Amin Haghani, Assignment 3

BISC 481, Fall 2016 Third Assignment for 1 st Section of the Course (Rohs) Total 100 pts.

Modeling of protein-DNA binding specificity/Statistical machine learning

1. Application of an open-source and distributed revision control project: (a) Create a public repository with a README at GitHub https://github.com. (b) Write your name in the file of README.md. (c) Add a collaborator (username: TsuPeiChiu) to the repository. You are required to push your report and R scripts to the repository. The example and file template are shown in https://github.com/TsuPeiChiu/BISC481. 5 pts.

The account is opened as instructed.

1. High-throughput binding assays: Briefly describe (a) the in vitro experiments SELEX-seq and PBM, and (b) the in vivo experiment ChIP-seq. (c) Compare and discuss the advantage and disadvantage of these methods. 10 pts.

a) SELEC-seq: this is a method to make the complete repertoire of binding site preferences for transcriptional factor complexes, or DNA binding proteins. In this method, different transcriptional factors are attached to the plate and a library of DNA sequences will be added on top of the plate. The attached DNA will be sequenced to study the binding sites. In this method the number of attachment of a DNA sequence can represent the affinity of the bind with transcriptional factor.

Protein binding microarray (PBM): In this method, different DNA sequences are attached to a microarray plate. A library of proteins will be added to these sequences and the affinity can be measured from the amount of binding signal.

b) In vivo ChIP-seq: This is a qualitative data that analyze all the regions that attached to histons. ChIP-seq combines chromatin immunoprecipitation (ChIP) with massively parallel DNA sequencing to identify the binding sites of DNA-associated proteins.

c)

|  |  |  |
| --- | --- | --- |
| Method | Advantages | Disadvantages |
| In vitro | It will give quantitative data.  The data has high quality. | The data is not from a biological system. |
| In vivo | It shows the interactions in a biological environment. | The data is qualitative and it cannot quantify the affinity. |

1. Preparation of high-throughput in vitro data analysis: (a) Download and install R (version >= 3.3.0) from https://www.r-project.org. (b) Install Bioconductor on your R platform. The installation instruction can be found at https://www.bioconductor.org/install/. (c) Install package DNAshapeR on your R platform. The installation instruction can be found at https://www.bioconductor.org/packages/devel/bioc/html/DNAshapeR.html (d) Install the machine learning package caret on your R platform. The installation instruction can be found at https://github.com/topepo/caret (e) Download the gcPBM in vitro experimental data of Mad, Max and Myc from https://github.com/TsuPeiChiu/BISC481/tree/master/gcPBM. 10 pts.

All packages are installed as instructed.

1. Build prediction models for in vitro data: (a) Use the DNAshapeR package to generate a feature vector for “1-mer” sequence model and a feature vector for “1-mer+shape” model with respect to the datasets of Mad, Max and Myc. (b) Use the caret package to build L2-regularized MLR models for “1-mer” and “1-mer+shape” features with 10-fold cross validation, and print out the average R 2 (coefficient of determination) for these two models with respect to the Mad, Max and Myc datasets. 20 pts.

|  |  |  |  |
| --- | --- | --- | --- |
| Names | Rsquared | data | model |
| Mad seq | 0.775428384 | Mad | seq |
| Mad seq+shape | 0.862707644 | Mad | seq+shape |
| Max seq | 0.78570044 | Max | seq |
| Max seq+shape | 0.864134936 | Max | seq+shape |
| Myc seq | 0.778402483 | Myc | seq |
| Myc seq+shape | 0.854771678 | Myc | seq+shape |

1. High-throughput in vitro data analysis: (a) Draw a plot for the comparison of two different models (1mer vs. 1mer+shape) as shown in Figure 1B of Zhou et al. PNAS 2015. (b) Briefly discuss what you have learned from the results. 15 pts.



The plot is showing in all three data, including shape increased the R2 of three models. While the R2 in the models with only the sequence is around 0.77, adding the shape would increase the R2 to around 0.86.

1. Preparation of high-throughput in vivo data analysis: (a) Download the ChIP-seq data (including “bound” and “non-bound” data) of CTCF transcription factor of Mus musculus from https://github.com/TsuPeiChiu/BISC481/tree/master/CTCF. (b) Install the R packages mentioned in question (3). 5 pts.

The packages installed as instructed.

1. High-throughput in vivo data analysis: (a) Use plotShape() or heatShape() functions of DNAshapeR to generate ensemble plots for the DNA shape parameters of minor groove width (MGW), propeller twist (ProT), Roll, and helix twist (HelT) based on the sequences downloaded for question (6). (b) Briefly discuss what you have learned from the results. 15 pts.









The results show distinct DNA shape in bounded regions compared to unbounded regions. The bounded regions show specific pattern at the center of the sequence. On the contrary, the unbounded regions do not show specific changes in any of the analyzed factors.

1. Build prediction models for in vitro data: (a) Build logistic regression models for “1-mer” and “1- mer+shape” features, draw a plot of the ROC curves, and calculate the AUC score for each curve. (b) Briefly discuss what you have learned from the results. 20 pts.

|  |  |  |
| --- | --- | --- |
|  | Models | AUC |
| **1** | Sequence | 0.8406051 |
| **2** | Sequence+Shape | 0.8408944 |



The AUC scores of both models show that these logistic models can predict the response successfully. While, adding shape factors into the models slightly increased the AUC, shape parameters seem to had only small effects on this predictive model.