Learning the topology of the genome from protein-DNA interactions


**Introduction**

A central problem in genetics is how the genome (which measures 2 meters from end-to-end when stretched out) fits inside the nucleus of a cell that has a diameter on the order of microns wide). We recently systematically mapped a number of structural features organizing the genome in three dimensions, including the positions of chromatin loops, genome-wide. However, mapping loops using experiments that measure DNA-DNA interactions is cost-prohibitive relative to the cost of measuring protein-DNA interactions. Here, we seek to learn and predict the three dimensional structure of the genome using protein-DNA interaction data.

**Datasets**

Gold standard chromatin loop annotations for the GM12878 cell line (a human B-lymphoblastoid cell line) were obtained from Rao and Huntley, et al. [1]. By combining these annotations with CTCF, RAD21 and SMC3 ChIP-Seq data from ENCODE [2], we identified 6987 high-confidence loop anchors (down to motif resolution) as well as 32,511 CTCF sites that do not participate in loop interactions [1,3].

For features, we downloaded ChIP-Seq data for 87 transcription factors from ENCODE [2]. By combining these annotations with CTCF, RAD21 and SMC3 ChIP-Seq data, we identified 6987 high-confidence loop anchors (down to motif resolution) as well as 32,511 CTCF sites that do not participate in loop interactions [1,3].

**Preliminary tests to predict chromatin loop anchors**

We first attempted to predict which CTCF binding sites were likely to be loop anchors using a number of different strategies [4]. In general, we found that logistic regression and support vector machines performed far better than a Naïve Bayes classifier. Our initial training data (blue columns) made an assumption about which examples were actually negative that we later realized was incorrect; when we corrected this in our training data (pink columns), the performance of all of our models improved.

**Grid Search**

To optimize the hyperparameters for our support vector classifier, we performed a parameter sweep through a range of values for the penalty parameter, C, and the kernel coefficient, γ.

**Feature Selection**

For features, we downloaded ChIP-Seq data for 87 transcription factors from ENCODE [2]. By combining these annotations with CTCF, RAD21 and SMC3 ChIP-Seq data, we identified 6987 high-confidence loop anchors (down to motif resolution) as well as 32,511 CTCF sites that do not participate in loop interactions [1,3].

**Predicting chromatin loops**

We are currently trying to use the predictions of our SVM to identify loop anchors to pair anchors together and predict the actual chromatin loops themselves.

In brief, we are operating under the assumption that loops form via a process of extrusion. By using the probabilities that a CTCF site is a loop anchor, we are trying to predict pairs of loop anchors that form a loop by using as features the probabilities that the two sites anchors as well as the probabilities of all the sites in between the two being anchors and how many intervening sites there are.

This work is in progress. There does not seem to be a major difference in loop anchor probability (distance to hyperplane) for loop anchors that are themselves contained within a loop or not, indicating that likely multiple consecutive loop anchors are needed to completely halt the extrusion complex.