MPRA-eQTL prediction method

Yue Li
Computational Biology Group
Computer Science and Artificial Intelligence Lab
Massachusetts Institute of Technology

December 10, 2015

1 Methods

1.1 Featurization

All for predictions, I complied based on the hg19 genomic coordinates (i.e., chromosome and position) of the centered variant the following 1978 features:

- 1-600: position-dependent nucleotide identity of the 150-mer sequence with 74 nt upstream and 75 nt downstream of the centered variant. For each nucleotide position, I used 4 columns corresponding to 'A/C/G/T' to encode the nucleotide identity where one and only one of the 4 entries is 1 and the other 3 entries are 0;
- 601-1964: position-independent nucleotide frequency of 1-mer, 2-mer, ..., 5-mer;
- 1965: phast 46way placental conservation score obtained from UCSC database;
- 1966-1978: 13 ENCODE epigenomic tracks in K562, namely E123-DNase, E123-H2A.Z, E123-H3K27ac, E123-H3K27me3, E123-H3K36me3, E123-H3K4me1, E123-H3K4me2, E123-H3K4me3, E123-H3K79me2, E123-H3K9ac, E123-H3K9me1, E123-H3K9me3, E123-H4K20me1.

1.2 Prediction methods

I used the following 6 different machine learning methods to predict Regulatory_Hit (classification) and C.A.log2FC/C.B.log2FC (regression). For all methods, I used an existing R library, and default settings were used unless noted otherwise.

1. elasticnet: I used R package `glmnet` [1] with alpha set to 0.5 and determined lambda by cross-validation (`cv.glmnet`); I set 'family' parameter to "binomial" for classification and "gaussian" for regression;

2. gbm: I used R package `gbm` [2] to train generalized boosted regression models with number of trees set to 500 and set 'distribution' parameter to "bernoulli" for classification and "gaussian" for regression;

3. lasso: same as elasticnet but with alpha set to 1;

4. ridge: same as elasticnet but with alpha set to 0;

5. randomForest: R package `randomForest` [3] was used with number of trees set to 500

6. svm: support vector machine (SVM) was applied to the data through R package `e1071` [4]

1.3 Method evaluations

To evaluate each method, I performed for each method cross-validation. In particular, I divided the training datasets containing 3044 examples into 22 folds each corresponding to one of the 22 chromosomes. Each method was trained on 21 folds and validated on the remaining fold. For classification task (i.e., predicting Regulatory_Hit and emVar_Hit), I used area under the curve of receiver operating characteristic (ROC) and precision-recall (PRC) to evaluate the model performance. For regression (i.e., predicting C.A.log2FC, C.B.log2FC, and LogSkew.Comb), Spearman correlation between the predicted fold-change and the observed fold-change was used. Additionally, I created a null model for each method by randomly shuffling every columns of the feature matrix and repeated the same CV procedure on the shuffled data.

2 Results

2.1 Part 1: prediction of regulatory hits and expression fold-change

Our results shows that the 3 regularized linear regression model confers favorable performance over the other 3 non-parametric methods in the classification task (Figure 1). In the regression task, however, SVM confers the best performance on both C.A.log2FC and C.B.log2FC predictions (Figure 3 upper and middle panels, respectively).

In the submission of part 1, I included the prediction results from all 6 methods for predicting "Regulatory_Hit", "C.A.log2FC", and "C.B.log2FC" on the 3006 testing cases using the 1978 features mentioned above. For the 3 regularized linear regression models (implemented in `glmnet`), I calculated the standard deviation based on square difference between the predicted values of the best lamdba and the worst lambda obtained from the internal cross-validation routine. For other methods, there is no confidence interval or standard deviation on the predicted values, so I placed zeros for all of the 3 "SD" columns as dummy values instead.

2.2 Part 2: prediction of expression modifying variants and expression fold-change

Same as Part 1, I applied the same 6 prediction methods on predicting "emVar_Hit". The best performing methods are ridge regression in terms of ROC and elasticnet in terms of PRC (Figure 2). Compared to regulatory hit predictions, all 6 methods performed relatively worse in terms of PRC mainly due to the 3 times smaller positive examples: 108 emVar versus 357 regulatory hits out of the 3044 training examples. None of the methods were able to predict "LogSkew.Comb" as the Spearman correlations are no different or even worsen than those that resulted from the permuted data (Figure 3 bottom panel). Nonetheless, I submitted the predictions of the 6 methods on "emVar_Hit" and "LogSkew.Comb" using the test data.

