Learning the topology of the genome from protein-DNA interactions

Suhas S.P. Rao – CS229 Final Project – 12/8/2015

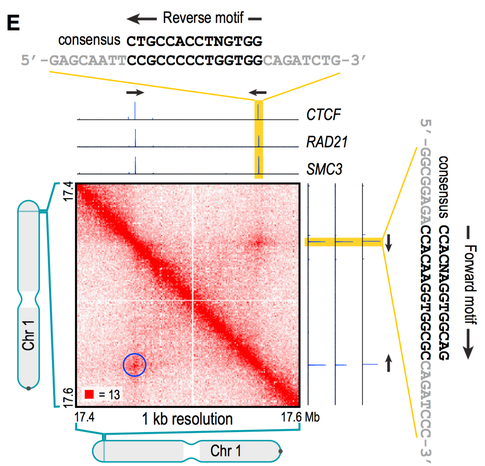
Introduction

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 and histone modifications as well as DNAse accessibility data from ENCODE. We also included features measuring CTCF motif strength, nucleosome occupancy around the site, and evolutionary conservation.



Legend: An example of what a chromatin loop looks like in a DNA-DNA interaction dataset (heatmap); the circled focal peak represents a chromatin loop. Protein-DNA interaction tracks shown for CTCF and cohesin above and to the right of the heatmap, illustrate the relationship between protein binding and 3D genome structure, and thus the potential to predict DNA-DNA interactions using only protein binding and sequence information. (reproduced from [1])

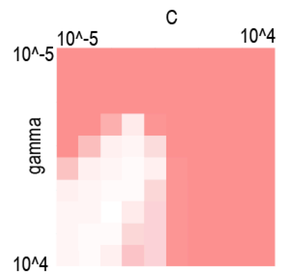
Preliminary tests to predict chromatin loop anchors

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.

Model	TrE	TE	S	P	TrE	TE	S	P
NB	0.249	0.251	0.955	0.349	0.227	0.233	0.941	0.430
LR	0.095	0.095	0.628	0.665	0.097	0.098	0.740	0.718
SVM (L)	0.097	0.097	0.635	0.694	0.101	0.101	0.815	0.681
SVM (P)	0.092	0.093	0.629	0.670	0.083	0.094	0.749	0.730
SVM (RBF)	0.095	0.094	0.602	0.675	0.089	0.093	0.763	0.729

Legend: NB: Naïve Bayes; LR: Logistic Regression (L2 regularization); SVM (L): SVM with linear kernel and C=1; SVM (P): SVM with polynomial kernel (degree 3) and C=1; SVM (RBF): SVM with RBF kernel and C=1. Blue: original training data; Pink: corrected training data

Grid Search



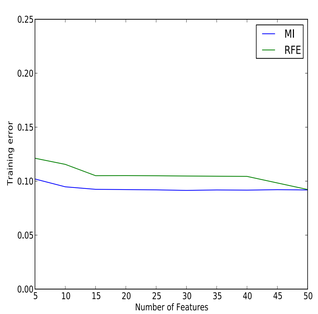
To optimize the hyperparameters for our support vector classifier, we performed a parameter sweep through a range of values for the penalty parameter, C, and the kernel coefficient, γ.

	Positive (True)	Negative (True)
Positive (Predict)	1643	594
Negative (predict)	495	9287

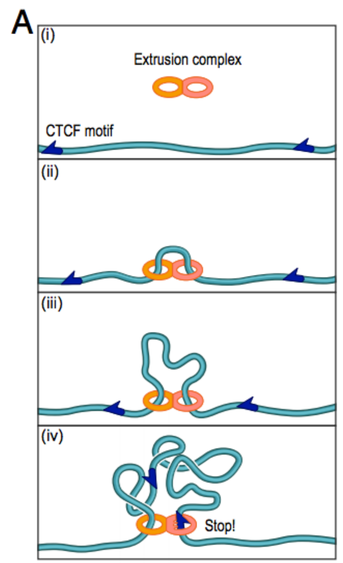
Feature Selection

Table 1: Feature Rankings

Ranking	Feature (MI score)	Feature (RFE))
1	SMC3 BS	SMC3 peak
2	RAD21 BS	CTCF peak
3	CTCF BS	SMC3 BS
4	SMC3 peak	PBX peak
5	RAD21 peak	BRCA1 peak
6	ZNF143 peak	EBF1 peak
7	conservation	POLR3G peak
8	motif 1 strength)	NFYA peak
9	DNase peak	H4K20me1 peak
10	motif 2 strength	RAD21 BS



Predicting chromatin loops



Legend: An illustration of the extrusion model for chromatin loop formation. A complex binds to two places on DNA that are very close together and then slides in opposite directions until it hits two correctly oriented CTCF sites that act as brake points, forming a loop. Reproduced from [3].

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 are anchors as well as a 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.

1. Rao, Suhas S.P., Huntley, Miriam H., et al. (2014). "A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping." Cell 159, 1665-1680.
2. ENCODE Project Consortium. (2012). "An integrated encyclopedia of DNA elements in the human genome." Nature 489, 57-74.
3. Sanborn, Adrian L., Rao, Suhas S.P., et al. (2015). "Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes." PNAS 112, E6456-6465.
4. Scikit-learn: Machine learning in Python." Journal of Machine Learning Research 12, 2825-2830.