Cheap Map: Hi-C from ChIP-Seq through Machine Learning

Arya Haji Taheri[a,b]

Ryan R. Cheng[a], Margaret S. Cheung[a,b], Vinícius G. Contessoto[a], Michele Di Pierro[a], José Onuchic[a]

[a]Center for Theoretical Biological Physics - CTBP - BRC, Rice University, Houston, TX, USA
[b]University of Houston, Houston, TX, USA

Summary

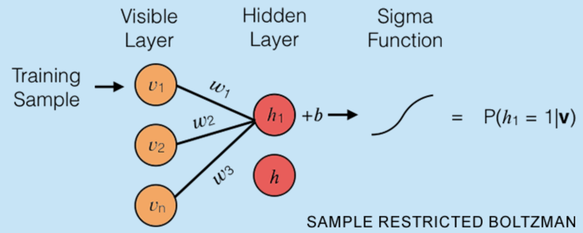
We present a software tool that predicts DNA-DNA ligation data from Chromatin ImmunoPrecipitation-sequencing data.

Abstract

In the nucleus of eukaryotic cells, chromatin is organized in specific conformations which depend on cell type; these conformations have been most recently studied through DNA-DNA ligation assays. Obtaining these data about the three-dimensional structure of the chromosome is an expensive and time consuming process. We exploit the idea that epigenetic data determine chromatin architecture[2] by developing a tool that can predict the chromatin contact map purely from epigenetic data without using any structural information. The model has been trained on chromosome 1 and 2 and can quickly predict the high-resolution contact maps (Hi-C) of chromosome 1-22 for the human lymphoblastoid cell line.

Method

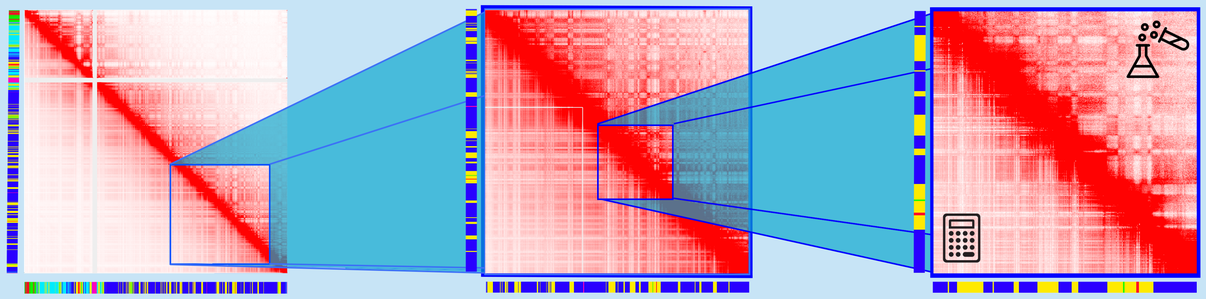
We built a restricted Boltzmann Machine to predict the probability of a contact forming between locus i and locus j given the epigenetic marking patterns at the same loci.

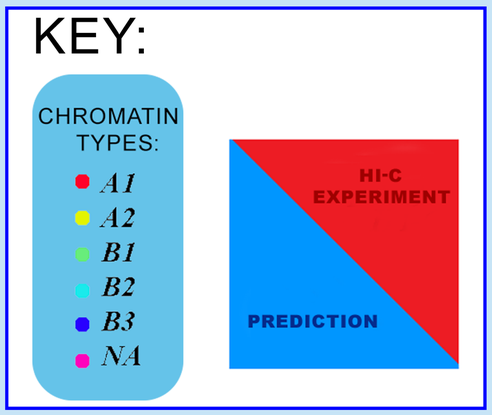


The network is trained on a dataset composed of the ChIP-Seq tracks from the two largest chromosomes in GM12878 cell line (1 and 2) together with their respective contact probabilities. Subsequently, we predict the contact probability maps for a test set composed of all the remaining chromosomes from epigenetic marks alone.

After the initial training, this network can predict the contact maps of all chromosomes of lymphoblastoid cells within 20 minutes. We hope to improve the accuracy of the predicted maps by using larger training data sets and introducing more sophisticated optimization methods.

Chromosome 4 predicted map





The panels above show the contact probability map of chromosome 4 of lymphoblastoid cells. The map obtained from our network is shown in the lower diagonal section while the upper diagonal region shows the experimental Hi-C maps. At this level, it is clearly visible that our network reproduces the patterns in Hi-C maps and resembles the chromatin type sequences.