3 Ongoing works

I would like to identify the predictive features among the 1978 features, which may not only improve the method performance but also perhaps give insights on the specific eQTL categories that the MPRA technique is capable of (not) detecting as well as further improvements of the experimental protocol.

References

[1] Friedman, J., Hastie, T., and Tibshirani, R. (2010) *Journal of Statistical Software* **33(1)**, 1–22.

[2] with contributions fromothers, G. R. gbm: Generalized Boosted Regression Models (2015) R package version 2.1.1.

[3] Liaw, A. and Wiener, M. (2002) *R News* **2(3)**, 18–22.

[4] Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien (2015) R package version 1.6-7.

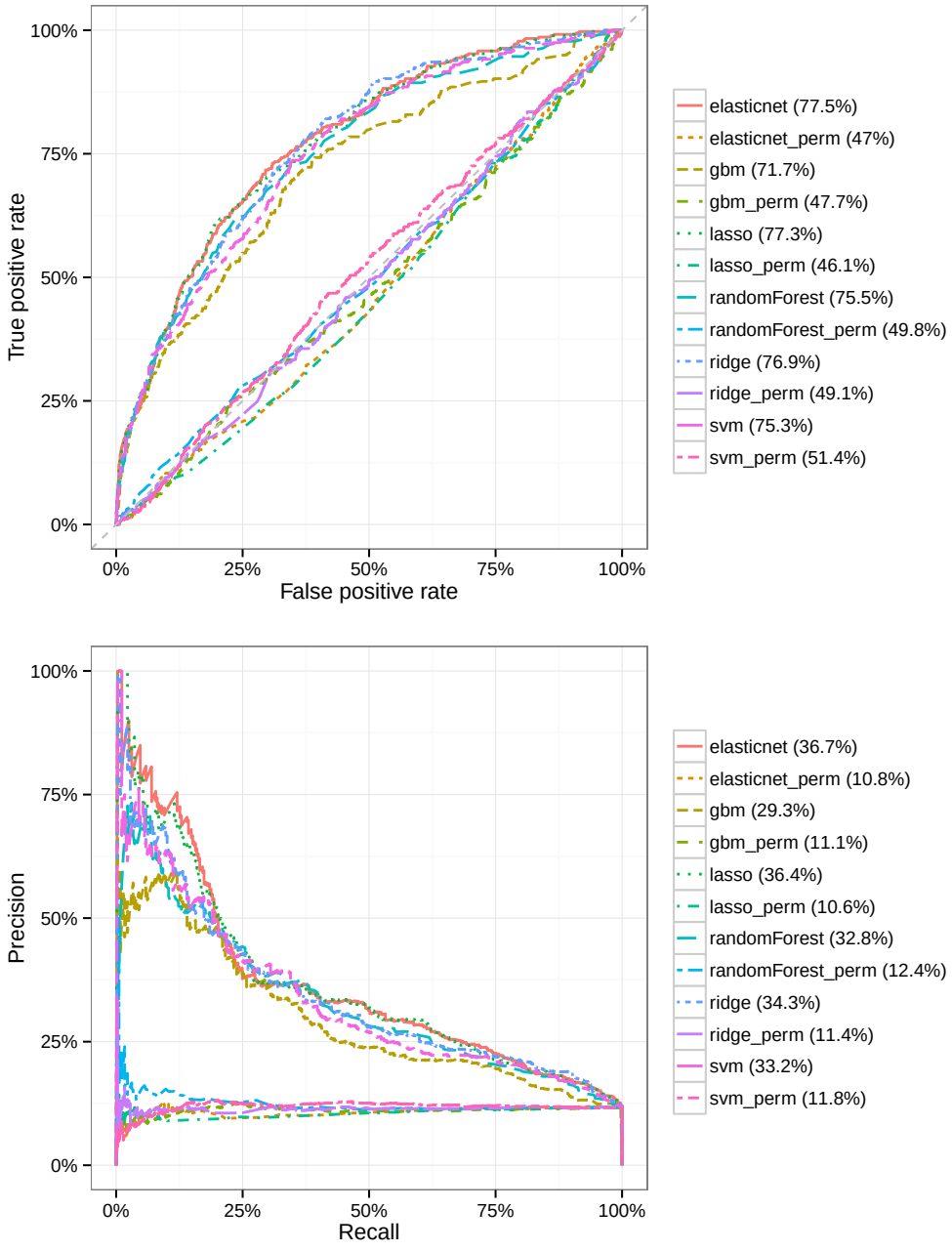

Figure 1: Power analysis on predicting regulatory hits. Six methods were assessed in their abilities to predict regulatory hits using the 1978 sequence and epigenomic features. Performance was assessed by leave-one-chromosome-out cross-validation. For each of the 6 methods, a null model was also constructed and evaluated using the permuted version of the training data.

