Predicting Genetic Interactions

Wenyan Li, October 18

Overview

In this project, yeast genotype is translated to cell growth phenotype by predicting pairwise gene interactions using featurization (ontotype) and random forest procedure described in [1].

Experimental Procedures

Building feature vectors and interaction matrix

According to [1], by interpreting genetic interactions among genes annotated to a term as interactions between the genes of different terms at a lower scale in GO, a feature vector (ontotype) whose size equals to number of terms (5125) can be built for each pair of genes from the gene-term GO data from Yu, et al[1]. That is, we add(or subtract) one at the position of the term for each gene in the pair if it is annotated under the term. A gene-term dictionary, which returns all related terms of a specific gene, is used to help build the ontotype. In this procedure, some genes in interactions are observed to be not annotated in any of the terms so they are discarded as they cannot be represented using feature vectors. It is also observed that using positive or negative (+1 or -1) representations does not necessarily influence the results. The reason is that at every decision split in a tree, one-third of all unused features were considered for the optimal split, defined by the minimum squared error, which won't be influenced if our feature vector is negative or not.

Meanwhile, an interaction matrix of pairwise gene interaction scores can be formed by utilizing the genetic interactions data from Collins, et al. (Nature, 2007)[2]. As genes are assumed that they can not interact with themselves, interactions scores on the diagonal of the matrix are ignored in the dataset. Also, gene pairs that do not have an interaction score can not be assumed to have non-interaction with a zero score as we do not know their interaction status, and they should not be formed into ontotypes. I misunderstood this point and naively assign 0 scores for the pairs which resulted in weird performance.

Then we use the feature vectors representing "ontotype" as input vectors, and pairwise interaction scores as targets to train and predict genetic interactions.

Predicting interaction with random forest regression

Genetic interaction scores are predicted with random forest regressor on simulated and real datasets. Four-fold-validation is used on both datasets and Pearson correlation coefficients are calculated for evaluation. Following the thresholds for the interaction scores, gene pairs are further categorized into negative interactions (score < -2.5), positive interactions (score > 2), or no interaction. During regression, instead of using 300 trees for each random forest as described in [1], which would result in a large cost in computation, 10 trees are used in this project. Trees are grown to maximal depth as explained in [1] and the data is shuffled before splitting into four folds in order to avoid any element of bias/patterns in the split datasets before training the model.

Datasets and Results

After implementing the aforementioned algorithm for predicting genetic interactions on both simulated data and real data, we can obtain the Pearson correlation coefficients as follows.

Table.1 Results for simulated and real data

Dataset		**Simulated data**	**Real data**
Number of unique genes		100	664
Number of terms		99	5125
Pearson correlation coefficients	**four-fold**	0.056, 0.0324,0.0430, 0.047	0.474, 0.486, 0.474, 0.484
	averaged	0.045	0.479

Shown in Table.1, the Pearson correlation coefficients of the real dataset (r = 0.479) is higher than the result presented in [1] where average r = 0.35. The main reason probably lies in the difference of the two interaction dataset. As the Costanzo, et al. (Science, 2010) data is too large to run in a reasonable amount of time, a curated smaller interaction dataset from Collins, et al. (Nature, 2007)[2] is used in this project. Also, as mentioned in [1], the terms of the annotated genes also affect the prediction.

As mentioned in the experimental procedures, the data is shuffled before splitting into four folds. Without shuffling, the correlation would be likely influenced by the patterns of split datasets as shown in Table.3.

The categorization result is summarized in Table.2 and it can be observed that compared to predicting positive and negative interactions, the method has a relative higher precision and recall on predicting no interaction, which is probably caused by the fact that we have most training samples in the no interaction category.