Accuracy

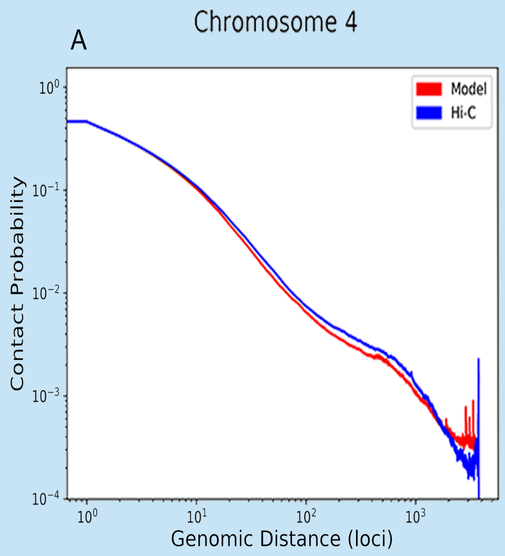


Figure A illustrates the probability of contacts as a function of genomic distance in both experimental and predicted maps. Our network accurately predicts the contact probability as the genomic distance increases.

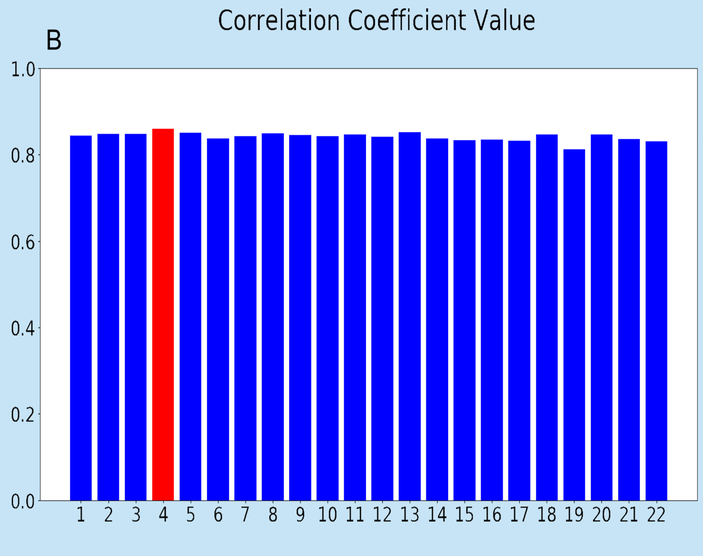
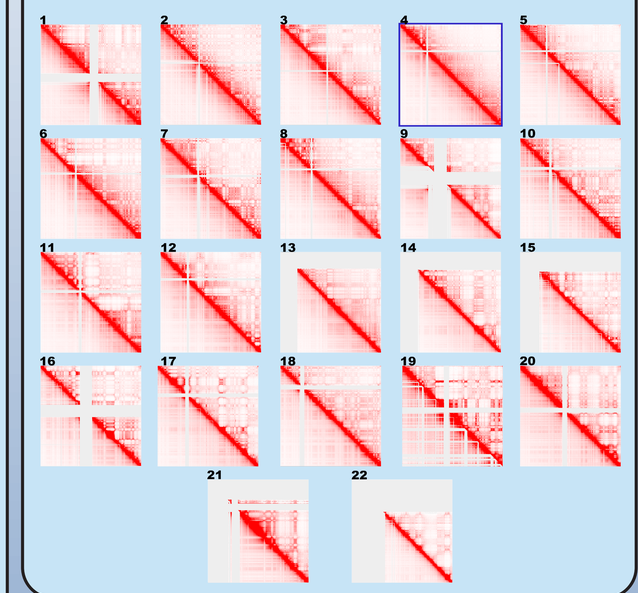


Figure B shows the Pearson's correlation coefficient between the contact probabilities generated by the network and measured by Hi-C for all the 22 chromosomes. The average correlation is 0.851. For chromosome 4, which is visualized above, the correlation is 0.867.

Contact Map Prediction



Conclusion

The cost of developing Hi-C maps have fallen drastically in recent years; however, the process is still lengthy and expensive. Additionally, for certain cell types, obtaining Hi-C data remains challenging. ChIP-Seq requires fewer cells to produce reliable data, and it is readily available through many public databases. We extend earlier work[1] by demonstrating that the structure of chromosomes can be predicted, de novo, purely from ChIP-Seq. The proposed model can predict Hi-C maps within minutes.

Acknowledgments

NSF PHY-1427654
NSF ACI-1531814

Bibliography

1. Di Pierro, M. *et al.* *Transferable model for chromosome architecture.* , 1613607113, (2016).

2. Di Pierro, M. *et al.* *De novo prediction of human chromosome structures: Epigenetic marking patterns encode genome architecture.* , 1714980114, (2017).