Explore factors that distinguish between informative contacts and uninformative contacts

Introduction

Chromosome folding has been observed for a long time. People have shown that chromatin contacts in the 3-dimensional space play important roles in the genomic regulatory network. Enhancer - promoter interaction is one important type of long range interaction, that has been proved to be crucial in controlling gene expression levels. Also it has been shown that disruption of chromatin architecture would change gene expression levels[1]. However, it is unclear yet how to distinguish between the contacting pairs involved in genetic regulation versus the contacting pairs of structural purpose. This question is very important in terms of understanding the role of folding in the genetic regulation network, and the mechanism behind chromatin architecture.

From previous research work, it was observed that long range eQTL pairs are enriched of contacting shown by non-zero HiC values. However, I found it very hard to leverage the contacting information to increase the power of identifying long range eQTL. One possible reason is that chromosomes fold for a variety of reasons, including performing genetic control. Thus the long range eQTL data provides a good dataset for assessing the regulation potential between two contacting regions.

Data

ATAC-seq data for adipose subcutaneous tissue was downloaded from ENCODE (https://www.encodeproject.org/experiments/ENCSR540BML/) using library ENCLB503MYH. The fastq files were mapped using bowtie2. Resulting bam files were converted to sam files using bamtools. Homer and MACS2 were both used to perform peak calling.

CTCF ChIP-seq data was downloaded from ENCODE where the same tissue from the same individual was sequenced (https://www.encodeproject.org/experiments/ENCSR762PCG/). The library ENCFF997ZMP was used where bam files were available.

[1] Zuin, Jessica, et al. Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *Proceedings of the National Academy of Sciences* 111.3 (2014): 996-1001.

HiC data was obtained for GM12878 cell line from GSE63525. Reads were normalized according to the original paper[2] .

Functionality score for chromosome regions were obtained from segway encyclopedia (ttp://noble.gs.washington.edu/proj/encyclopedia/). This feature represent how important the sequences are, and is assumed to be a surrogate for the sequence information as well as histone modifications.

Long-range eQTL pairs were derived from previous research work, where the pairs were defined to have distance > 100kb and matrixeQTL was used. Genes with mappability <= 0.8 were removed because of the potentially incorrect mapping of HiC reads.

Data was generated for 10 chromosomes (chromosome 13 - chromosome 22) from adipose subcutaneous.

Methods

A machine learning framework was built to reach the goal of identifying the informative factor that explains why some interacting pairs are associated while some are not. The validity of the model was tested, and then the important features were extracted to help understand the potential mechanism behind interacting associations.

First analysis was restricted to pairs that have contact based on non-zero HiC value between the 10kb window of the SNP and the 10kb window of the gene, because the goal of this project is to research the factors that make interactions functional instead of studying the role of interaction itself.

Then features including ATAC peak reads, CTCF peak reads and functionality scores were annotated to both the variants and genes. For variants, peak reads were obtained by overlapping the peaks and the window of 10kb with the variant as the center. For genes, peak reads were obtained by overlapping the peaks and the window from 2kb upstream of the gene start to the gene end. In this way, information of the promoter regions could be included. Both the variants and the genes were annotated with the closest TAD (topological domains).

Pairs were marked as eQTL if the BH corrected FDR was smaller than 0.05, while pairs with top high p-values were matched for distance between the variant and the gene with the eQTL pairs and marked as negative sets. Matching for distance is an important step in picking the negative sets, because distance is known to impact the association potential and thus would be a strong confounder in the model. One thing to notice is that for eQTL pairs, genes could be associated

[2] Rao, Suhas SP, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 159.7 (2014): 1665-1680.

with multiple SNPs. To avoid bias caused by genes with a large number of SNPs, the genes in eQTL pairs were restricted to have at most 10 associated SNPs by randomly selecting 10 associated SNPs. An equal number of pairs were randomly sampled from the negative set.

Data from 5 of the chromosomes were used for training and data from the other 5 for testing. Among the training data, there are two ways of deriving the validation set. One was to use cross validation with 10 folds, and the other was to use data from one chromosome as validation set and data from the other three chromosomes as training. The second method helps to alleviate overfitting in the training data.