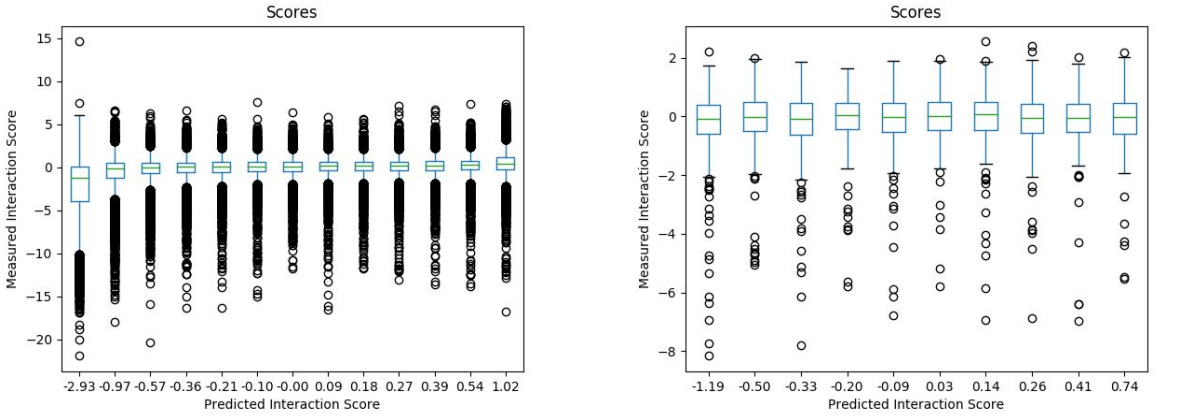
Measured genetic interaction scores versus predicted scores of simulated and real datasets are shown in Figure.1. For the real dataset, the bins are divided in [-16.07, -1.57, -0.91, -0.6, -0.4, -0.26, -0.15, -0=.06, 0.02, 0.1, 0.18, 0.27, 0.36, 0.48, 0.69, 5.66]. Rather than presenting the predicted score in evenly distributed spans as Figure.3C in [1], Figure.1 ensures same number of predictions for each bin and the columns are the mean values. Thus its more direct to analyze the distribution of the predicted scores. The class imbalance displayed in Table.2 is consistent with Figure.1(a), from where we can also easily observe that we have most gene pairs measured and predicted as no interaction. In addition, Figure.1(a) also shares a similar trend in Figure.3C in [1]. In both figures, measured interaction score matches better with the predicted score and has a smaller variance(tighter distribution) around no interaction category, while presenting a loose match for other two types of interactions. For simulated data, no obvious class bias is observed.

Table.2 Precision and recall on real data prediction

Real Data	**Precision**	**Recall**
Negative interactions (score<-2.5)	0.29	0.56
No interactions	0.98	0.93
Positive interactions (score>2)	0.07	0.58

Table.3 Effects of shuffling during 4-fold cross validation:

Simulated data	Without shuffling	0.047, 0.026, -0.019, 0.07 (average: 0.031)
	After shuffling	0.056, 0.032, 0.043, 0.047 (average: 0.045)
Real data	Without shuffling	0.376, 0.433, 0.372, 0.461 (average: 0.411)
	After shuffling	0.474, 0.486, 0.474, 0.484 (average: 0.479)



(a) Measured genetic interaction scores versus predicted scores of real dataset

(b) Measured genetic interaction scores versus predicted scores of simulated dataset

Figure.1 Measured genetic interaction scores versus predicted scores

Reference

[1] Yu M. et al. (2016) Translation of genotype to phenotype by a hierarchy of cell systems. *Cell System* 24;2(2):77-88

[2] Collins, S. R. et al. Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. Nature 446, 806–810 (2007).

Appendix

Four-fold validation results on real data and corresponding confusion matrix

Pearson correlation coefficients are calculated using scipy.stats.pearsonr
-training size: 109924 testing size: 36642
For fold: 0 , the correlations is: 0.474094159198

Predicted interaction	0	1	2
Actual interaction			
0	436	1060	1240

1 1973 8211 5544
2 2572 6522 9084

-training size: 109924 testing size: 36642
For fold: 1 , the correlations is: 0.48556516452
Predicted interaction 0 1 2
Actual interaction
0 421 1113 1231
1 1989 8126 5421
2 2597 6751 8993

-training size: 109925 testing size: 36641
For fold: 2 , the correlations is: 0.474034305647
Predicted interaction 0 1 2
Actual interaction
0 428 1080 1194
1 1942 8087 5543
2 2448 6877 9042

-training size: 109925 testing size: 36641
For fold: 3 , the correlations is: 0.483916062072
Predicted interaction 0 1 2
Actual interaction
0 398 1065 1291
1 1936 8232 5523
2 2420 6704 9072