Meet-EU project report
M2 BIM-info 2021-2022

TADs borders detection based on consensus deep learning approach

Maxime Christophe, Antoine Szatkownik, Wiam Mansouri

Contents

1. Context

The genomes of eukaryotic cells of mammalian organisms is made up of about three billions base pairs. If we were to stretch it end to end it would be two meters long yet the size of the nucleus where it lies is at the order of the micrometer scale. This is equivalent to packing a 40km long thin wire into a tennis ball. The cell solves this genome folding problem by a layered organization structure. Histone protein structure helps the DNA to adopt a compact conformation, called chromatin, while being at the same time in a biologically active state allowing for replication, repair, gene expression... In order to express a gene, certain areas of less dense level of compaction needs to be spatially close. These self-interacting chromatin regions are called Topologically Associated Domains (TAD). The assertion that there is a correlation between the expression of genes and the presence of TADs has been reconsidered [1] yet we can question about their role in gene regulation and more generally their functions. Thus the identification of patterns of chromatin structures is crucial for understanding how the spatial organization of DNA affects genome functionality.

Methods based on chromosome conformation capture techniques such as Hi-C, High Chromosome Contact map, enables one to analyze the spatial proximity of DNA segments within a genome. It determines the average number of contacts between distant genomic regions with a resolution down to few kilobases. TADs are characterized by groups of genomic loci that have high levels of interaction within the group and minimal levels of interaction outside of the group [5]. Such contact matrices harbor different kinds of patterns like TADs, loops, that remains to be detected. Since there is a large number of TADs even for a single chromosome, the task of identifying them needs to be automatized. For this matter, several computational methods have been established, among them figures Arrowhead [2], TopDom [3] just to name a few. Despite the vast range of TAD prediction tools, there is no consensus way to define a TAD [4], no universal protocol was brought out to identify them. Each method is built on a different set of assumptions about TADs like TAD size, biological signification of TADs overlap, TAD boundaries are enriched in CTCF binding sites [5]... They are sensitive in a different ways to parameters like resolution, coverage. Thus the problem of TAD prediction is still open.

