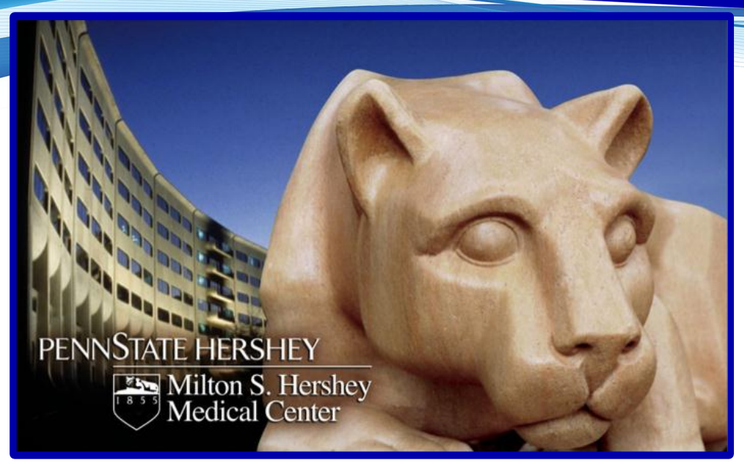

A Supervised Learning Approach to the Prediction of Hi-C Data

Tyler Derr, Yanli Wang, Feng Yue

The Department of Biochemistry & Molecular Biology, Pennsylvania State University College of Medicine, Hershey, PA 17033

Introduction

The state of the art method for studying genome-wide chromatin structures is Hi-C, which is a high-throughput chromosome conformation capture (3C) based technology. However, due to the cost of performing Hi-C experiments, only a small subset of all possible species, tissue, and cell type data are currently available. We propose a supervised learning method for the prediction of the entire intra-chromosomal Hi-C interaction data using a Random Forest (RF) approach.

Methods

The RF is trained using a known Hi-C matrix and a set of organized features for each region in the chromosome, which includes the GC content, mappability, number of restriction enzyme cut sites, histone modifications (H3K4me3, H3K36me3, etc.), and transcription factor binding sites (Pol2, Ctcf, etc.), which are freely available thanks to the recent efforts of the ENCODE and Roadmap Epigenomics projects, for each region. Figure 1 shows the stages of our current implementation.

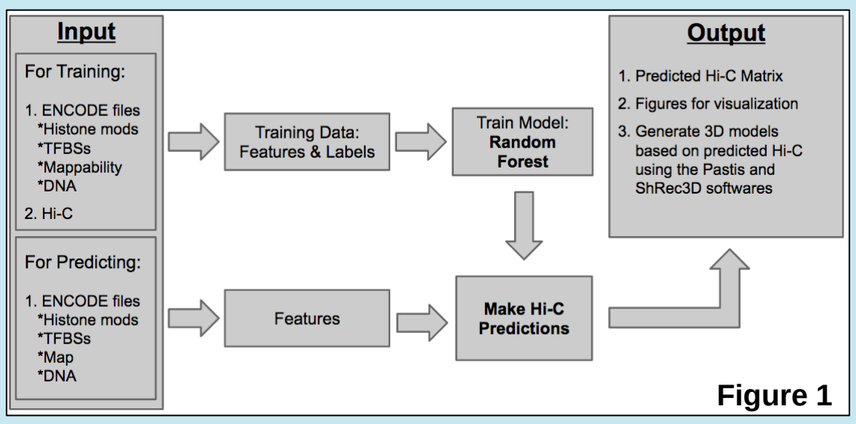

Figure 1

Figure 2 shows an example of how the training instance $i_{\alpha\beta}$ is constructed for the interaction between regions α and β. $i_{\alpha\beta}$ consists of a feature set, F, which consists of $(2k+1)$ features, where k is the number of features we have for each region, and the plus one comes from $\delta_{\alpha\beta}$, which is the genomic base pair distance between the two regions. The label, L, is the Hi-C interaction frequency value associated with the two regions, α and β.

