Cross-Platform Prediction of Regulatory Activities

Runzhe Li

JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

July 16, 2019

Content

1. Introducution
2. Methods
3. Results
4. Application
5. Conclusion

Content

1. Introducution
2. Methods
3. Results
4. Application
5. Conclusion

Introduction

- Understand the complex regulome-transcriptome relationship
- Leverage gene expression profile to predict DNase I level

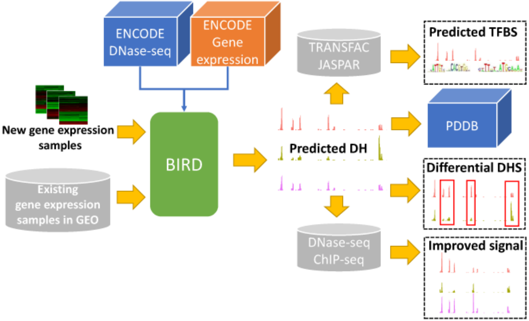


Figure: The work flow of BIRD

Introduction

- Cross-Platform?
- Expand the utility of GEO samples
- e.g. Use the human exon array data to train the prediction model and apply it to the microarray data

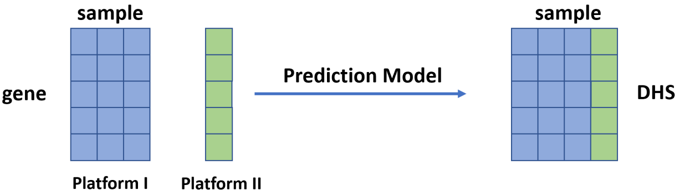


Figure: The sketch of cross-platform prediction

Content

1. Introducution

2. Methods

3. Results

4. Application

5. Conclusion

Methods: Data Preprocessing

- DNase-seq data: Bowtie
- Exon array data: The Affymetrix Human Exon 1.0 ST Array, GeneBASE
- Microarray data: Gene Expression BARCODE, GPL 96
- Training and test datasets partition

Methods: Problem Formulation and Notations

- Goal: use gene expression to predict DH level
- Let $Y^{(l)}$ be the DH level of genomic locus l, and X be the design matrix $C \times G$ of gene expression data, where C and G is the number of samples and genes respectively.
- Consider the prediction model:

$$Y^{(l)} = X\beta^{(l)} + \epsilon$$

where $l = 1, 2...L$, therefore we fit L separate regression models.

Methods: Problem Formulation and Notations

- Goal: cross-platform prediction
- Idea: apply BIRD model
- Problem: platform effect
- How to solve: normalization

Methods: Problem Formulation and Notations

- Suppose there is a new sample $\tilde{x}$ drawn from another platform, we want to apply the fitted model to predict the DH level : $\tilde{Y}^{(l)} = \tilde{x}\beta^{(l)}$?
- Assume $f(\cdot)$ is the normalization function, then we get the regression model

$$\tilde{Y}^{(l)}_{norm} = f(\tilde{x})\beta^{(l)}$$

Normalization

We introduce 4 normalization methods

- Sample-quantile method
- All-sample method
- Neighboring-sample method
- Gene-quantile method

Sample-quantile normalization

Algorithm 1 Sample-quantile normalization

```
Require: Exon array data matrix $X_e : G \times C_1$, microarray data matrix $X_m : G \times C_2$
Compute the mean quantile vector of the exon array data $X_e^q : G \times 1$
Assign the mean vector $X_e^q$ to each column of the microarray data according to its order
Obtain the normalized microarray data $X_m^{norm}$
```

All-sample normalization

Algorithm 2 All-sample normalization

```
Require: Exon array data matrix $X_e : G \times C_1$, microarray data matrix $X_m : G \times C_2$
Estimate $\mu_1, ..., \mu_G$ and $\sigma_1, ..., \sigma_G$, such that the linear transformation $T_i = \frac{x-\mu_i}{\sigma_i}, i = 1, ..., G$ applied to each row of the whole microarray samples can generate the same mean and standard deviation as exon array data.
```

Neighboring-sample normalization

Algorithm 3 Neighboring-sample normalization

```
Require: Exon array data matrix $X_e : G \times C_1$, microarray data matrix $X_m : G \times C_2$
1: For each exon array sample, select $k$ largest cross-gene correlation microarray samples and remove the duplicate samples.
2: Obtain the neighboring samples data matrix $X_m^{nbr}$
3: Estimate $\mu_1, ..., \mu_G$ and $\sigma_1, ..., \sigma_G$, such that the scaling $T_i = \frac{x-\mu_i}{\sigma_i}, i = 1, ..., G$ applied to each row of the neighboring samples $X_m^{nbr}$ can generate the same mean and standard deviation as exon array data.
4: Apply the linear transformation in step 3 to the microarray data matrix $X_m$.
```