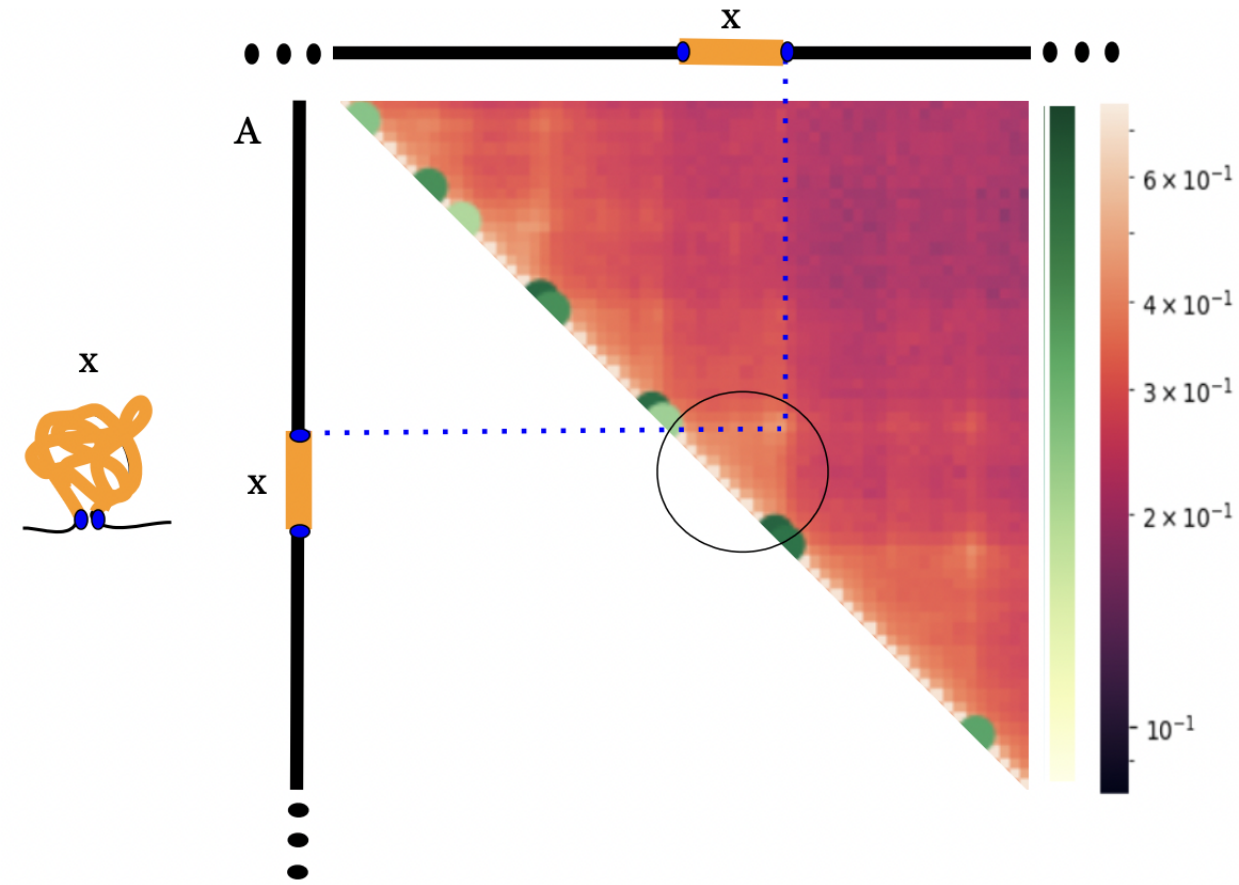


Figure 1. Outline of the problem. Extract from a Hi-C contact map. Legend : Black thick lines are the same portion of chromosome A. Orange segment called "x" is a genomic region of A, the boundaries of x are delimited by blue points. The dashed blue lines intersect at a loop. The circle encompass the TAD represented as a triangle (or square, the Hi-C matrix is symmetric). On the left is the non linear representation of x. A TAD is a localized cluster of contacts, hence the bright triangle. The green dots on the diagonal represent predictions of TADs boundaries, the color represent the score of the prediction.

2. Material and Methods

Following our reading of the literature we found that Chromosight tool [4] was standing out from the others since it treated the TADs detection problem as a Computer Vision problem. It inspired us and gave us hope that a new method based on convolutions between learned kernels and the HiC matrix could work since it did work with a priori defined kernels. That is why we decided to tackle this problem by creating a convolutional neural network (CNN).

