SU Team 2 - Digest

TADs borders detection based on consensus deep learning approach

Maxime Christophe, Antoine Szatkownik, Wiam Mansouri

Project Aim

The eukaryotic genome is organized into a hierarchy of structures within the 3D nuclear space. Chromatin, the complex formed by the association between DNA and proteins, is organized into higher order structures such as Topologically Associating Domains which are localized clusters of interacting chromatin regions. In an effort to understand the interplay between the spatial organization of DNA and genome functionality methods to identify patterns of chromatin structures were developed. Here we will focus on the High Chromosome Contact map, Hi-C technique that 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 a few kilobases. Such contact matrices harbor different kinds of patterns like TADs, loops, that remain to be detected in an automatic manner.

Material and Methods

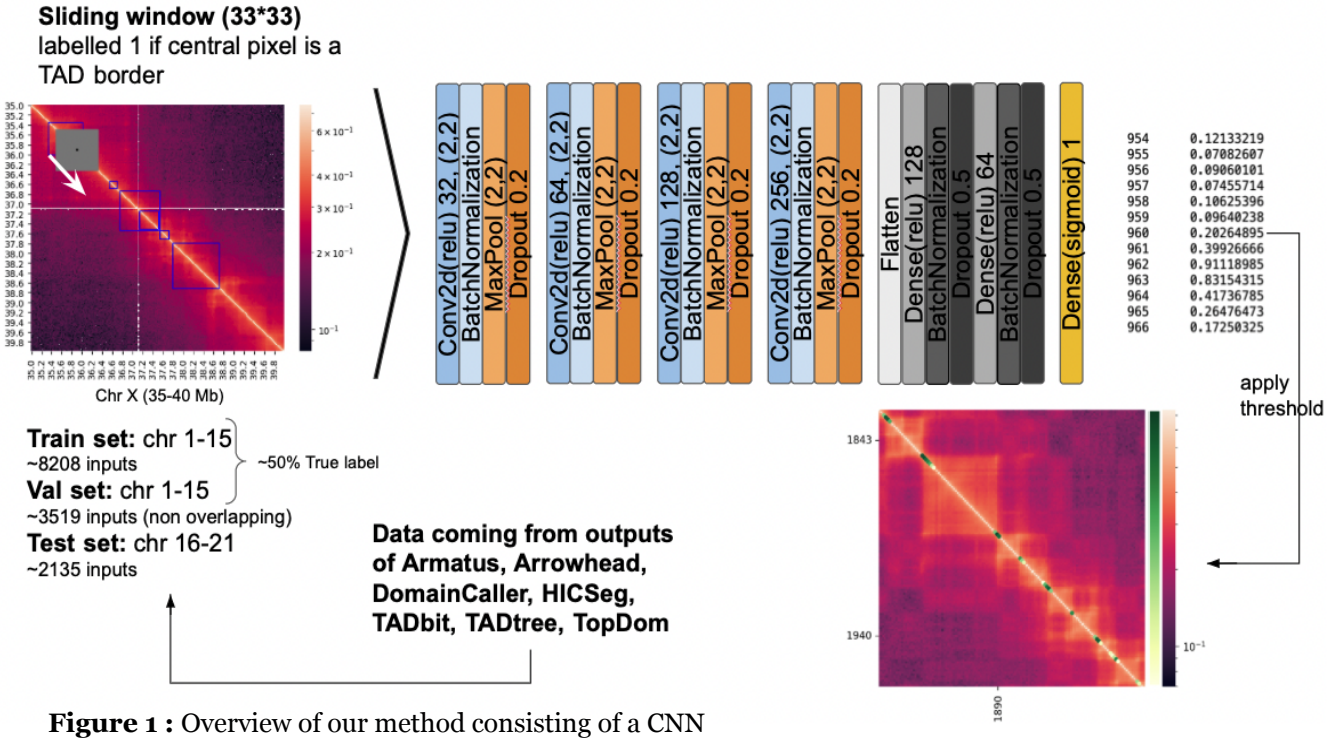

Figure 1 : Overview of our method consisting of a CNN

Following our reading of the literature we found that Chromosight tool was standing out from the others [1] 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).
The material we used was the outputs from the benchmark [1] of Armatus, DomainCaller, HiCSeg, TADbit, TADtree and TopDom. In order to label the data, the provided benchmark has been pre-processed so that each position is 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 methods. 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 labeled as 0 (not a tad border) have been randomly selected among positions that were not detected by any tools in the benchmark.

Results

In order to evaluate our results with a tool that was not involved in labeling our data, we did use Chromosight output (Matthey-Doret et al. 2020) as an independent evaluation tool. As evaluation data, we selected the chromosomes that was not used to train the model: chromosome 1 GM12878, chromosome 17 GM12878, chromosome 18 GM12878, chromosome 19 GM12878, chromosome 20 GM12878 and add chromosome 1 IMR90 to test the model on another cell type. Despite the short training, we can observe nice results with a ROC curve with 85% area under the curve.

Conclusion

As far as the challenges goes 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.
All in all, 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 TAD 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. 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.

Ressources

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