Finding enhancer-promoter interactions from Hi-C data

Dror Moran, Gil Ron and Tommy Kaplan
The Rachel and Selim Benin School of Computer Science and Engineering, Hebrew University, Jerusalem, Israel

1. Enhancer-promoter interactions

Hi-C
- Count chromatin interaction
- Repeated million of time
- Hi-C Heatmap(i,j) = #interactions between blocks i and j
- #interactions decrease as the distance from the diagonal increase.

TAD
- Strong self-interaction

Enhancers
- Short (50-1500 bp) DNA regulatory element
- Enhancer-promoter interaction are hard to find
- #Enhancer-promoter interactions > #Expected interactions

2. Build Hi-C model

Partition into TAD:
- Two TAD fold to bigger TAD
- Different cells have different TAD

We use TAD hierarchy:

Predict TAD interactions

3. Hi-C model Comparison

- Six Hi-C model where tested:
  Bi-Linear, linear, Hierarchy, No-hierarchy, Dixon, Gil
- RMSE – Distance function between a model prediction and the values actually observed:

\[
\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2}
\]

4. Residual Matrix

- Matrix of the difference between the original Hi-C and the model
- Containing 2M bp before and after gene promoter
- Most of the interactions are not promoter-enhancer interaction. So, this is a normal distribution of residuals for non enhancer-promoter interactions

p-Value
- \( p(x) = P(X \geq x | H_0) \)
- If \( p(x) \) is small, \( x \in H_1 \)
- Calculate p-value for the gene promoter on a window of size 2M bp from each side

5. Foxg1

- Residual matrix of mouse cortex
- Foxg1 gene express mainly in the brain
- Found two over-represented interactions
- Compatible to different enhancer detectors

6. Pol2

- List of 23,000 genes
- Normal distribution for every gene
- Calculate p-value of the promoter on every index closer then 2M bp
- Mean pol2 per block of size 10,000

7. Summary

1. Gil’s hierarchy and bi-linear regression, give better fit for the data.
2. We can use Hi-C data for finding promoter-enhancer interactions using the residual matrix