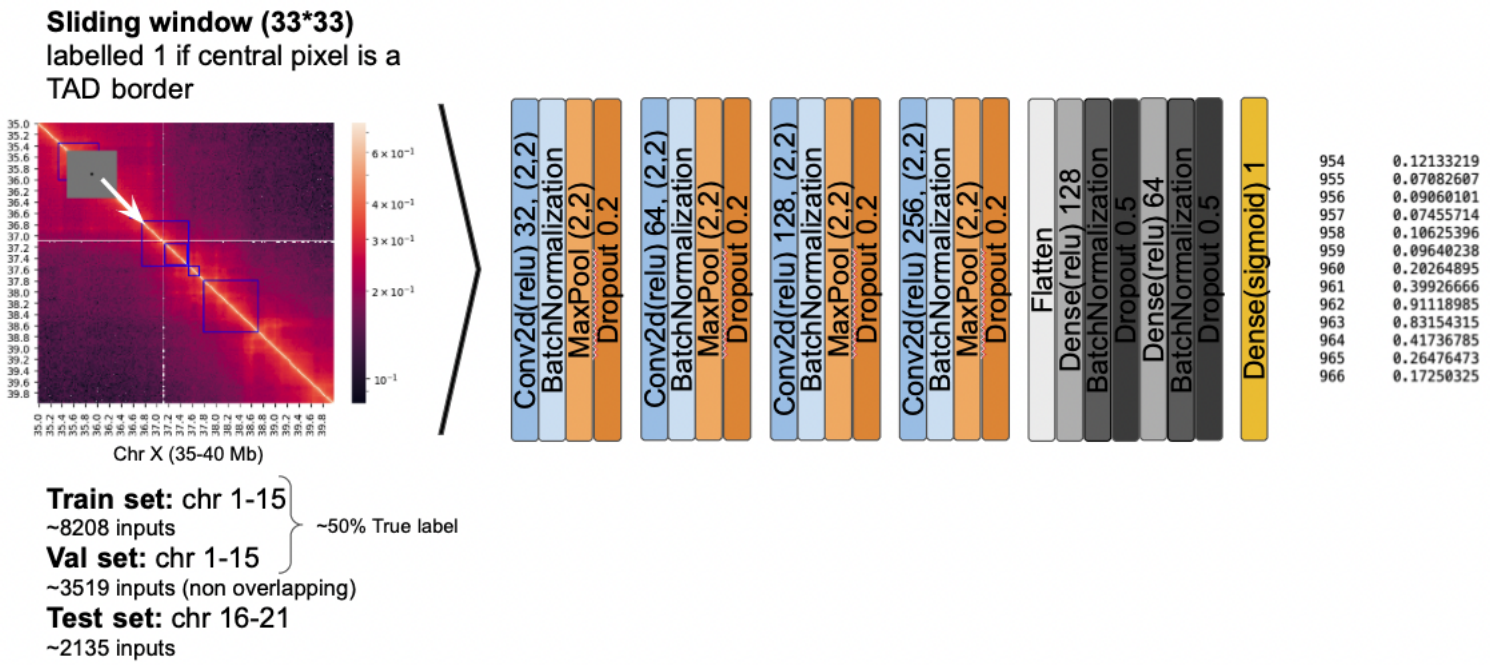


Figure 2. Overview of the method. A threshold is applied on the output to filter what we consider as a TAD border.

2.1. Data

We were given :

- Hi-C Data for five different cells of the human body, GM12878, HMEC, HUVEC, IMR90, NHEK. For each cell, we have data for all chromosomes at 25kb and 100kb resolution, of intrachromosomal type. A raw Hi-C matrix is an integer $n \times n$ matrix M, such that n is the number of bins of a chromosome and M_{ij} is the average contact frequency of the genomic bin i and j. For example, for a chromosome of 141 100 000 base pairs, at a resolution of 25kb, the Hi-C matrix will be of dimension $\frac{141100000}{25000} = 5644$
- a validation set consisting of the output of Arrowhead algorithm over the given Hi-C data as input. It contains the positions of the TADs.

Complementary data were used, namely the materials used in [3]. We thus also integrated the results of Armatus, DomainCaller, HiCSeg, TADbit, TADtree and TopDom into our method. For evaluation we chose Chromosight as groundtruth because it wasn't present in [3]. We converted the test set to cool format so that we could generate the results of Chromosight.

2.2. Pre-processing

In their benchmark, Dali and Blanchette used the following tools to detect TADs from 25kb resolution HI-C data, in the GM12878 human cell type: Armatus, Arrowhead, DomainCaller, HICSeg, TADbit, TADtree, TopDom. In order to label the data, the provided benchmark has been pre-processed so that each position matched with a unique TAD. When multiple TADs were detected with a common start position or end position we kept the TAD detected by the highest number of method. On top of that, the TADs detected by less than six methods were discarded. We ended up with a total of 3266 TADs among the whole genome. The 3266

positions labelled as 0 (not a tad border) have been randomly selected among positions that were not detected by any tools in the benchmark.

2.3. CNN architecture

We use a convolutional neural network, which is an efficient way to deal with this kind of matrix data. The CNN was built as shown in Figure 2.

The visible part of the network is built with four blocks composed of:

- A convolutional layer
- A batch normalization layer
- A pool layer
- A dropout layer

The hidden part of the network is composed of two blocks composed of:

- A dense layer
- A batch normalization layer
- A dropout layer

2.4. Training parameter tuning

The training has been done on the chromosomes 1 to 15 from the GM12878 cell type. 30% of these data were used as validation set. In order to avoid overfitting, several strategies have been set up:

- Dropout layers: the dropout layer randomly deactivates neurones to drop some information between layers
- Learning rate: The learning rate from the adam optimizer have been parametrized to 0.005 (default parameter is 0.01) to decrease the impact of gradient back-propagation.
- Early stopping: To avoid the network to learn too much from the network noise, the validation loss was monitored during the training, so that the training would be stopped when the performance decreases.

