3D genome Hi-C contact maps topological data analysis, a brief review

Youssef OUADGHIRI
Ouadghiri.yo@gmail.com
Université Côte d'Azur

February 7, 2022

1 Introduction

1.1 Chromatin interaction

The chromatin is a complex protein and DNA structure that packs long DNA molecules together in a dense form to protect DNA molecules from damage, reinforce them during cell division and regulates their replication. It is packaged in the nucleus (see figure 1). It tries to package the DNA in the. right way for the correct interactions between DNA sections (enhancer to activate genes), this is very important as any wrong interaction can cause disease (ex: Deformities or cancer [FDL+16]).

In order to look at interactions in chromatin, we have some techniques (see figure 2): 2 regions of the chromatin interact, we use restriction digest to cut the regions into small sections (few Kb), then an enzyme is used to ligate the segments, we obtain a reverse cross-linked region. Then we apply a specific method:

- 3C: Chromatin Conformation Capture, 1 to 1, We only keep the fragment around the ligation.
- 4C: 1 to many, We take a particular region and. see how it interacts with all the nearby regions.
- 5C: Many to many, it is useful for long sequences analysis, for example, taking a 1MB region, we look at all the interactions inside the region.
- Hi-C: All to all, we see this method in the next section.

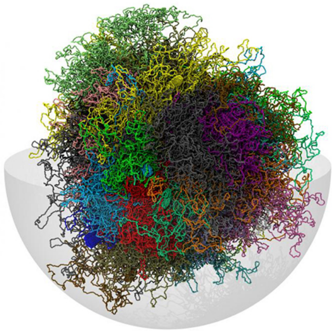

Figure 1: Chromatin inside nucleus, here we see the long DNA molecule in the Chromatin and the way it is packed makes some regions collide and interact.

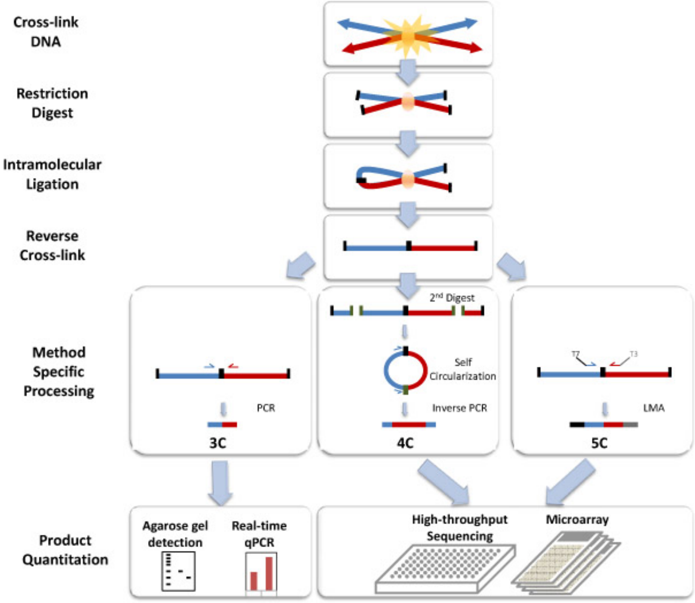

Figure 2: Here we see some techniques used to look at the interactions, the chromatin is ligated, cut into smaller fragments, then the extremes are stuck together through a reverse cross-link, then a specific method is used to get depending on the way we want to look at the interaction 3C (1 to 1), 4C (1 to many) or 5C (many to many).

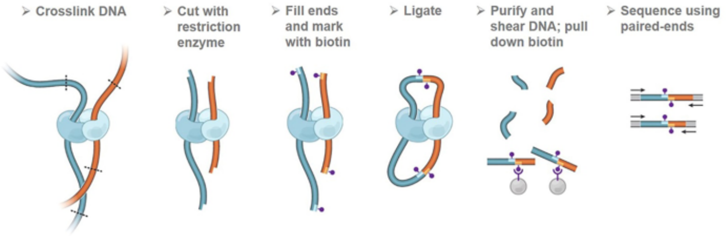
Crosslink DNA
Cut with restriction enzyme
Fill ends and mark with biotin
Ligate
Purify and shear DNA; pull down biotin
Sequence using paired-ends

Figure 3: These are the Hi-C analysis steps as described in the Lieberman-Aiden 2009 paper [LAVBW+09], 1- we cross-link spatially close DNA fragments, 2- we cut sections with an enzyme, 3- we fill the ends, 4- we ligate the segments, 5- we purify the DNA, 6- we obtain paired sequences .

