

Prediction of Disease Phenotype & Inference of Gene Networks: Revisting the DREAM5 Systems Genetics Challenge

Vankiripalli Mohammad Aaftab* Mohammad Faiz Khan**

** Department of Mechanical Engineering, IIT
Madras(email:aaftaabv@gmail.com)*

*** Department of Chemical Engineering
(mohammedfaiz13786@gmail.com)*

Abstract: The problem of Infering a Gene Regulatory Network from Microarray Data has been around in the community of Bioinformatics for a long time. Accurate prediction of Gene Regulatory Network is a very crucial step in creating model of a cell good enough to be used in Precise Drug screening of compounds. Our project is an attempt at revisting the 2014 DREAM5 challenge of Infering Gene Regulatory Network from Gene Expression Data of insilico soybeans and Predicting Phenotype of a Disease based on Genotype and Gene Expression Data of insilico soybeans.

In the last 5 years the fields of Machine Learning and Artificial Intelligence have seen a huge growth, thus bringing faster and more accurate algorithms into the front for inference of Gene Networks and Prediction of disease Phenotypes. This is our attempt to find some of the best available algorithms for these purposes and implement them on the 2014 DREAM5 dataset and see if we can build better models than those that existed at the time this competetion took place.

For the first part of the problem statement, we try to implement some Gene Regulatory Network Inference Algorithms like Partial Correlation, Bayesian Networks and some Multiple Ensemble Feature Importance Algorithms like GENIE3 (GEne Network Inference with Ensemble of Trees) and compare them based on accuracy, run time. The Gold standard Network has been included with the DREAM5 dataset.

For the second part of the problem statement, we plan on training various Machine learning models available like Regression, Random Forest, Deep Learning on the given training data and predict the Disease Phenotype of testing data based on Genotype data, Gene Expression Values. The Gold Standard has been given in the problem statement. Then we will compare various models based on accuracy and run time. We also present the best set of parameters for each model.