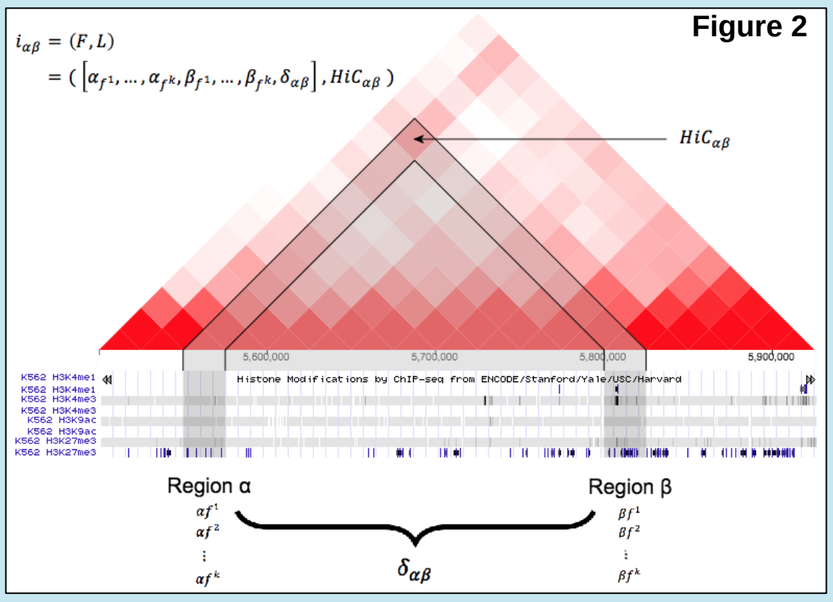

Figure 2

Results

The heatmaps in Figure 3, which were generated using the 3D Genome Browser that was developed in YUE Lab, visually shows that the topologically associating domains (TADs) have been preserved during the prediction process.

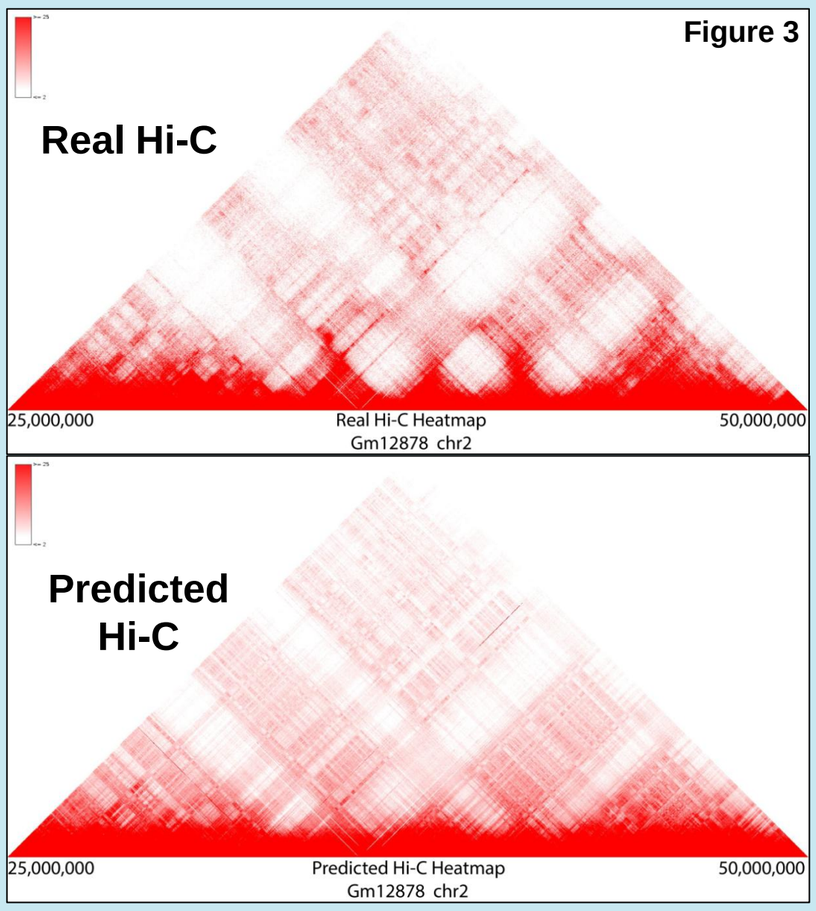

Figure 3

Our method has shown in Figure 4 that the distribution of the predicted Hi-C data is almost identical when predicting in the same cell type.

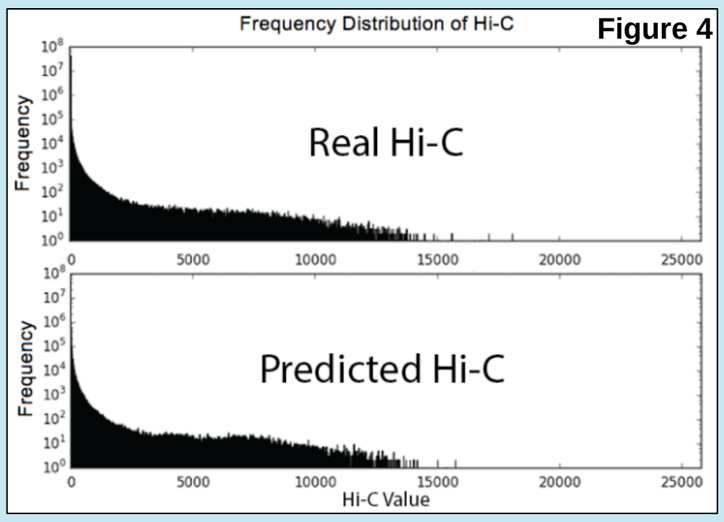

Figure 4

In Table 1 are the statistical values for predictions made in A) The same cell type (Gm12878), B) The same species, but different cell type (Gm12878 and K562), and C) Different species (Gm12878 and Ch12)