Decision tree, random forest, SVM and logistic regression were applied to fit the data. The tree algorithms were used because they not only provide non-linear models, but also can be used to assess the importance of features. For decision tree and random forest, the parameter of maximum depth was explored using validation data. Accuracy was computed as the proportion of correct prediction. This was used as the evaluation matrix because the negative set and the positive set has equal number of pairs. To determine the validity of the models, the labels of being eQTL or negative were permuted. Accuracy using the permuted data was also computed to compare to the accuracy using the true data. Importance of the features were extracted to understand the role of these factors in distinguish information interactions and non-informative ones.

Results

1. Exploratory plots

The distribution of features were plotted for visualization. It is not surprising that for the two features hic value and functionality score of the gene, some chromosomes show quite obvious pattern of different distributions, while some chromosomes show very little difference (Fig 1). This is partly because of the small sample size of positive and negative sets (100 - xxx for all the chromosomes), while arises naturally from the long range eQTL analysis. Wilcoxon rank-sum test was used to find the significantly different distributions (excluding the zero values because they are not informative). The results show that chromosome x, x, and xx has significantly different distributions of functionality score of genes (p-value < 0.05), and chromosome x, x, and xx has significantly different distributions of HiC values (p-value < 0.05). Besides, fisher's exact test was used to compare the proportion of zero values. It shows that chromosome x, x, and xx has significantly lower proportion of functionality score of genes being zero (p-value < 0.05).

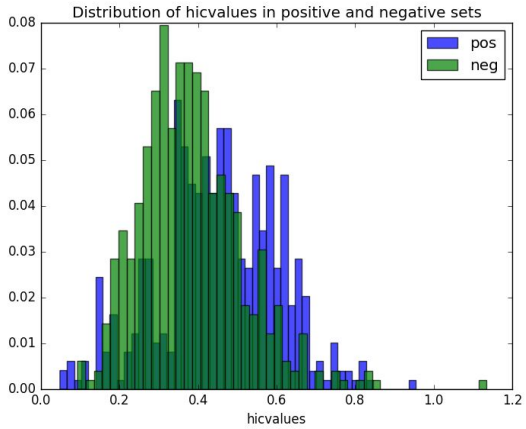


Fig 1A. Distribution of hic contacting values (chr15)

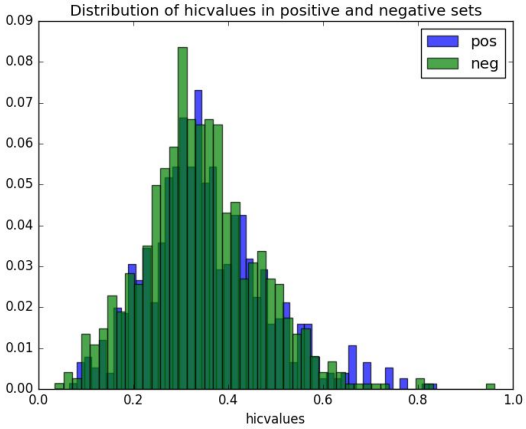


Fig 1B. Distribution of hic contacting values (chr17)

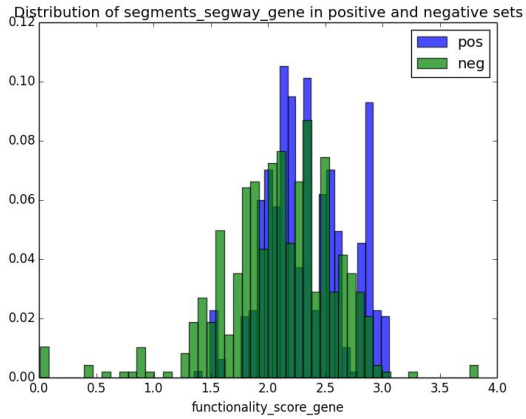


Fig 1C. Distribution of functionality score over genes (chr14)

