**Paper : Ensembling of Gene Clusters utilizing Protein-protein Interaction Information**

The current paper is a combination of multiobjective optimization and deep-learning models to develop cluster ensemble system for gene clustering to generate consensus partitioning using protein protein interaction information. The proposed approach is used for clustering the genes by maintaining both biological relevance and statistical significance of the genes.

The proposed framework works in four phases:

1. filtering out the irrelevant genes from the microarray dataset
2. generation of diverse base partitioning to optimize three different cluster quality measures and generates a set of partitioning solutions on the Pareto optimal front
3. generation of a consensus partitioning using mentha scores to generating a weighted incidence matrix
4. finally, two approaches are used to generate a consensus partitioning from the obtained incidence matrix. The first approach is based on a traditional machine learning method, and another approach exploited a linkage graph and two deep neural models to generate the final clustering and thus, further providing labels to genes based on clusters.

Then these datasets are used to train two deep learning models, and those trained models are then used to generate final consensus partitionings.

## Prerequisities

* keras
* pandas
* Python 2.7+
* Sklearn
* Matplotlib 2.0+
* Numpy
* Cytoscape

[**Preprocessed\_dataset**](https://github.com/sduttap16/DeepEnsm/tree/master/1.Preprocessed_dataset)

This Folder contains python file for Preprocessing of datasets.   
Need of preprocessing : Among all the genes of microarray dataset, all the genes are not differentially expressed over the two sample classes. Differentially expressed (DE) genes are relevant enough for clinical purpose which are very useful for diagnosis and prognosis of a particular disease. For this purpose, We have filtered genes. which had variances greater than 10 percentile. we have calculated the t-statistic , which compares the difference between two sample means, in relation to variation in the data.