

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

Vankiripalli Mohammad Aaftab *

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

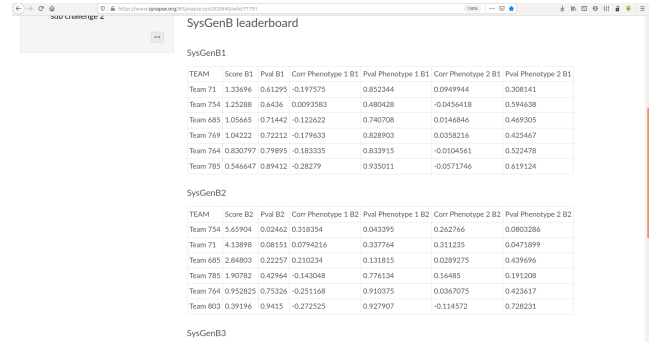
0.1 First Report

The DREAM5 Systems Genetics Challenge Part-B consists of 3 parts. In the First part, the problem is to predict the extent of disease phenotype existence based on the Genotype of a particular plant. There are 941 genotype variants in 200 plants (genotype variants represent either the presence or absence of a gene in each of the 200 plants). Hence, the training data for this problem consisted of a 200X941 boolean type matrix and the extent of disease presence in each plant is given as a separate 200X2 matrix with each plant having 2 different phenotype values both representing the effect a particular pathogen (*Phytophthora sojae*). Hence we had 200 plants with each plant having 941 features which can be used to predict 2 different phenotype values. So the Machine Learning model took 941 features from each of the 200 plants and fitted a model (separately) using these features to predict outcomes (phenotype1, phenotype2).

The prediction of phenotype was to be done on a testing set that consisted of 30 plants with variations across 941 genes given. Machine Learning Models were trained on the training set and these models were used to then predict the phenotype values (real numbers) for the 30 plants in the test data set. The predicted phenotype value was compared with the given gold standard phenotype values using spearman correlation coefficient and p-value.

Similarly in the second sub-challenge, the training data consisted of a real 200X28,395 matrix of gene expression values and a 200X2 real matrix representing phenotype values in each plant. And the prediction of phenotype had to be done only based on the given gene expression values in each plant. Again the models were fit on the training data and predicted on the test dataset that consisted of gene expression values of 28,395 genes in 30 different plants. The predicted phenotype value was compared with the given gold standard phenotype values using spearman correlation coefficient and p-value. Here also we had 200 plants with each plant having 28,395 features that can be used to predict 2 different outcomes. So the Machine Learning model took 28,395 features from each of the 200 plants and fitted a model using these features to predict outcomes (phenotype1, phenotype2).

Finally the score of each model was calculated as was done by the DREAM5 Systems Genetic Challenge i.e. sum of natural logarithms of p-values of spearman correlation between given gold standard phenotype values and pre-



SysGenB1							
TEAM	Score B1	Pval B1	Corr Phenotype 1 B1	Pval Phenotype 1 B1	Corr Phenotype 2 B1	Pval Phenotype 2 B1	
Team 71	1.33696	0.61295	-0.197575	0.852344	0.0949944	0.308541	
Team 754	1.25288	0.6436	0.0095983	0.880408	-0.0454618	0.394608	
Team 685	1.05665	0.71442	-0.122622	0.740708	0.0146846	0.469905	
Team 769	1.04222	0.72212	-0.179633	0.829903	0.0268256	0.425467	
Team 764	0.830797	0.79895	-0.183335	0.833915	-0.0504561	0.523478	
Team 785	0.546647	0.89412	-0.28279	0.935011	-0.0571746	0.619124	

SysGenB2							
TEAM	Score B2	Pval B2	Corr Phenotype 1 B2	Pval Phenotype 1 B2	Corr Phenotype 2 B2	Pval Phenotype 2 B2	
Team 754	5.65904	0.02642	0.338354	0.043395	0.262766	0.0803286	
Team 71	4.13898	0.08151	0.0794216	0.337764	0.311235	0.0473899	
Team 685	2.84803	0.22257	0.230234	0.131815	0.0289275	0.439696	
Team 785	1.90782	0.42964	-0.143048	0.776134	0.16485	0.191208	
Team 764	0.952825	0.75326	-0.251168	0.910375	0.0367075	0.423617	
Team 803	0.39196	0.9415	-0.272525	0.927907	-0.114572	0.728231	