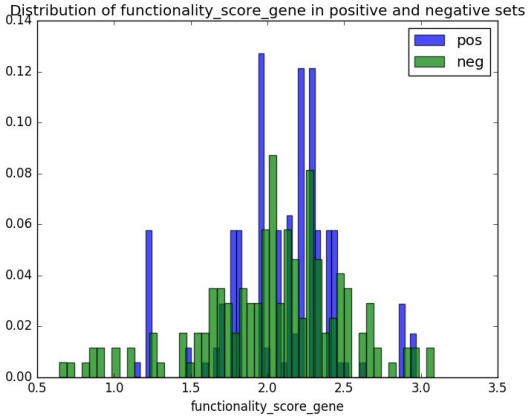


Fig 1D. Distribution of functionality score over genes (chr20)

2. To determine the optimal parameter for decision tree and random forest

Two way of deriving validation data sets as described in Methods were used to determine the optimal maximum depth of the tree in decision tree and random forest model. Fig. 2A and Fig. 2C show the results of using cross validation, where we can see clearly that the accuracy in the validation set increases with larger tree depth, while the accuracy in the test set decreases. This means that the validation data set is not generalized well to the test set, and using this way of validation to obtain the optimal tree depth would be more likely to result in overfitting. Fig. 2B and Fig. 2D show the results of using another chromosome as the validation set. The trend of accuracy is similar in the validation and in the test data sets, indicating better generalization.

The accuracy of test data using the optimal tree depth that maximizes the accuracy in the validation data are summarized in table 1.

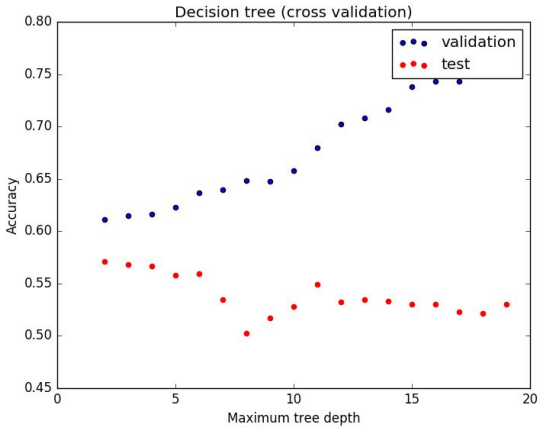


Fig 2A. Accuracy vs. maximum depth in decision tree (uses cross validation)

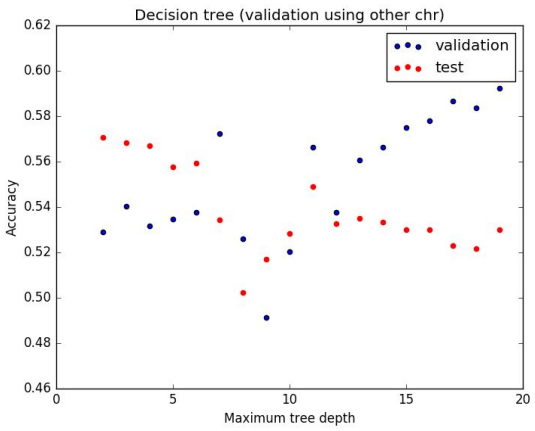


Fig 2B. Accuracy vs. maximum depth in decision tree (validation set using another chr)

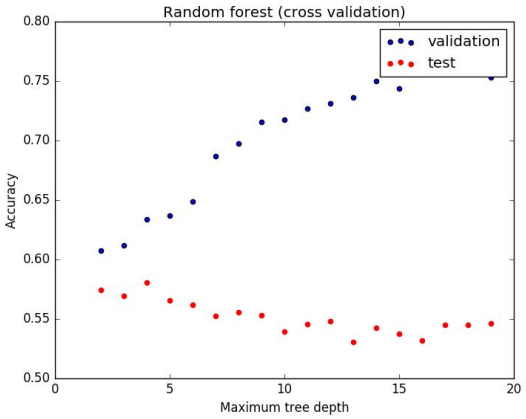


Fig 2C. Accuracy vs. maximum depth in random forest (uses cross validation)

