TB-TAD

Tyler Collins, Brian Palmer

Contents

Our updated project statement

Previously we decided to use an unsupervised learning algorithm to find topologically associated domains (TADs) in chromatin. This was going to be done by using epigenetic information which is commonly found around TADs to implicitly call tads at those regions. We have shifted to creating a supervised model by using some epigentic features to identify labeled data ie. tad boundaries. There is data out there already on GEO for tad boundaries. Alternatively, we can calculate TADs on our own using one of many published approaches (z score, insulation score, etc.). These algorithms take an intra chromosomal loci contact matrix as inputs, which we have available to us from GEO. Our approach will be to pass epigenetic data in as inputs use our labeled data to see if the model accurately predicts whether the location is a tad or not. To cross validate our data set, we will use epigenetic information from another cell line to see if it can predict tads or not.

Our collaboration strategy

Our strategy has been to keep a single space on our talapas (the HPC at U of O that we have been approved to use), and use github to work on the data processing and model building code in a distributed topology. We have a conda environment that we both use to keep our packages and workspaces consistent. Additionally, we have predefined variables pointing to shared resources so that that the code can symbolically reference locations and can be updated with ease if needed.

We generally have meetings over zoom and conversations on slack to stay in touch. Currently, we are meeting around 2-3 times per week. The responsibilities have been somewhat divided to:

- Brian - data processing and preparation
- Tyler - building the model

Although cross pollination is indeed occuring.

Our approach

Because we are basing our work on previously published research, we are relying on the following resources:

1. J. Chyr, Z. Zhang, X. Chen, X. Zhou, PredTAD: A machine learning framework that models 3D chromatin organization alterations leading to oncogene dysregulation in breast cancer cell lines. Computational and Structural Biotechnology Journal. 19, 2870–2880 (2021).

Github: jchyr-sbmi/PredTAD

We are using a subset and slightly revised version of the data processing and preparation scripts they have published here: * Processing: (1_gen_pre_info_08132020.r)[https://github.com/jchyr-sbmi/PredTAD/blob/master/Codes/1_gen_pre_info_08132020.r] * Preparation: (2_sample_preparation.R)[http
sbmi/PredTAD/blob/master/Codes/2_sample_preparation.R]

These scripts are helping us format the data to be passed into the model script, which we will base off of: * Model: (3_predTAD_model.R)[https://github.com/jchyr-sbmi/PredTAD/blob/master/Codes/3_predTAD_mod

The data

Essentially, our goal is to use the data used in the aforementioned paper stored on geo encode and process the data so that in fits in GenomicRanges objects. We are following the introduction found on the Bioconductor webpage to better understand these objects. Currently, our methylation data and tad boundary data is stored in these objects. We are able to view these objects using R studio by loading in the .Rdata file that is output at the end of the script. This .Rdata file is used in the data preparation step, which uses parameters store in the genomics ranges object to pool together data together rows of observations of methylation and tad data. The final .Rdata file will contain all features of interest. The data preparation step will additionally ensure that the observations have a strong significance value (P <= 0.05) where applicable.

Due to the size of the dataset (currently 2.4 GB), we cannot show the entire data set that we have. Here are some snippets though (specifically of the methylation data):

```
>
   load("C:\\Users\\14436\\projects\\ML\\TB-TAD\\Preprocessing\\pre_info_08132020.

> window(me)
GRanges object with 485577 ranges and 2 metadata columns:
           seqnames    ranges strand |        v1        p1
              <Rle> <IRanges>  <Rle> | <numeric> <numeric>
       [1]    chr16  53468112      * |   0.91828         0
       [2]     chr3  37459206      * |   0.94137         0
       [3]     chr3 171916037      * |   0.91857         0
       [4]     chr1  91194674      * |   0.06746         0
       [5]     chr8  42263294      * |   0.91514         0
       ...      ...       ...    ... .       ...       ...
  [485573]     chr6  68232042      * |   0.54641         0
  [485574]     chr3  14658866      * |   0.96195         0
  [485575]    chr15  45999823      * |   0.95774         0
  [485576]     chr2  12148220      * |   0.96256         0
  [485577]     chr3  86662155      * |   0.94185         0
  -------
  seqinfo: 24 sequences from an unspecified genome; no seqlengths
```

Initial results

We are currently working through data collection. We found out that some of the GEO IDs were not reported at the end of the paper and had to be retreived from other papers. We feel confident that we will complete data collection soon. A larger concern we have is properly formatting the data. While preprocessing the methylation data went smoothly, we are finding the TAD data is failing to convert to the GenomicRanges object. We have several data examples so our hope is a combination of different data sets and debugging the r script will elucidate answers.

Current reservations

Our largest concern is data preparation. While we have a script to work off of, we have found that:

1. Data retrieval has not been clear cut
2. Data preparation has had a stumbling block (likely because of the data)

We are pushing through and think that we can have a successful outcome. We are excited to exposure to GenomicRanges objects, as these are used in our field of study (bioinformatics) here at the University of Oregon.

Internal: To compile for submission

```
pandoc README.md --listings -H listings-setup.tex --toc -V geometry:"left=1cm, top=1cm, right=1cm, bottom=2cm" -V fontsize=12pt --pdf-engine=xelatex -f markdown -t pdf -o tyler_brian_ML_progress.pdf --metadata title="TB-TAD" --metadata author="Tyler Collins, Brian Palmer"
```