  We did train several models that over-fitted each times and produce unsatisfying results. The strategies cited above leads us to a model with fine training scheme, some other corrections can be done to improve the performances by doing longer training without decreasing performance.

Here we can see that the model was trained on only 13 epochs, which is a really small amount. The early stopping did detect the decrease of the model performance and stopped the training there. The problem here is that it is difficult to observe any convergence and be sure that the model reached is maximal training performance.

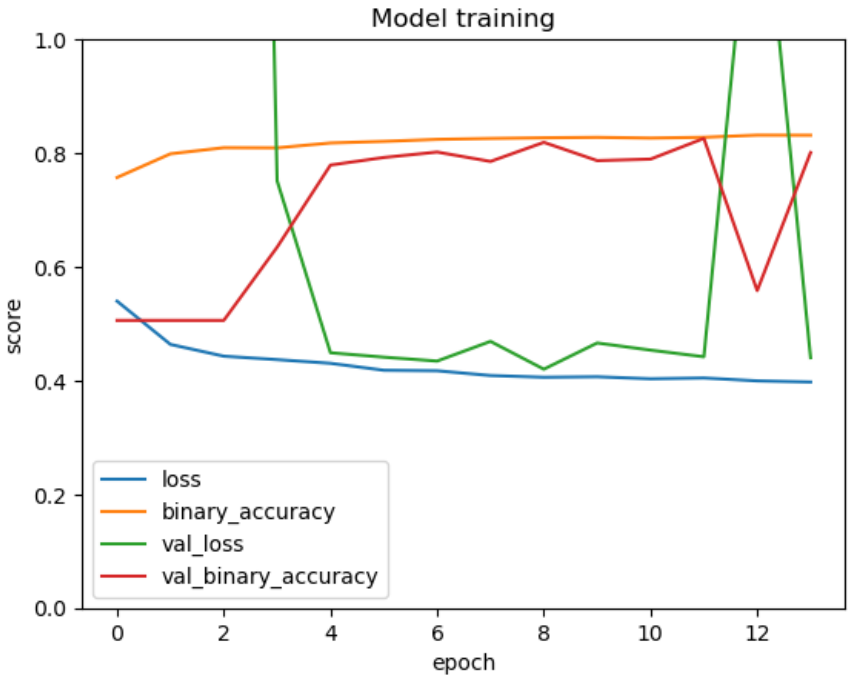


Figure 3. Training history of the model, stopped at epoch 13th to conserve the model's performance, we can indeed observe that validation and train metrics stay really close

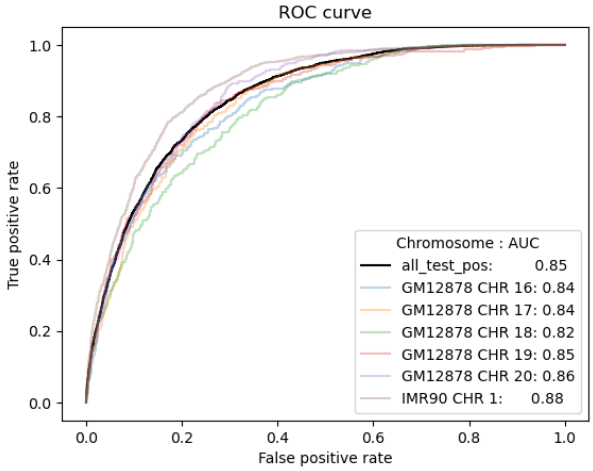
3. **Results**

In order to evaluate our results with a tool that was not involved in the labeling of our data, we used Chromosight outputs [4] as an independent evaluation tool.

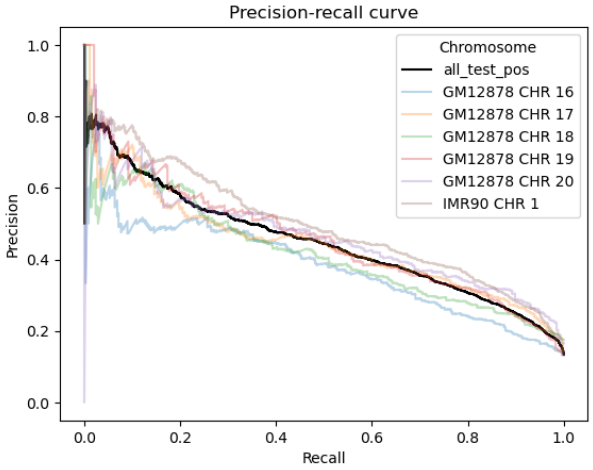
As evaluation data, we selected the chromosomes that were not used to train the model: chromosome 16 GM12878, chromosome 17 GM12878, chromosome 18 GM12878, chromosome 19 GM12878, chromosome 20 GM12878 and added chromosome 1 IMR90 to test the model on an other cell type.