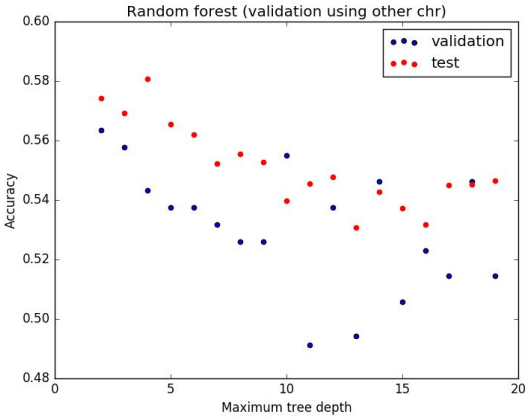


Fig 2D. Accuracy vs. maximum depth in random forest (validation set using another chr)

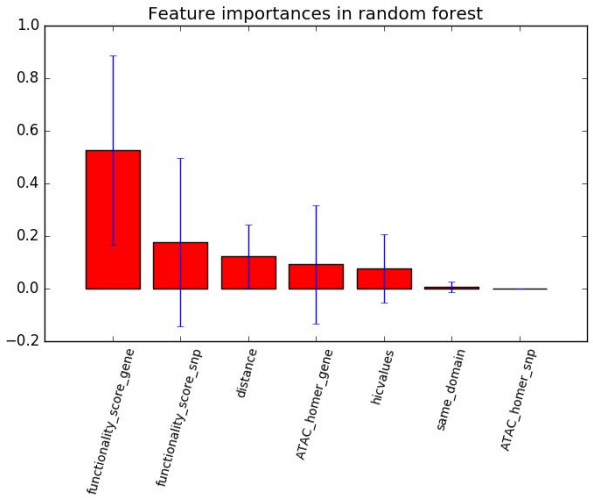
3. Compare accuracy rate to permuted data

After permuting the labels, the models were fit again. The distribution of accuracy for permuted data was shown in Fig. 3A and Fig. 3B. Empirical p-value for the accuracy is 1 in both random forest and decision tree, showing that the model did improve the prediction power. Table 1 shows the accuracy of test data for both true data and permuted data (mean value).

Table 1. Accuracy in the test data				
Model	Validation set	Permuted or not	Maximum depth	Accuracy
Decision Tree	Cross validation	True data	18	0.50
		Permuted data (mean)	4	0.50
	Validation using different chromosome	True data	8	0.56
		Permuted data (mean)	5	0.50
Random Forest	Cross validation	True data	15	0.545
		Permuted data (mean)	11	0.50
	Validation using different chromosome	True data	4	0.58
		Permuted data (mean)	11	0.48

4. Importance of features

The importance of features is defined as the fraction of the decisions they contribute to. The intuition is that features that rank higher on the tree have larger impact on the split.



Conclusions

This project shows that among the several features tested, segway functionality score plays an important role in assigning association potentials for interacting pairs. This makes sense because genes that are more actively expressed or more relevant to functions in this specific tissue are more likely to be involved in the genetic regulatory network. Also the effect is not linear, that interactions between the features are identified by the tree algorithm. However, a very limited number of features is tested in this project, that may partly explain the small increase in accuracy compared to 0.5.

From the comparison between two validation sets, we can see that it is crucial to use a well generalized validation set in order to pick a hyper-parameter for the model. Also, the better performance of using a different chromosome as validation shows that the chromosome architecture may behave differently between chromosomes.

Further work could restrict the variants' regions to the regulatory regions, as identified from GWAS, eQTL studies and experiments. Because theoretically there should be structural contacts for any regions including the regulatory ones, while regulatory regions could provide more enriched signal in regulation potential and thus increase the power of the model.

Besides the tree algorithms, hierarchical graphical models could be used in this scenario to account for information specific to genes, and also chromosomes. In the hierarchical model, nodes should present the potential of two contacting regions being regulatory associated. The potential on a specific level should be controlled by the potential from the higher level and covariates on the same level. In this way, we can have a sense of the importance of features on different levels.

Also, the model could involve more features, and even the sequence information.

Reference

Course project statements

This project is a side project for the long range regulation potential research project. I observed this situation of some interactions have associations while some don't, and have been wondering about this for a while.