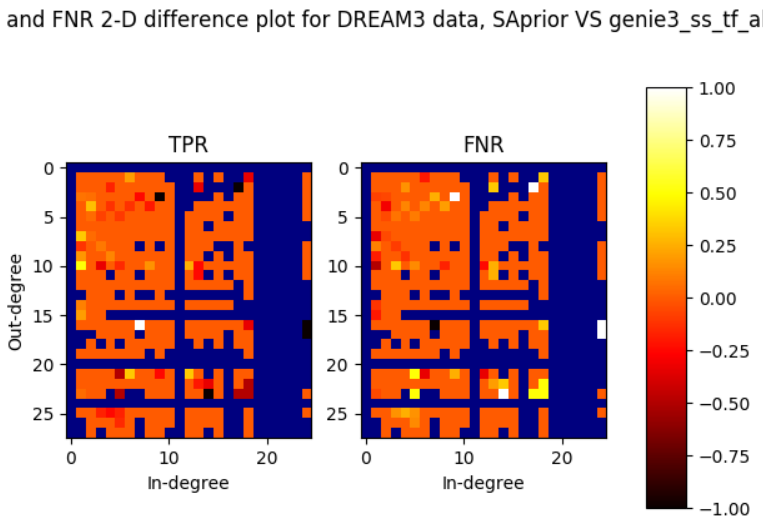

Figure 12: Comparing on DREAM3 data, TPR and FNR difference plot

4 Results, comparisons and future work

There were observed patterns in the plots when we analyzed the SAprior inferred networks.

- False positives were relatively sparse in the 2-D heatmap with low values of FPR. There was a single strong peak observed in all the network types, where the in and outdegree was approximately half of max possible in-degree and out-degree.
- True positives were present in the left half of the heatmap, with the inference methods performing better for lower in-degree.
- False negatives were present in a large fraction of the heatmap, and were quite high for higher values of in-degree.
- This seems to suggest that SAprior can guess links with lower in-degree correctly, and can be combined with a method that guesses higher in-degrees well.
- FNR trends for in-degree (viewed in 1D) were similar across network types. This implies a consistent failure of the method to pick up many edges, viewed from the perspective of in-coming links to a node.
- Similarly, TPR trends for in-degree (viewed in 1D) were similar across network types.

When comparing the SAprior method to the GENIE3 method, following were the observations.

- In GENIE3 2-D FPR heatmap, the one spike in FPR value was seen in the lower and/or right portion of the heatmap, as opposed to the central part of the heatmap for SAprior.
- TPR was higher for lower in-degrees, but in this case the 'trend' was not so clear as for SAprior. It was difficult to characterize as the behaviour was not consistent across network types as was for SAprior.
- Regarding FPR difference heatmap, the difference between FPR values was quite close to zero for much of the heatmap. In all the three network types, SAprior had much higher FPR in one spot on the heatmap.

- For DREAM3 and EIPO modular data, the differences in TPR and FNR were also close to zero for much of the heatmap. However, for EIPO data, SAprior had consistently lower TPR than GENIE3, while having similar FNR.

The ideas derived from our work for future directions of investigation are as follows.

- Currently, SAprior only looks at the in-degrees of nodes to impose a prior. This can be extended by imposing a prior on both the in-degree and the out-degree distribution. Since out-degrees can't be included in the current framework of binary programming optimization for sampling of networks, we need to devise a new sampling approach.
- Specifically, we can look at an MCMC approach to sampling networks, after applying a prior on the in as well as out-degree distribution. The binary programming optimization can be replaced by an MCMC sampling. Similar work has been done in [8], which we can draw on to devise the new sampling.
- Observing that different algorithms predict different kinds of links more accurately, ensembling can be used to combine predictions of different models, each of which identifies different network patterns. This idea is also explored in [7] and is shown to be more effective than individual learners.
- Finally, we will also need to validate the obtained model on real data, to test its applicability to real biological datasets.

References

[1] Tarun Mahajan. *Augmenting gene network inference using meta-analysis and structural priors*. Master's thesis, Department of Electrical Engineering, IIT Delhi, 2017.

[2] Xue Zhang, Marcio Luis and AcencioNey Lemke. *Predicting Essential Genes and Proteins Based on Machine Learning and Network Topological Features: A Comprehensive Review*. Front. Physiol. 7:75. doi: 10.3389/fphys.2016.00075

[3] C. Angermueller et. al. *Deep learning for computational biology*. Mol Syst Biol. (2016) 12: 878

[4] Vn Anh Huynh-Thu, Alexandre Irrthum, Louis Wehenkel, Pierre Geurts. *Inferring Regulatory Networks from Expression Data Using Tree-Based Methods*. PLOS One. https://doi.org/10.1371/journal.pone.0012776

[5] Riet De Smet, Kathleen Marchal. *Advantages and limitations of current network inference methods*. Nature Reviews in Microbiology, doi:10.1038/nrmicro2419.

[6] Daniel Marbach , Robert J. Prill , Thomas Schaffter , Claudio Mattiussi , Dario Floreano and Gustavo Stolovitzky. *Revealing strengths and weaknesses of methods for gene network inference*. PNAS. www.pnas.org/cgi/doi/10.1073/pnas.0913357107

[7] Daniel Marbach, James C Costello, Robert Kffner, Nicole M Vega, Robert J Prill, Diogo M Camacho, Kyle R Allison , The DREAM5 Consortium , Manolis Kellis , James J Collins & Gustavo Stolovitzky. *Wisdom of crowds for robust gene network inference*. Nature Methods, doi:10.1038/nmeth.2016.

[8] Sheridan P, Kamimura T, Shimodaira H. *A scale-free structure prior for graphical models with applications in functional genomics*. PLoS One. 2010;5(11):e13580.

[9] Qian X and Dougherty ER (2013) *Validation of gene regulatory network inference based on controllability*. Front. Genet. 4:272. doi: 10.3389/fgene.2013.00272