The model outputs are scores between 0 and 1, the best results was obtained for a threshold of 0.5 (>0.5 positions score are considered as TAD border.)

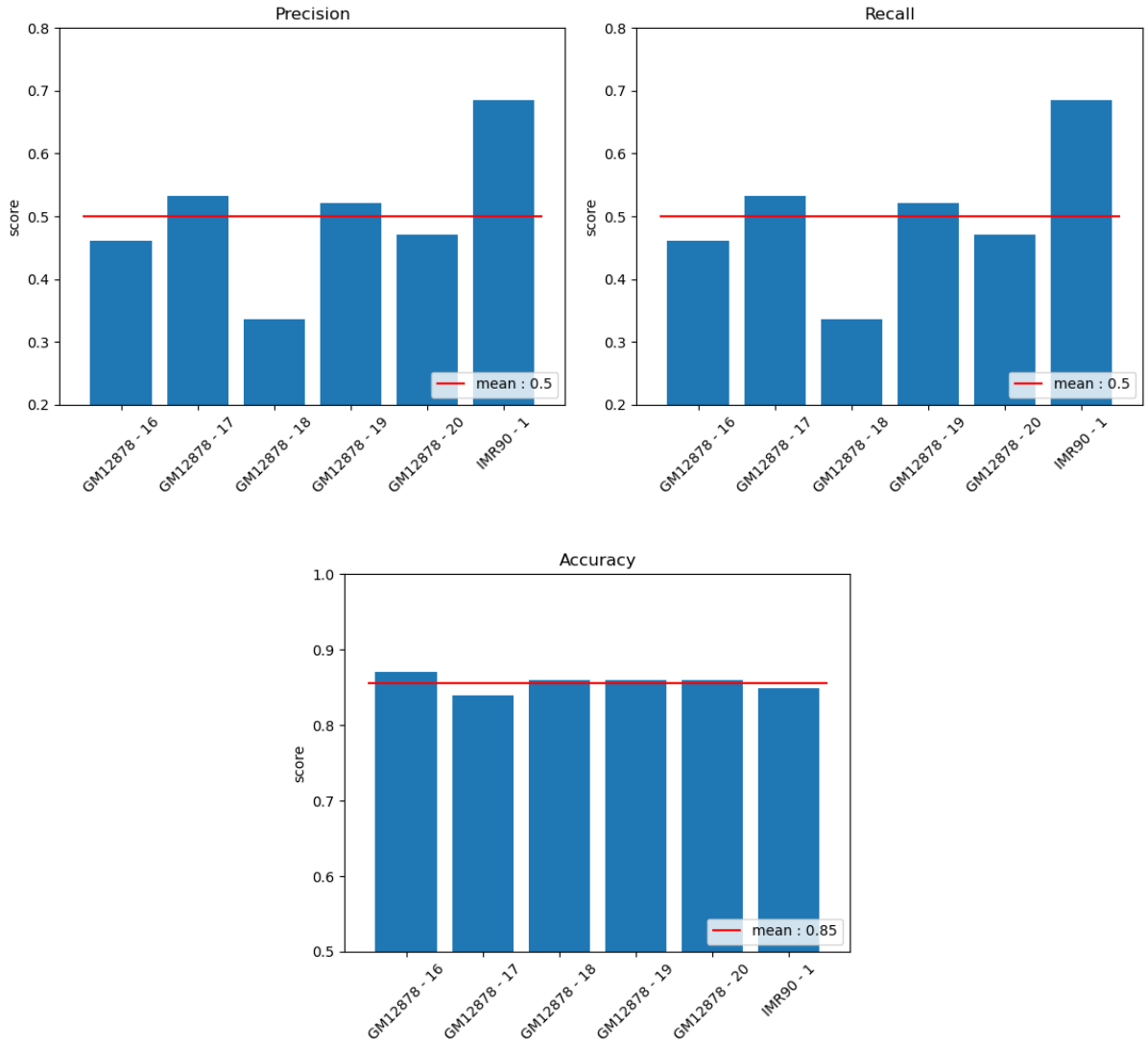
Despite the short training, we can observe nice results with a Receiving Operator Curve with 85% area under the curve.



