From 3C to 3D: Constructing 3D models of Murine Olfactory Receptor gene related interactions from Hi-C 3C data using Multidimensional Scaling and Clustering Algorithms

Siddhant Sanghi COMS 4762 Final Project
Columbia University
New York, NY 10027
ss5943@columbia.edu

Abstract

The murine olfactory three-dimensional nuclear architecture is essential in order to understand olfactory receptor (OR) choice: in particular, the LDB1 adaptor protein is claimed to control OR choice through trans interactions[1]. This paper intends to use Hi-C data from wild-type (WT) and knockout (KO) mice for the LDB1 gene, in order to construct 3D models for murine nuclear architecture specific to olfaction using Multidimensional scaling, and compare the differences in model architecture between WT and KO mice using k-means clustering on the Multidimensionally scaled 3D data with the goal of understanding the role of LDB1 in changing three-dimensional chromatin configuration for OR expression. Results from this paper align with previous papers, to demonstrate drastic 3D conformational changes due to LDB1 protein transcribed from the LDB1 gene.

1 Introduction

My project aims to use Multidimensional Scaling Algorithms (MDS) [2] in order to create 3D models by reconstructing genomic spatial structure from Hi-C Contact matrices, to capture all OR gene related functional interactions using the Olfactory Marker Protein (OMP)-sorted bulk Hi-C datasets from WT and KO mice for genes of critical adaptor protein LDB1 relevant to murine nuclear architecture, as discussed by Monahan[1].

I wish to specifically understand how 3D spatial genome models of murine OR gene related functional interactions (constructed from Hi-C contact matrices by extending an existing MDS) aid in the investigation into the differences in WT vs. KO mice for LDB1 gene. This investigation will help quantify the significance of this gene to murine nuclear architecture relevant to olfaction.

Building my 3D models on Hi-C datasets for both WT and KO mice is a worthwhile endeavor that aims at carefully investigating how OR gene nuclear architecture changes with removal of said proteins in order to see their functional relevance to the murine OR gene nuclear architecture. This project is one of its kind, in that it builds on previous papers to chart out uncharted paths in 3D modelling of murine nuclear architecture relevant to olfaction.

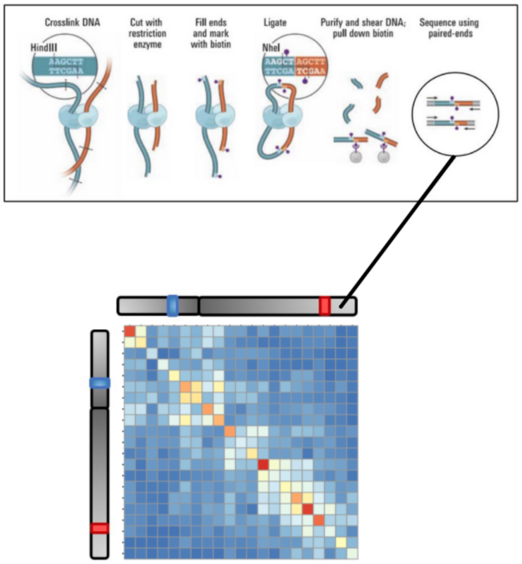
Crosslink DNA
HindIII
AAGCTT
TTCGAA
Cut with restriction enzyme
Fill ends and mark with biotin
Ligate
NheI
AAGCT AGCTT
TTCGA TCGAA
Purify and shear DNA; pull down biotin
Sequence using paired-ends

Figure 1: Hi-C process[1] that can be represented by contact matrix

I will be implementing Classical MDS[1], which is a dimensionality reduction strategy popularly used in optimization problems, in order to reconstruct 3D spatial genomic models from contact matrices processed from Hi-C Data. After using the MDS algorithm, I will run my implementation of k-means clustering algorithm to better understand the spatial organisation of my 3D models in order to compare WT vs KO mouse nuclear architectures more specifically.

I will be using publicly available Hi-C data on (`https://data.4dnucleome.org/`) under the following accession numbers: 4DNES18BMU79 and 4DNESEPDL6KY).[2]

2 Methods

Multidimensional scaling (MDS)[2] refers to a group of techniques that, given a set of noisy distances, find the best fitting point conformation.[2] MDS makes up a class of unsupervised estimators that seeks to describe datasets as low-dimensional manifolds embedded in high-dimensional spaces.[2] Many cost functions can be optimised, but the classical MDS method minimizes the Frobenius norm of the difference between the input Euclidean Distance Matrix (EDM) and the Gram matrix of the points in the target embedding dimension k [2](in our case k = 3 for three dimensions). In this way, MDS carries out manifold-learning.