SysGenB3							
----------	--	--	--	--	--	--	--

Fig. 1. DREAM5 SysGenB Challenge Leaderboard

dicted phenotype values. (There were two predictions in each subchallenge one corresponding to phenotype 1 and the other corresponding to phenotype2)

The models used for the subchallenge1 include: Linear Regression, Lasso Regression, Decision Tree, Random Forest, SVM, AdaBoost, Bagging Regressor, Gradient Boosting Machine (GBM), Homogeneous Voting Regression (using Random Forest Models and GBM), Cat -Boost Regressor.

The models used for the subchallenge2 include: Linear Regression, Lasso Regression, ElasticNet regression, Decision Tree, Random Forest, SVM, AdaBoost, Bagging Regressor, Gradient Boosting Machine (GBM), Homogeneous Voting Regression (using Random Forest Models and GBM).

From the above leaderboard, it can be seen that the best score in the challenge B1 (i.e. inference of Phenotype only from Genotype data) is 1.33696, whereas my score for Linear Regression is 2.2044070348685127. Till this point of time, the given data was scaled using standard scaler method in Sklearn and was used to fit a model using various machine learning algorithms. The other ML algorithms that beat the then best score are Ridge Regression, Bagging Regression, GBM regression, Voting Regression using only GBM models, and Catboost Regressor.

I plan make further improvements to the model by using PCA (linear dimensionality reduction), Manifold Learning (Non-linear dimensionality reduction) etc. Right now the number of features is about 941 which might introduce curse of dimensionality to some models. Next step will be to analyse data using dimensionality reduction.

The score was calculated by this procedure:

Spearman correlation coefficient and P-value was found between the real values of phenotype1 and predicted values of phenotype1 and the same 2 values were found between real values of phenotype2 and predicted values of phenotype2. The score is $-1 * (\text{natural logarithm of p-value of spearman correlation between phenotype1's real and predicted values} + \text{natural logarithm of p-value of spearman correlation between phenotype2's real and predicted values})$.

For the subchallenge B2 (i.e. prediction of phenotype based only expression values of 28,395 genes), the best score in the leaderboard is 5.65904. My decision tree regressor model had a score of 6.283454289075021. Other models that surpassed the best score on leaderboard include Random Forest, Gradient Boosting Machine (GBM), GBM voting regressor.

The B2 challenge had an even larger feature space (28,395 features) than B1 challenge hence we expect the models to perform even better on this data after linear or non-linear dimensionality reduction is applied.

The score was calculated by the same procedure as indicated above. The complete scores of each model on SysGenB1 and SysGenB2 is summarised in the below table.

The following table illustrates the Model vs problem statement information of used models.

Table 1. Scores Table

Model	B1 Score	B2 Score
Linear Regression	2.2044070348685127	1.4372172661318754
Ridge Regression	2.1819492446231603	1.4371453159087157
Decision Tree Regression	0.4965799903304826	6.283454289075021
Random Forest Regression	0.45537125620964347	5.792798653297446
SVM Regression	0.6627204323246874	1.7120837632245094
Adaboost Regression	0.42658865020883785	1.5024139668918883
Bagging Regression	1.7619184978612388	3.1444887594932824
GBM Regression	1.6468578041050397	5.986443810068452
Random Forest Voting Regression	0.528437716104597	3.939148750324295
GBM Voting Regression	1.5982751592676547	5.965925663149765
Catboost Regression	1.428776661540361	nan

0.2 Further Work

One noteworthy point is while simple linear models worked very well in SysGenB1 they performed poorly in SysGenB2, similarly the more complex models failed in SysGenB1 and succeeded in SysGenB2. This might indicate a more complex relation between Phenotype and Gene Expression as compared to Phenotype and Geneotype Variations.

Final aim of the SysGenB challenge will be to try and explain the type of relation between Phenotype and Genotype and Gene expression (i.e. maybe determine which particular genotype variant or Gene expression value is having a huge impact on Phenotype value.) based on the performance of different models on the data.

The model building of SysGenB3 (prediction of Phenotype based on both Genotype Variants and Gene expressions) is to be done along with Gene Network Inference (SysGenA).