(a) Receiving operator curve: We can observe a good performance considering the false positive and true positive rates



(b) Precision recall curve: Despite the good accuracy, the recall increases quickly



(a) We can observe singular results on the metrics. The accuracy is good (around 0.85) but precision and recall are only at 0.5, in this case we can see that the precision is strictly equal to the recall, meaning that we have the same amount of false positives and false negatives. An interesting point to note is the really good results on the chromosome 1 from IMR90 cell type. We can imagine that chromosomes 1 from IMR90 and GM12878, share substantial identity, leading to good results, as chromosome 1 GM12878 was part of the training set.

4. Comparison with paired team

The Heidelberg team HA1 who implemented MinCutTAD approached the problem completely differently. They did not predicted TAD borders but computed the likelihood of a pixel to be in a TAD. Moreover their test set and our test set overlapped on one chromosome only, the chromosome 20.

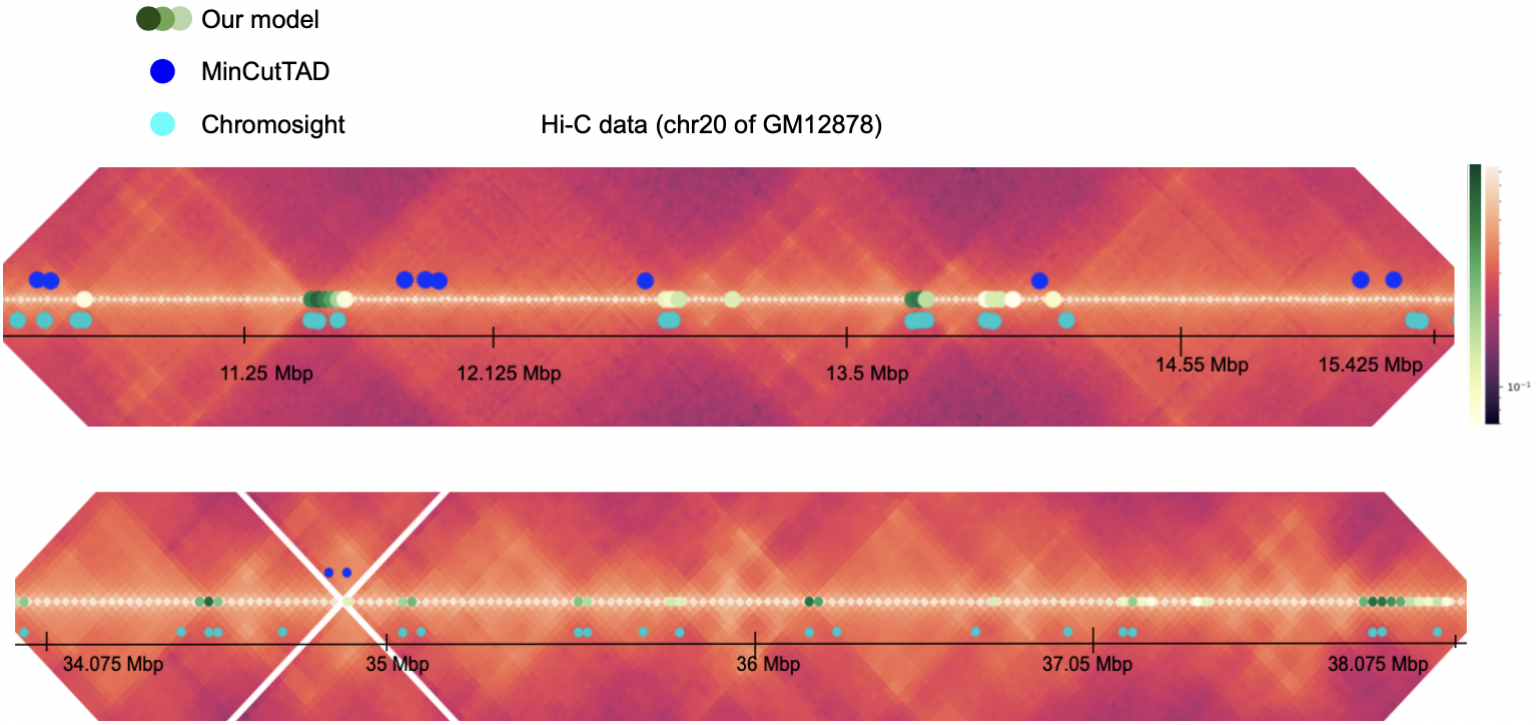


Figure 6. Sample comparison of our model, MinCutTAD the model of Heidelberg's team HA1, and Chromosight on chromosome 20 of cell line GM12878 at 25kb resolution. Dots corresponds to predicted TADs borders. The green color represent the score of ours predictions from 0.5 (threshold value) to 1.

5. Challenges

We encountered a glitch in the predictions of our model that hasn't been fully resolved. At regions between two TADs we have multiple borders detected. This might be due to the fact that TADs can intersect and that the end of a TAD and the beginning of another one are close to one another. A partial approach was carried out by clustering the aggregates and choosing as representative the point with highest score. This approach isn't completely right because a region between two TADs has at least two borders as mentioned before, yet a clustering method will flatten them to one single point. Aside technical problems, a member of our team decided to quit the Master hence leaving only the three of us working on this project.