We get the EDM from the approximated distance matrix which is derived from the Hi-C contact matrix extracted from the Hi-C data (.hic files from 4DNucleome).

The approximated distance matrix is simply an element-wise transformation of the Hi-C contact matrix, using the fact that there is an inverse relation between contact frequency and spatial distance[4], i.e., the more the distance between regions on the DNA, the higher the frequency of contact.

$D_{ij} = \frac{1}{F_{ij}^{\alpha}}$ ...(1)

In the above equation[4], α is the conversion factor, which in my case I have chosen as 0.5. The optimum range for alpha is [0.1, 2][1], because any lower or higher and the inverse trend will be underrepresented or overrepresented.

Now, to get the EDM from the approximated distance matrix, we need to take pairwise euclidean distances from elements in the approximated distance matrix, representing the spacing of points in Euclidean space.This EDM now is the input for our MDS algorithm.

After using the MDS algorithm, we can run k-means clustering to better understand the spatial organisation of my 3D models in order to compare WT vs KO models to understand the significance of LDB1 protein.

3 Results

I have written code `https://github.com/Sid01123/HiC.git` in the R programming language, that extracts Hi-C matrix information from my downloaded datasets (from 4Dnucleome), then applies element-wise transformation to get the approximated distance matrix, and finally converts this matrix into the EDM using the dist function, which is then the input my myCMDScale function, which runs the classical MDS algorithm to get a 3D model.

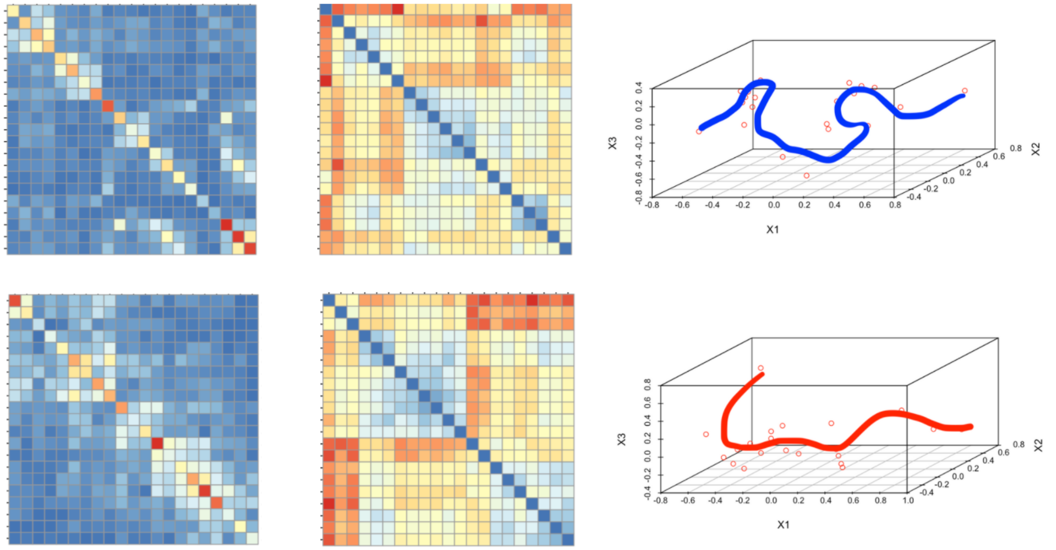

Figure 2: Hi-C (3C) heatmap → EDM heatmap → 3D model for intrachromosomal interactions between bases 45700000-46700000 in chromosome 19 of WT mouse (above) and LDB1-KO mouse (below)

Figure 2 shows the Hi-C (3C) heatmap → EDM heatmap → 3D model transformation for intrachromosomal interactions between the same region (bases 45700000-46700000) in chromosome 19 of both WT and KO mice. I have chosen this section to compare between the two mice, because this region is the location[3] of the LDB1 gene that if transcribed, is translated into LDB1 adaptor protein, necessary for trans interactions that guide Olfactory receptor choice. I have used the heatmap function in R to visualise the matrices handling any NA/NAN/INF values by replacing them with 0, and the scatterplot3d package in R for visualising the 3D model of nuclear genomic architecture. The two 3D models (both very differently oriented) show that the conformational looping facilitates transcription of the LDB1 gene in WT, but the absence of looping in KO causes no transcription of LDB1 gene.