2 Hi-C and TADs

2.1 Hi-C

Hi-C uses high throughput sequencing, it captures the nearby DNA fragments in the 3D genome, we do it in a similar way we did for previous methods(see3), sequencing using paired ends lets us see the interaction between genomes, the resolution get smaller the more we sequence, the first resolution gave a resolution of 40Kb. With this experiment, we can see the chromatin interactions, and [LAVBW+09] obtained the first interaction map by plotting a heat-map taking the chromosomes, ordered spatially, as features. (see figure 4)

2.2 Topological Association Domains

In many experiments, we notice the little squares around the diagonal. line in the matrix represented in the heat-map in figure 4, those are called Topological Association Domains TADs, they show highly self-interactive sub-regions, at around 100Kb resolution. A Pearson correlation evaluation shows that indeed, the correlation within the TAD is higher than the correlation across different TADs, even for the same distance (same resolution). Based on the interaction heat-map and TADs results, [DSY+12]

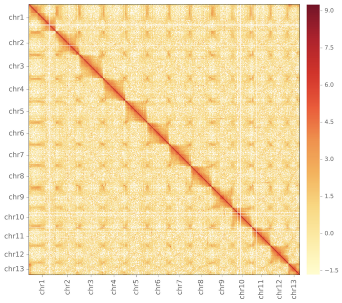

Figure 4: Hi-C interaction map proposed by Lieberman-Aiden in 2009 [LAVBW+09], we can see that the heat-map is warmer in the diagonal, meaning the interaction is happens mostly between chromosomes close to each other.

proposed a Directionality Index DI, the index tries to see how most chromatin interact, they noticed that the chromatin interact with neighbouring chromatin in a very organized way, forming well shaped noticeable triangles on the superior matrix of the data (the data matrix is symmetrical). Using the symmetry, they computed TADs boundaries, and found that a disruption in TADs and their boundaries could be responsible of some diseases [LKH+15] like Brachydactyly, F-syndrome or (hands and fingers deformations).

3 Topological analysis

Topological data analysis TDA is a set of mathematical methods to extract structure and topology of complex data. It is derived from Algebraic topology and it showed good results in many applications in the biological field [SSK19]. It is suiting the of 3D genomes analysis and more specifically our Hi-C data. For instance, in this analysis, we use a Mapper [SMC+07], it is a visualization method that highlights the topological and plots a graph of nodes, we will explore it more in the next sections.

3.1 Contact maps

There is a number of tools to visualize the Hi-C data:

- Juicebox(Aiden) [Jui]: This tool was produced by the inventors of the Hi-C experiment, it can be used to look at the matrices interactively.
- HiGlass [KAL+18]: developed by Harvard, it provides some pre-loaded data that can be looked through in a smooth interactive way.

3.2 Comparison

As the raw Hi-C data is likely to be unreliable, [YZY+17] proposed a two-stage approach to assess data:

1. 2D mean filtering: this method has already shown good result for Hi-C data normalization, it tries to solve the problem of large interaction space in Hi-C experiment. The filter is used to smooth the contact map, it is fast and effective for this data, it replaces the mean of each contact by the mean counts of contacts in its genomic neighbourhood.
2. Stratification: The contacts in the contact map are stratified to keep the distance importance in the reproducibility assessment.

Then they applied a Stratum-adjust correlation coefficient SCC. SCC is a reproducibility metric, it is also the best statistic to compare Hi-C contact maps data. It tests if variables keep similarities when they are stratified to a third-part variable [Agr03]. SCC uses the Pearson correlation, and can be computed with the following relation:

$$SCC = \frac{\Sigma_{i=1}^{s} r_i (m-i) \sqrt{var(v_i^x) var(v_i^y)}}{\Sigma_{i=1}^{s} (m-i) \sqrt{var(v_i^x) var(v_i^y)}}$$

With v_i the i-th stratum, (m-i) the length of v_i, r_i the Pearson correlation coefficient of v_i^x and v_i^y. And $var(v_i) = \frac{\Sigma_{j=1}^{m-i}(v_{i,j} - \bar{v_i})^2}{m-i}$.

3.3 Mapper

A mapper tries to create a graph (also called simplicial complex) that conserves the topological properties of the space. It is a great way to visualize the important connectivities of a structure of data. The mapper algorithm first projects a dataset, then covers the projection with overlapping hyper-cubes, then performs local clustering.

4 Results

4.1 Challenges and difficulties

For our application, the data is in a 500kB resolution, divided into 1171 folders npz (numpy matrix) format, the format is proper to the numpy library in Python and is hard to convert. The Hi-C specific libraries in Python mostly work with Cooler, h5py, Hic or matrix formats:

- cool/mcool: An efficient storage format for high resolution genomic interaction matrices.
- h5py: Lets you store huge amounts of numerical data, and easily manipulate that data from numpy.
- Hi-c: a=An indexed binary format designed to permit fast random access to contact matrix heatmaps.
- matrix: Represents a matrix, mostly used with R.