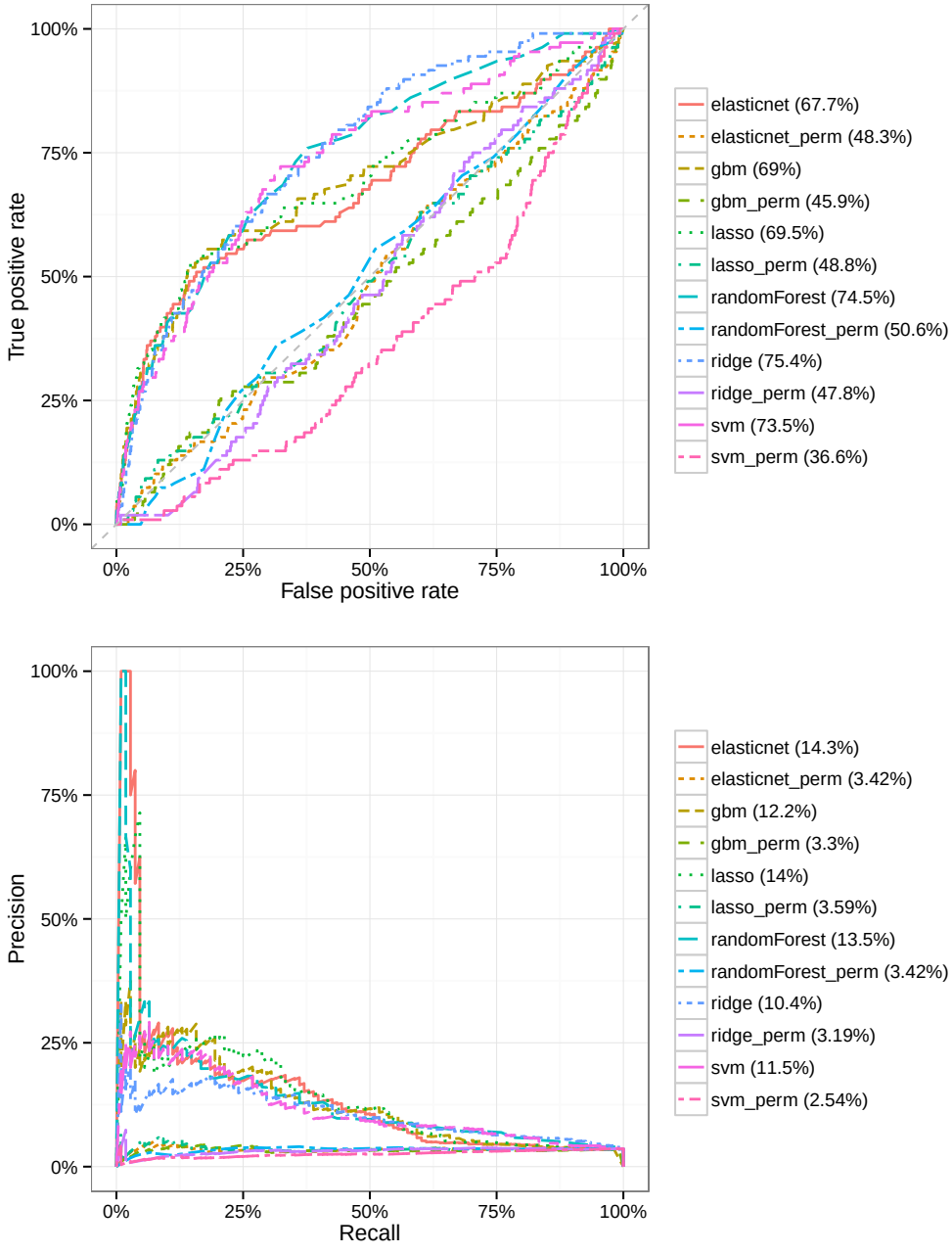

Figure 2: Power analysis on predicting expression modifying variants. Same as 1 except for predicting emVar_Hit.

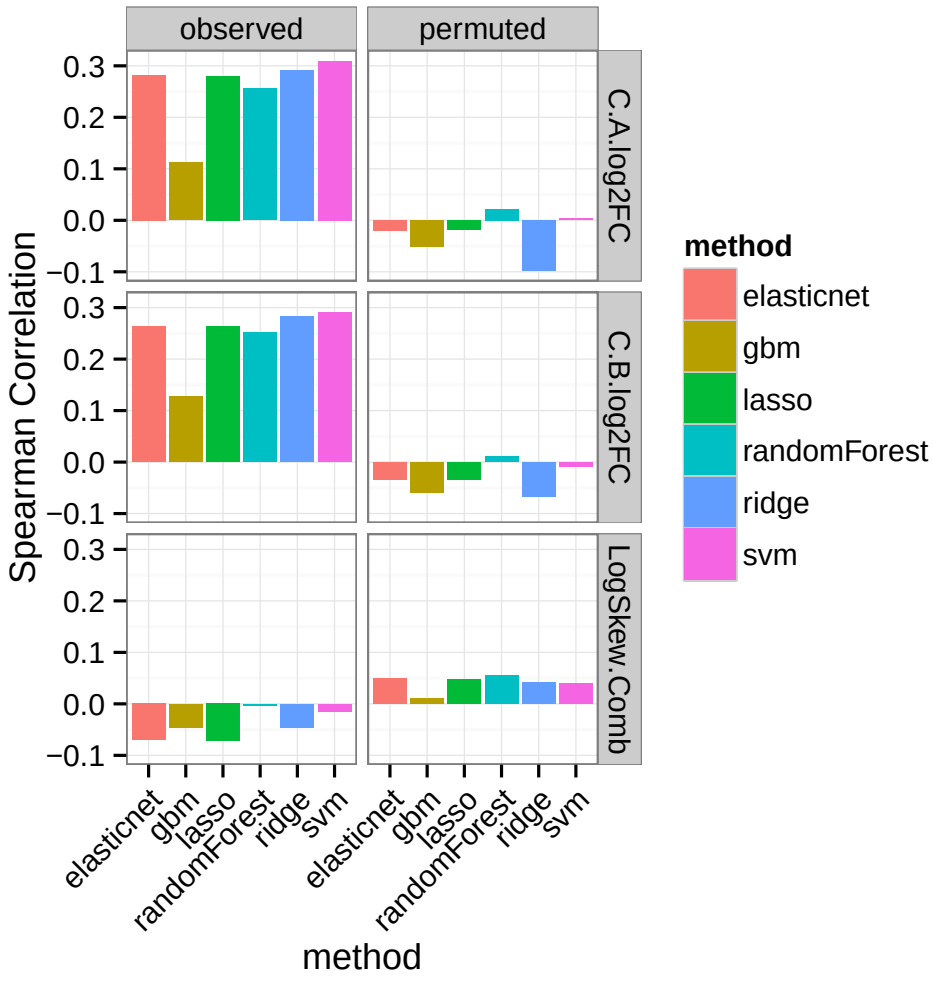

Figure 3: Spearman correlation on the training dataset. Six methods were assessed in their abilities to predict expression fold-change in the control and mutated sequences using the 1978 features.