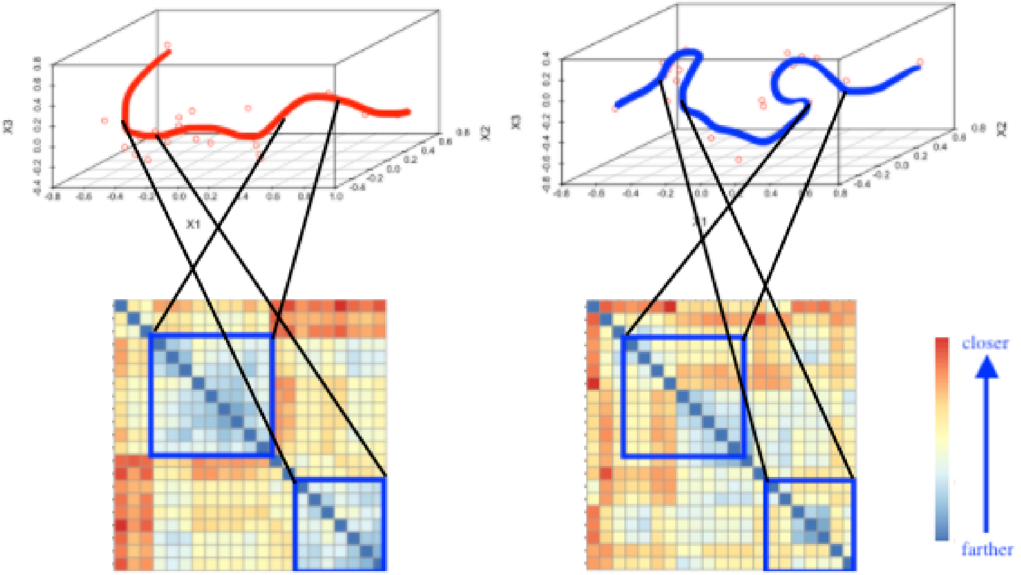

Figure 3: Mapping Euclidean Distance Matrix onto 3D model for KO (left) and WT (right) in order to zoom in on the drastic conformational differences between the 3D models

Figure 3 clearly zooms in on the conformational differences between the two 3D models, mapped directly from the Euclidean distance matrix. These difference conformations are the difference between expression and non-expression of LDB1 gene, which in turn cause trans conformational changes that affect Olfactory receptor choice, as seen in Figure 4[5].

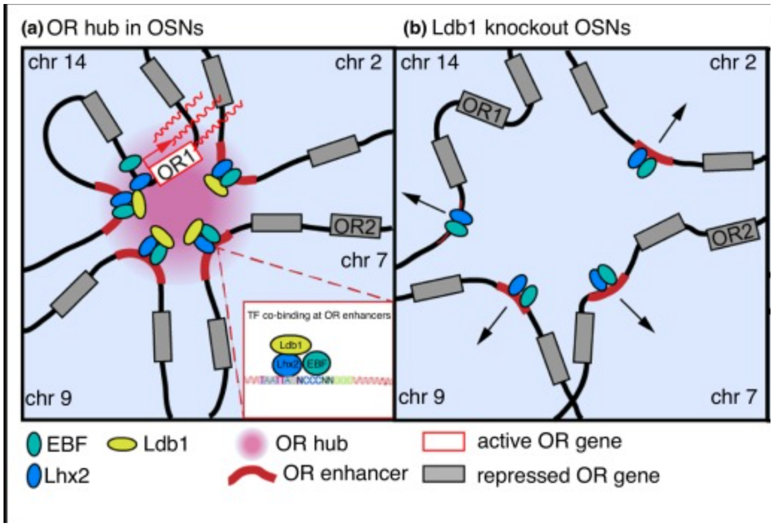

Figure 4: Trans interactions controlled by LDB1 protein[5]

Finally, I implemented myKMeans, my k-means clustering function that flattened my 3D points into a 2D plot, clustering spatially close domains. I varied my values for k, and ran random restarts, in order to get the best number of clusters that described my 3D data. I got k=4 clusters for WT and k=2 clusters for KO, for the same DNA region (chromosome 19, bases 45700000-46700000). Figure 5 demonstrates the mapping from my clustering plots to my 3D models for both WT and KO mice.