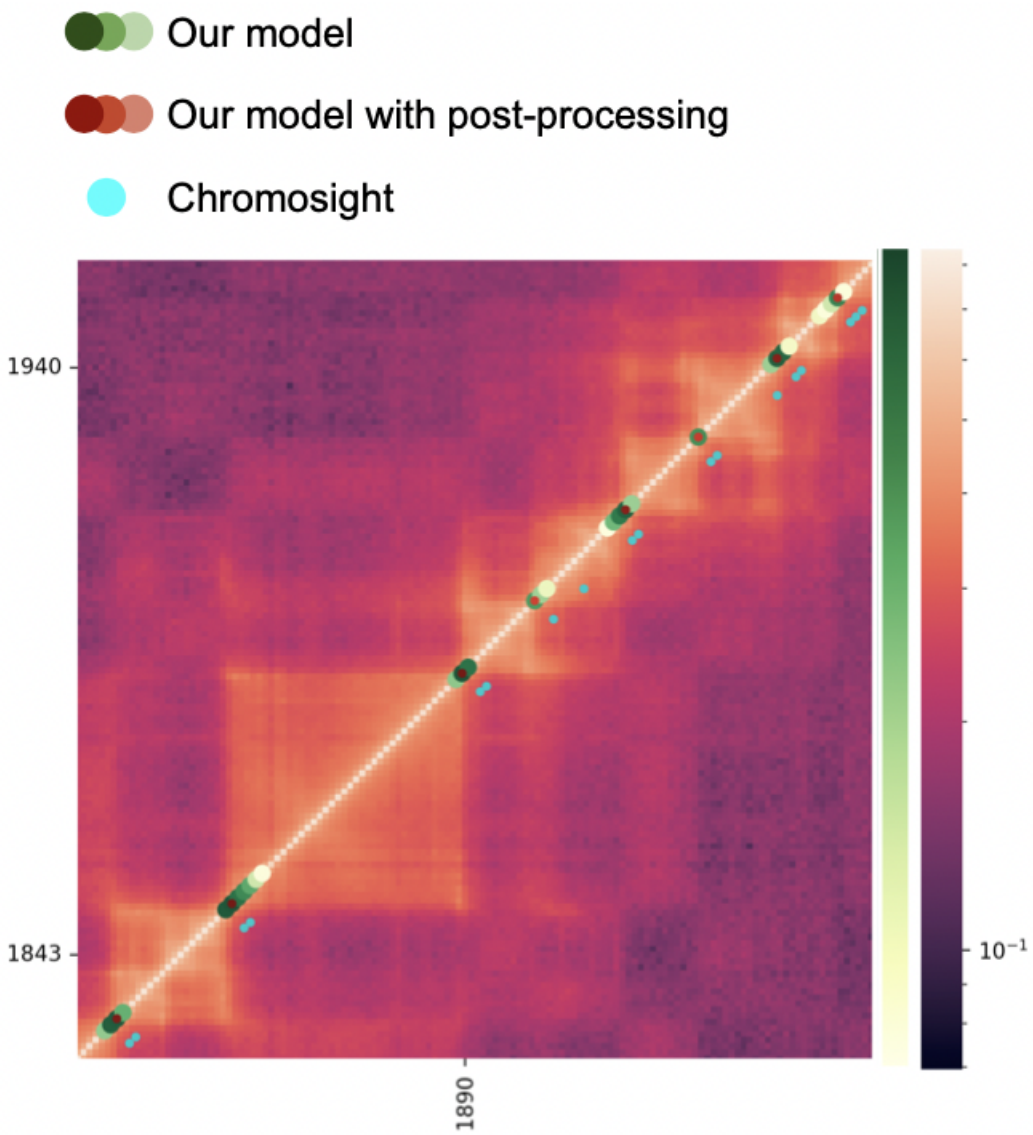


Figure 7. Aggregation problem. Multiple borders are predicted at a given region. The green dots on the diagonal represent predictions of TADs boundaries, the color represent the score of the prediction from 0.5 (threshold value) to 1, the gradient of red dots represent the representative and in cyan is Chromosight predictions.

6. Discussion

The model introduced in this paper cannot be seen as a final product due to the results that cannot achieve as much as the tools we cite. However this work is a good proof of concept of the application of deep learning in TADs border detection. The main points to improve are the model architecture and training, in order to train longer to reduce the number of false negatives: as we only trained on "obvious" borders (detected by a large number of tools), it might be difficult for the model to interpret the windows corresponding to borders detected by a few tools, that did not exist in our sets (we only included pixels that were never detected by any tools as 0 labels). The other point to improve is the post-processing of the data, particularly the clustering of really close detected borders, it would reduce the number of false negatives. We can still observe on the figures 6 and 7 that our results tend to match with Chromosight's results, which is a good achievement and demonstrate the potential of deep learning approaches.

7. Meet-EU experience