Table 1	A	B	C
Pearson	0.98	0.96	0.94
Spearman	0.63	0.45	0.33
r^2	0.96	0.92	0.87
rmse	22.2	5.2	11.6

After the RF model has been trained we are able to determine a ranking of which features are more important relative to each other. In Figure 5 we show the differences in these rankings from mm9 mesc predictions that were made using 40kb resolution.

Results

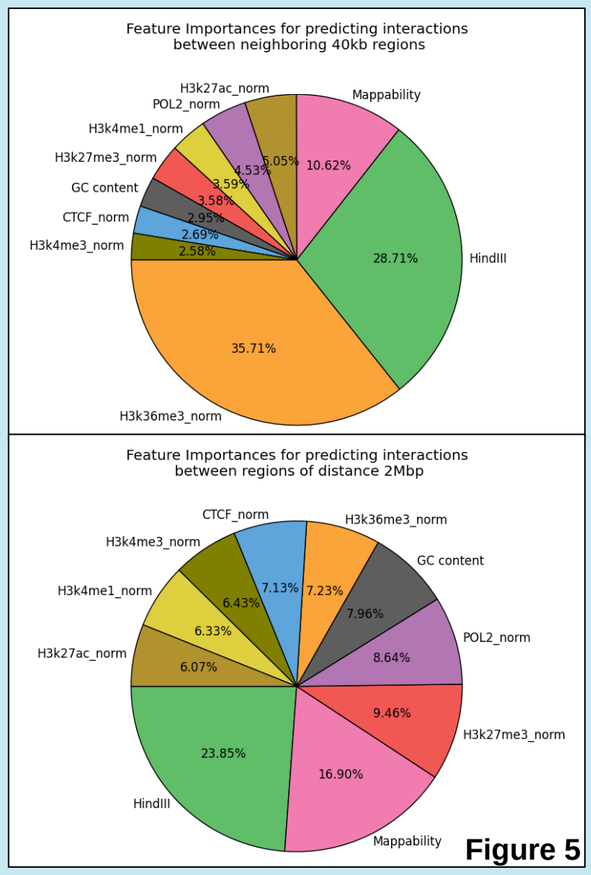

Figure 5

In Figure 6 are the 3D predictions returned by the MDS method found in Pastis[1].

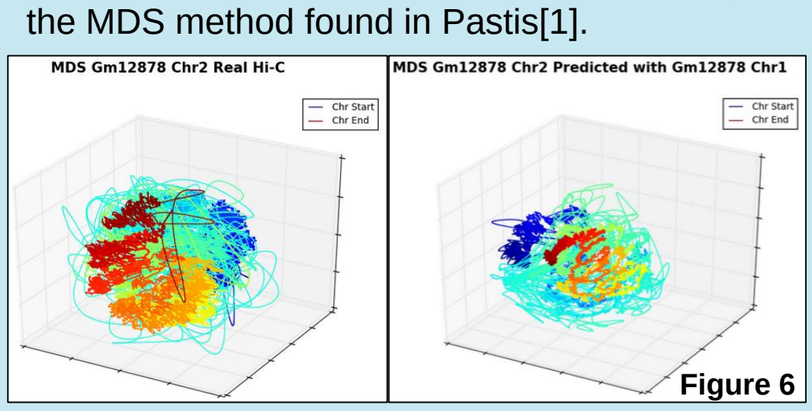

Figure 6

Discussion

Our model has shown to be quite robust in its predictions and further research into this idea of predicting Hi-C we believe will be very fruitful. An analysis on the feature importances returned by the RF has allowed us to determine which of the features are more meaningful for the interactions between regions at varying distances and as hypothesized, they are different for longer interactions.

Conclusions

We have devised a method for the prediction of Hi-C using ENCODE data. The predictions have shown to not only perform well for predictions in the same cell type, but also across cell types, and even across species. The RF has provided insight into what are the more meaningful/causal features for the interactions between regions of varying distances.

References

1. Varoquaux, N. *et. al.* A statistical approach for inferring the 3D structure of the genome. *Bioinformatics*, 30(12):i26-i33, 2014.

PENNSTATE HERSHEY College of Medicine 1855