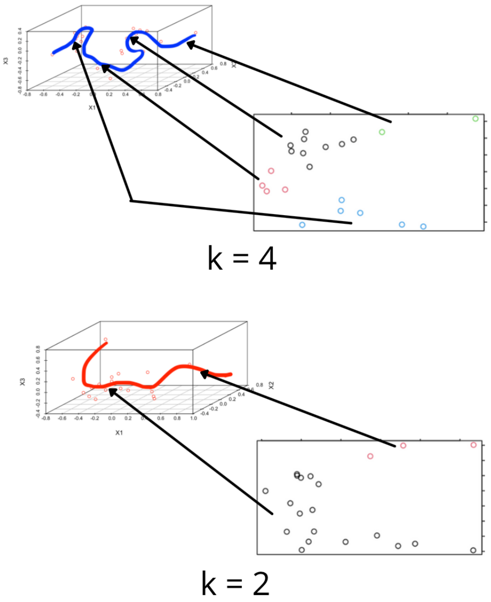

Figure 5: My K-means clustering algorithm, that isolates spatially clustered domains, mapped onto my generated 3D models for WT(above) and KO(below)

Thus, I was able to use Classical MDS along with K-means clustering in order to clearly diffentiate WT vs. KO mice for the LDB1 gene on the basis of the 3D conformation of murine nuclear architecture related to olfaction.

4 Discussion

After a lot of struggling, things did turn out better than I expected. I was able to implement Classical MDS and K-means clustering for my Hi-C data and convert 3C to 3D!

One challenge I faced was the vastness of Hi-C data, which forced me to choose small segments to run my code on, else run-times would get too large. I wish to be able to apply my strategy to multiple other regions in the genome to understand more about LDB1 protein's effects on the mouse genome. Another challenge I faced was understanding the algorithms in order to implement them, and I am still working on making my code more robust, and extending it to have larger scope for different functionality, like k-mer[6] sequence space to try to understand genomic distance space and its relation to Euclidean space. Further, I would like to investigate phase separation[7] and multi-way binding through Hi-C data.

5 Acknowledgements

I would firstly like to thank my peer-reviewer Peter Halmos, for giving extremely detailed and constructive feedback. Next, I would like to thank Professor Knowles for his excellent and clear teaching style, that equipped me with the skills I needed to take on this project. Lastly, I would like to thank Ariel Pourmorady, an MD-PhD student at the Lomvardas Lab at Columbia, for helping me come up with my topic of exploration.

References

[1] Monahan, K., Horta, A. Lomvardas, S. LHX2- and LDB1-mediated trans interactions regulate olfactory receptor choice. Nature 565, 448–453 (2019). https://doi.org/10.1038/s41586-018-0845-0

[2] Dokmanic, I et al.(2015). Euclidean Distance Matrices: Essential theory, algorithms, and applications. Institute of Electrical and Electronics Engineers (IEEE). Retrieved from http://dx.doi.org/10.1109/MSP.2015.2398954

[3] https://genome.ucsc.edu/cgi-bin/hgTracks?db=mm10&lastVirtModeType=default&lastVirtModeExtraState=&virtModeType=default&virtMode=0&nonVirtPosition=&position=chr19%3A46031570%2D46045214&hgsid=1095968569_QWOYviW1MKchR7JqaHpI7ug1iiRA

[4] Lieberman-Aiden E et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009 Oct 9;326(5950):289-93. https://doi.org/10.1126/science.1181369

[5] Spitz F. Chromosomes come together to help mice distinguish odours. Nature. 2019 Jan;565(7740):439-440. https://doi.org/10.1038/d41586-019-00010-6

[6] Trofimov, A et al. (2018) Towards the Latent Transcriptome. https://arxiv.org/pdf/1810.03442.pdf

[7] Liu, T., Wang, Z. (2018). Reconstructing high-resolution chromosome three-dimensional structures by Hi-C complex networks. BMC bioinformatics, 19(Suppl 17), 496. https://doi.org/10.1186/s12859-018-2464-z