The meet-eu experience was a nice opportunity to work on open problems and share the results with a community of scientists and students. We feel a bit frustrated because we would have liked to have more time to go deeper, develop and present some more ideas in this report. However we are willing to keep working on this project to improve our model and maybe give birth to a robust and useful tool in TADs border detection.

References

[1] Yad Ghavi-Helm; Aleksander Jankowski; Sascha Meiers; Rebecca R. Viales; Jan O. Korbel; Eileen E.M. Furlong; Highly rearranged chromosomes reveal uncoupling between genome topology and gene expression ,*Nat Genet.*, **2019**, doi:10.1038/s41588-019-0462-3.

[2] Suhas S.P. Rao; Miriam H. Huntley; Neva C. Durand; Elena K. Stamenova; Ivan D. Bochkov,; James T. Robinson; Adrian L. Sanborn; Ido Machol; Arina D. Omer; Eric S. Lander; Erez Lieberman Aiden; A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping ,*Cell*, **2014**, doi:10.1016/j.cell.2014.11.021

[3] Rola Dali; Mathieu Blanchette; A critical assessment of topologically associating domain prediction tools ,*Nucleic Acids Research*, **2017**, doi:10.1093/nar/gkx145

[4] Cyril Matthey-Doret; Lyam Baudry; Axel Breuer; Rémi Montagne1; Nadège Guiglielmoni; Vittore Scolari; Etienne Jean; Arnaud Campeas; Philippe Henri Chanut; Edgar Oriol; Adrien Méot; Laurent Politis; Antoine Vigouroux; Pierrick Moreau; Romain Koszul; Axel Cournac; Computer vision for pattern detection in chromosome contact maps ,*Nature Communications*, **2020**, doi:10.1038/s41467-020-19562-7.

[5] Kellen G. Cresswell; John C. Stansfield; Mikhail G. Dozmorov; SpectralTAD: an R package for defining a hierarchy of topologically associated domains using spectral clustering ,*BMC Bioinformatics*, **2020**, doi:10.1186/s12859-020-03652-w.