Gene-quantile normalization

Algorithm 4 Gene-quantile normalization

```
Require: Exon array data matrix $X_e : G \times C_1$, microarray data matrix $X_m : G \times C_2$
1: Obtain the neighboring samples data matrix $X_m^{nbr}$ using algorithm 2
2: Sort the exon vector $X_{e,i}^{sort}$ and the neighboring microarray vector $X_{m,i}^{nbr,sort}$ for each row $i$.
3: Fit the LOESS regression model to $(X_{e,i}^{sort}, X_{m,i}^{nbr,sort})$
4: Given a new microarray sample, predict the normalized value for each row according to the LOESS model in step 3
```

Methods: BIRD model

Consider the prediction model:

$$Y^{(l)} = X\beta^{(l)} + \epsilon$$

- Step One: Variable Clustering
- Step Two: Fast Variable Screening

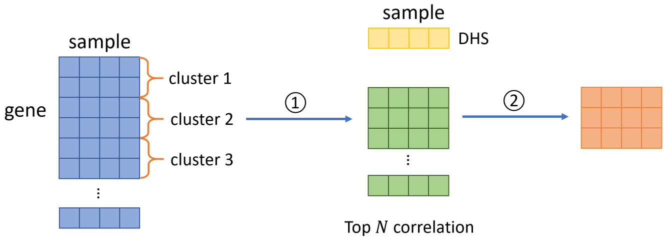


Figure: The BIRD model

Methods: Parameter Tuning

- Consist of three hyper parameters
  - the cluster number K
  - the predictor number N
  - the gene number M
- Determined by 5-fold cross-validation
- Select the genes with high cross-cell-type correlation

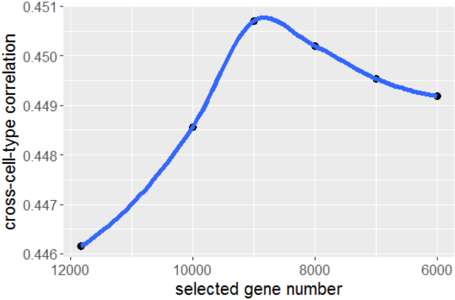


Figure: The relations between cross-cell-type correlation and gene number

Content

1. Introducution
2. Methods
3. Results
4. Application
5. Conclusion

Results

We use the following metrics to evaluate the model performance.

- Cross-locus correlation
- Cross-cell-type correlation
- Prediction squared error

