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

Interim Report
Team #20, Los Alamos High School

Genes are expressed at different levels in each cell. Gene activity determines the proteins that the cell makes, which consequently controls the cell type and every function in the cell. There is currently very little understanding of what causes differences in gene expression. A leading theory is that genes are activated by enhancers located in open chromatin regions physically near the gene (Figure 1) (Pennacchio et al.). 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). Right now, most researchers use raw genomic distance to predict enhancer-gene connections.

When a mutation occurs in the genome, it may change the folding of the DNA, change the size of enhancers, or relocate genes. After a mutation, enhancers may change is size or control genes they are not supposed to. Overexpresson of a gene can lead to uncontrolled cell growth and cancer. 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.

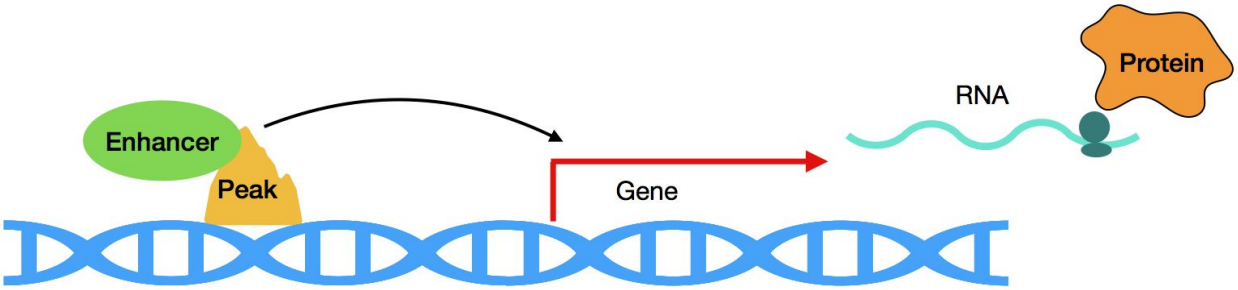

Figure 1. Enhancers bind to the genome at open chromatin regions. They activate nearby genes by recruiting general transcription factors (GTFs) and RNA polymerase II. The gene is subsequently transcribed into mRNA and translated into protein, which carries out functions in the cell.

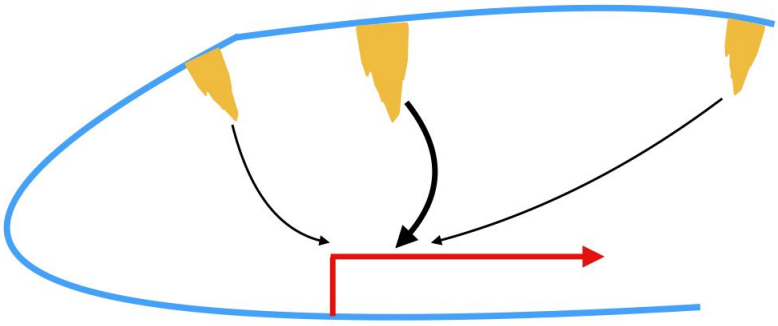

Figure 2. A single gene (red arrow) may be influenced by multiple enhancers (yellow) by varying amounts. Enhancer-gene connections may also span large genomic distances and a single enhancer may control multiple genes, making these connections difficult to predict.

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:

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

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 to train the model on is given by RNA-seq data. In an alternate definition of ABC, the activity of the enhancer is simply given by the ATAC, referred to here as "ABC (no H3K27ac)".

The model was first validated using sequencing data from the cell line K562, a human myelogenous leukemia cell line. Validation data of 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 obtained from the McVicker Lab at the Salk Institute for Biological Sciences. After standard processing and normalization steps of each of these indices, I calculated the two ABC score for peak-gene connections and the genomic distance between the peak and gene. Only peaks and genes within one megabase were considered. Figures 3 and 4 display the performance differences between ABC, ABC (no H3K27ac), and distance.