The npz data format pushes us to work in a Python framework from scratch, the SCC function has been written according to the mathematical formula. Unfortunately, the computation time is very high and we only take a subsample data of each cell cycle (matrices between numbers 1 and 14). The direct solution is using specific libraries, among many we cite:

- HicRep [YZY+17]: R package to evaluate the reproducibility of Hi-C data, the Python version was developped by Dejunlin in his Github.
- HicPro [SVL+15]: Designed to process Hi-C data, from raw fastq files (paired-end Illumina data) to normalized contact maps.
- Cooler [AM19]: Python library handling the cool format

The libraries do not take npz data in consideration. Moreover the conversion requires some metadata (ex: bins_id1...). Another solution is to work with frameworks containing data such as Juicebox or Higlass [DRS+16, KAL+18], however, the project obliges us to work with a certain data, that is pre-processed and non-identifiable because of a lack of biological background.

Finally, due to the limited computational power, further tests could not be done locally.

Consequently, although a python notebook will be provided to show the trials we implemented. The following results will be taken from literature for more correct and concluding interpretations.

4.2 SCC

After getting the contact pairs, the data is smoothed with a moving average of size 1, and then the SCC pairwise matrix is computed, the distance is then calculated[LLYN18]: Distance :

$$D(M_k, M_{k'}) = \sqrt{k(M_k, M_k) - 2k(M_k, M_{k'}) + k(M_{k'}, M_{k'})}$$

We then use this distance matrix to draw the mapper.

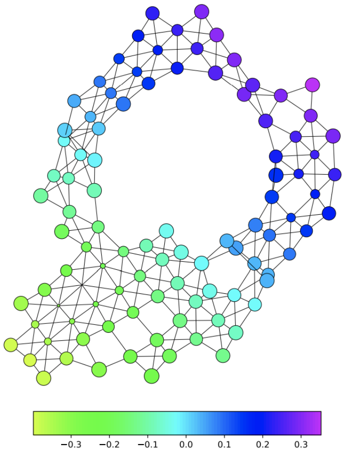

Figure 5: This mapper is obtained with the smoothed SCC matrix done by [CR20], we notice the circle in the middle, hinting to the existence of a correlation with the cell cycle.

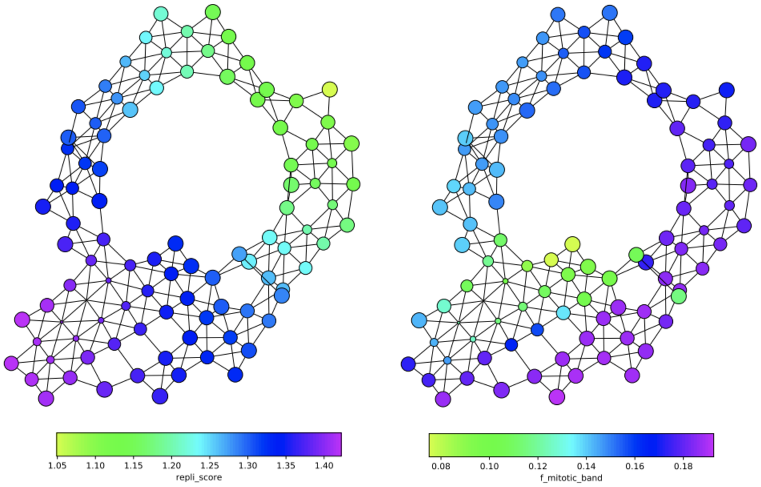

Figure 6: Mappers colored according to some cell features done by [CR20]. The correlation is confirmed, some values increase in some nodes of the circle and decrease in other nodes.

4.3 Mapper and interpretation

The figure 5 shows the mapper we obtained after filtering, projecting, covering overlaying data with hyper-cubes, clustering and drawing the nodes in a topological way. Notice the presence of a circle, this is probably due to the data being extracted during different steps of the cell cycle: $G_1, S, G_2 and M$.

To confirm the theory, [NLV$^+$17] studied the correlation and then colored the nodes in the mapper according to some features related to some precise phases of the cell cycle, as seen in the figure **??**, we take the example of ”repli-score” and ”f mitotic band”, same results can be shown for ”f near band” and ”mean insulation of TADs”, the coloration confirms the correlation, some values increase in some particular phases, and their topological position gets smaller (dense regions).