Results: cross-locus correlation

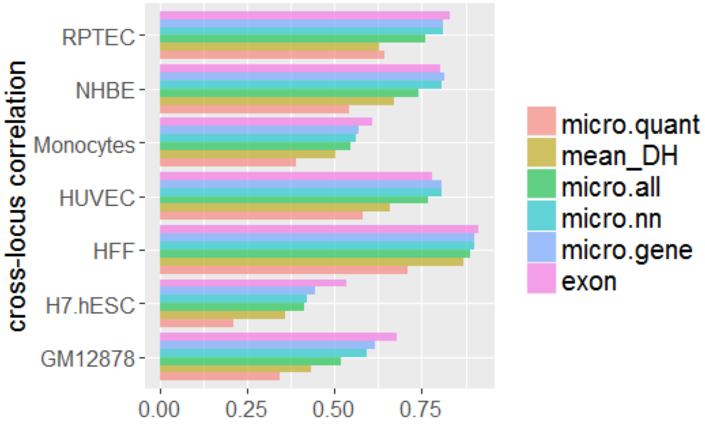


Figure: Cross-locus correlation

Results: cross-cell-type correlation

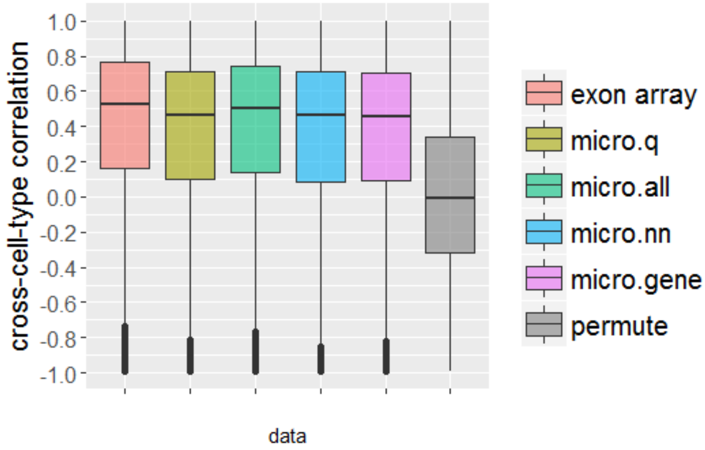


Figure: Cross-cell-type correlation

Results: prediction squared error

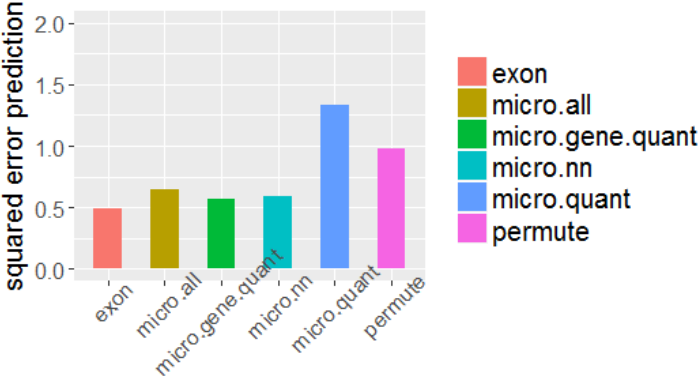


Figure: Prediction squared error

Content

1. Introducution
2. Methods
3. Results
4. Application
5. Conclusion

Application

- RNA-seq test
- Analysis of Pou5f1 binding sites

Application: RNA-seq test

- Goal: apply cross-platform prediction upon RNA sequence data
- Six samples which appear in both exon array and RNA-seq data are served as the gold standard, and the cross-cell type correlation is the evaluation metric.

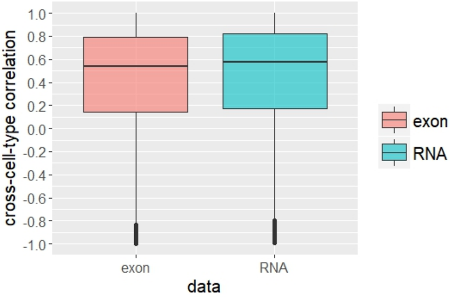


Figure: The cross-cell type correlation for exon array and RNA-seq data

Application: Pou5f1 binding sites

- Pou5f1 CHIP-seq peaks in H1-hesc
- Apply the cross-platform BIRD upon all public available GPL96 samples, and select the loci that
  - overlapped with Pou5f1 CHIP-seq peaks
  - high variability of the predicted value
- The ultimate data : 2490 DHS $\times$ 11865 samples

Application: Pou5f1 binding sites

Group both the DHS and samples into 10 clusters, and create a heatmap

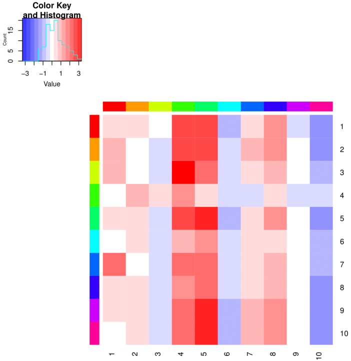


Figure: The heatmap of predicted DH level at Pou5f1 binding sites

Look up the sample annotation table

cluster id	samples
4	brain cortex hippocampus cerebellum
5	embryonic stem cells mesenchymal stem cells lung

Table: cluster id and samples

Content

1. Introducution
2. Methods
3. Results
4. Application
5. Conclusion

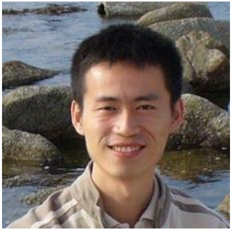
Conclusion

- The motivation of BIRD model is to predict chromatin accessibility using gene expression.
- Some normalization methods are proposed to deal with cross-platform prediction.
- Our method is further applied to some other examples.

Acknowledgement

This is joint work with Dr.Weiqiang Zhou and Dr.Hongkai Ji.

Weiqiang Zhou



Hongkai Ji

Reference I

Weiqiang Zhou, Ben Sherwood, Zhicheng Ji, Yingchao Xue, Fang Du, Jiawei Bai, Mingyao Ying, and Hongkai Ji.
Genome-wide prediction of dnase i hypersensitivity using gene expression.
Nature communications, 8(1):1038, 2017.

Cross-Platform Prediction of Regulatory Activities

Runzhe Li

Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

July 16, 2019