

Invitation to M.Tech. Thesis Defense of Ayush Garg: July 27, 2020 (Monday): 14.00-15.00 IST

In Partial Fulfillment of the Requirements for the Degree of

M.Tech. CB

Ayush Garg (MT18243)

Will defend his thesis

Title: **“Developing Gene Interaction Network from Public Literature and Genomic Data to Learn Gene Regulations”**

IIIT-D Faculty and Students are invited

Date: July 27, 2020 (Monday)

Time: 14.00-15.00 IST

Google meet link; <https://meet.google.com/bpn-hqwi-hyd>

Examiner:	Internal:	Ganesh Bagler
	External/Internal:	Arjun Ray
	Advisor:	Tavpritesh Sethi (IIITD), Shruti Srinivasan (Elucidata)

Abstract

"Public Literature has immense knowledge which goes undermined due to current approaches. The reusability of this information has not been utilized to its full potential. Data mining and Natural Language Processing processes rely only on text publications^{1,2,3} while other processes rely only on the gene expression data (metadata) for analysis and result, eg. CREEDS⁴ and GeneQuery⁵.

Integrating the knowledge of 40,000 GEO datasets (Gene Expression Matrix) with their corresponding PubMed literature (abstracts) provides much powerful insight. This technique has an edge over others, as the textual data is endorsed by the genomic information of the metadata. Gene signature is taken as the input of a disease condition, to filter the publications which have a significant overlap of differentially expressed genes in metadata with the input signature. The biomedical keywords (disease, genes) are linked among each other using a defined genomic score (generated by p-value among the gene list in metadata) of each publication. These links are used to construct a dense disease-gene network. Only the important links are retained in the network on the basis of the p-value obtained by the disparity filter algorithm⁶. Node embeddings are calculated over the network using the node2vec algorithm⁷. Cosine Similarity of the embeddings with the master regulator gene in the disease condition is calculated to obtain the top 15 positive and negative keywords.

For the case of Hypoxia, "HIF1A" is considered as the master transcriptional regulator of cellular and developmental response⁸. The results of the pipeline are validated such that the "HIF1A" positively regulates 12 genes out of 15 keywords (12 genes and 3 disorders), and negatively regulates 5 genes out of 15 keywords (11 genes and 4 concepts)."