4.4 Formal Statistical test: Bootstrap

To quantify the cell cycle statistical robustness, and formally prove it is not an artifact of a computation, we can use some statistical methods. For this case, we bootstrap the data: bootstrapping the data means sub-sampling the dataset many times with replacement. Then we see how common and reproducible are our result. The bootstrap done by [CR20], show that the confidence level for the point corresponding to the loop is 93%. This high score validates the result of the experiment.

5 Conclusion

This experiment have captures successfully biologically important variation in the chromatin by capturing its topological structure. TDA has indeed proven extremely powerful in this kind of study cases, however, it is vital to not ignore the importance of the pre-processing, SCC and the other statistical processing the data went through to become conclusive. The raw interaction data (contact maps) are inexpressive if we try to model them directly. Linear dimensionality reduction does not yield any result, and non-linear data reduction does not preserve well the spatial data, hence the importance of TDA in this application.

References

[Agr03] Alan Agresti. *Categorical data analysis.* John Wiley & Sons, 2003.

[AM19] Nezar Abdennur and Leonid A Mirny. Cooler: scalable storage for Hi-C data and other genomically labeled arrays. *Bioinformatics*, 07 2019.

[CR20] Mathieu Carrière and Raúl Rabadán. Topological data analysis of single-cell hi-c contact maps. In *Topological Data Analysis*, pages 147–162. Springer, 2020.

[DRS+16] Neva C Durand, James T Robinson, Muhammad S Shamim, Ido Machol, Jill P Mesirov, Eric S Lander, and Erez Lieberman Aiden. Juicebox provides a visualization system for hi-c contact maps with unlimited zoom. *Cell systems*, 3(1):99–101, 2016.

[DSY+12] Jesse R Dixon, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S Liu, and Bing Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380, 2012.

[FDL+16] William A Flavahan, Yotam Drier, Brian B Liau, Shawn M Gillespie, Andrew S Venteicher, Anat O Stemmer-Rachamimov, Mario L Suvà, and Bradley E Bernstein. Insulator dysfunction and oncogene activation in idh mutant gliomas. *Nature*, 529(7584):110–114, 2016.

[Jui] Juicebox.js provides a cloud-based visualization system for hi-c data.

[KAL+18] Peter Kerpedjiev, Nezar Abdennur, Fritz Lekschas, Chuck McCallum, Kasper Dinkla, Hendrik Strobelt, Jacob M Luber, Scott B Ouellette, Alaleh Ahzir, Nikhil Kumar, Jeewon Hwang, Burak H Alver, Hanspeter Pfister, Leonid A Mirny, Peter J Park, and Nils Gehlenborg. Higlass: Web-based visual exploration and analysis of genome interaction maps. *Genome Biology*, 19(1):125, 8 2018.

[LAVBW+09] Erez Lieberman-Aiden, Nynke L Van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, Michael O Dorschner, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science*, 326(5950):289–293, 2009.

[LKH+15] Darío G Lupiáñez, Katerina Kraft, Verena Heinrich, Peter Krawitz, Francesco Brancati, Eva Klopocki, Denise Horn, Hülya Kayserili, John M Opitz, Renata Laxova, et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*, 161(5):1012–1025, 2015.

[LLYN18] Jie Liu, Dejun Lin, Galip Gürkan Yardımcı, and William Stafford Noble. Unsupervised embedding of single-cell hi-c data. *Bioinformatics*, 34(13):i96–i104, 2018.

[NLV+17] Takashi Nagano, Yaniv Lubling, Csilla Várnai, Carmel Dudley, Wing Leung, Yael Baran, Netta Mendelson Cohen, Steven Wingett, Peter Fraser, and Amos Tanay. Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature*, 547(7661):61–67, 2017.

[SMC^{+}07] Gurjeet Singh, Facundo Mémoli, Gunnar E Carlsson, et al. Topological methods for the analysis of high dimensional data sets and 3d object recognition. *PBG@ Eurographics*, 2, 2007.

[SSK19] Natalie Sauerwald, Yihang Shen, and Carl Kingsford. Topological data analysis reveals principles of chromosome structure throughout cellular differentiation. *bioRxiv*, page 540716, 2019.

[SVL^{+}15] Nicolas Servant, Nelle Varoquaux, Bryan R Lajoie, Eric Viara, Chong-Jian Chen, Jean-Philippe Vert, Edith Heard, Job Dekker, and Emmanuel Barillot. Hic-pro: an optimized and flexible pipeline for hi-c data processing. *Genome biology*, 16(1):1–11, 2015.

[YZY^{+}17] Tao Yang, Feipeng Zhang, Galip Gürkan Yardımcı, Fan Song, Ross C Hardison, William Stafford Noble, Feng Yue, and Qunhua Li. Hicrep: assessing the reproducibility of hi-c data using a stratum-adjusted correlation coefficient. *Genome research*, 27(11):1939–1949, 2017.