Activity-by-Contact Model to Predict Enhancer-Gene Connections

Overview

Genes are expressed at different levels in every cell, but there is currently a limited understanding of what causes differences in gene expression. A leading theory is that genes are activated by enhancers located in open chromatin regions near the gene (Figure 1). The larger the enhancer, the higher the expression of the gene. However, testing this theory is complicated because multiple enhancers may control one gene, a single enhancer may control many genes, and connections can span large genomic distances (Figure 2).

When a mutation occurs in the genome, it may change the folding of the DNA, change the size of enhancers, or relocate genes. Over or under-expression of a gene can lead to uncontrolled cell growth and loss of function. It is vital to predict gene expression and identify enhancers-gene connections to form a better understanding of the activation of oncogenes, identify transcription factor binding sites, and possibly identify kinases for drug targets.

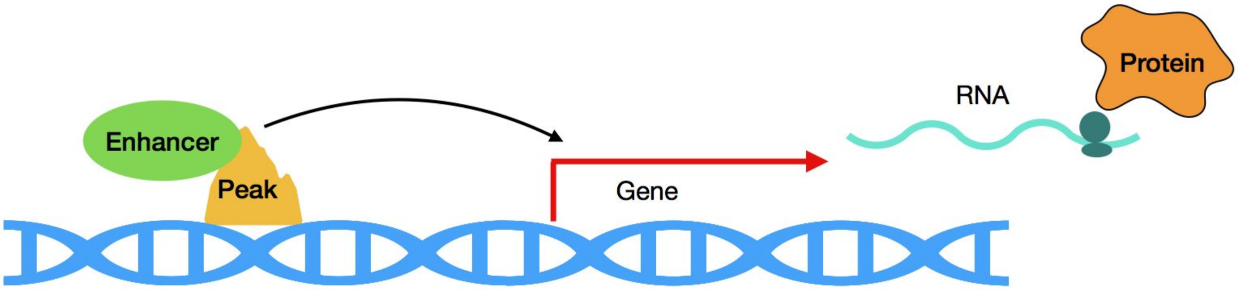

Figure 1. Enhancers bind to the genome at open chromatin regions. They activate nearby genes by recruiting general transcription factors (GTFs) and RNA polymerase II. The gene is subsequently transcribed into mRNA and translated into protein, which carries out functions in the cell.

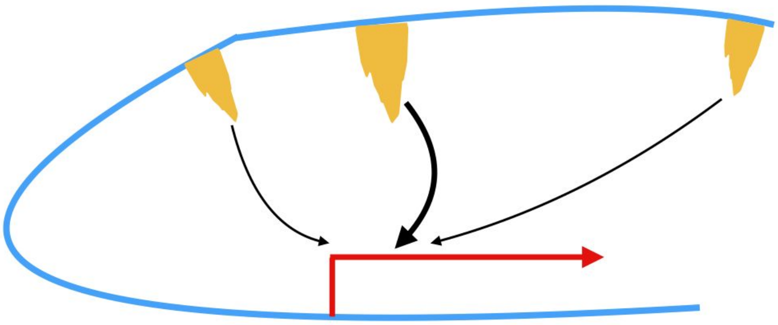

Figure 2. A single gene (red arrow) may be influenced by multiple enhancers (yellow) by varying amounts. Enhancer-gene connections may also span large genomic distances and a single enhancer may control multiple genes, making these connections difficult to predict.

The goal is to create a model of gene activation and predict enhancer-gene connections based on enhancer activity and the 3D structure of the genome. This Activity-by-Contact model is defined as:

$$\text{ABC score}_{E\text{-}G} = \frac{A_E \times C_{E\text{-}G}}{\sum_{e \text{ within } 5 \text{ Mb}} A_e \times C_{e\text{-}G}}$$

Where A_E is the activity of the enhancer and $C_{E\text{-}G}$ is the contact between the enhancer and the gene (Fulco et al 2019). Enhancer activity is defined as the geometric mean between ATAC-seq data and H3K27ac CHiP-seq data of the open chromatin peak, and Contact is the contact frequency measured by HiC. The gene expression is given by RNA-seq data.

The goal of this model is to predict enhancer-gene connections across the genome. After validation, the model will be used to analyze mutations and oncogenes in leukemia patients.

The model was first validated using sequencing data from the cell line K562, a human myelogenous leukemia cell line. The connection between peaks and genes was obtained from Gasperini et al (2019), H3K27ac and RNA-seq data for K562 was obtained from ENCODE, and ATAC-seq data for K562 and general HiC data was sequenced at the McVicker Lab at the Salk Institute for Biological Sciences. After standard processing and normalization steps of each of these indices, I calculated the ABC score for known peak-gene connections, known negatives, and unknown connections. A precision/recall curve was then generated to test how well the model can differentiate connections.

After validation, the model will be applied to genetic data from 16 B ALL Leukemia patients at a hospital in San Diego, California. ATAC-seq, HiC, and RNA-seq data were obtained for each patient and sequenced in the McVicker Lab at Salk. A general H3K27ac CHiP-seq dataset for pre-B cells will be used. The ABC model will then be used to study probable peak-gene connections and analyze how mutated enhancers affect the expression of oncogenes.

All analysis and data processing will be done in python, R, and shell script.