BANANA-TAD:
OBTAINING A CONSENSUS FROM MULTIPLE TAD CALLING ALGORITHMS

Team SA-1 Digest

Abakarova Marina, Crouzet Simon, LeGoffic Liam, Trang Alexis, Zhong Yann

January 28th 2021

Overview

DNA is a greatly complex and important molecule that is often referred to as the building block of life. The entire human genome is able to fit within a single nucleus, and this is in part thanks to how efficiently DNA is able to pack and compress itself while still retaining high levels of necessary interactions for gene expression. One such unit necessary for this function is known as a Topologically Associating Domain (TAD) – a genomic region capable of high self-interaction through the presence of chromatin and loop formation. TADs are a discovery made within the last decade after the advent of chromosome conformation techniques such as Hi-C sequencing.

The recency in TADs' discoveries has come with many questions and challenges associated to it. TADs are important because their presence is often associated to the regulation of gene expression, and disruption in TADs has been linked to oncogenic onset, for example. However, TADs are not trivial to detect. There have been many algorithms and methods such as Topdom or Arrowhead, to cite the two most known, but because of a general lack of consensus on what exactly a TAD is, this field still faces many challenges in the vital aspect of "calling" (or identifying) a TAD from raw sequencing data.

Our method, Banana-TAD, seeks to help solve this challenge by the re-implementation of four existing algorithms: Topdom, TADTree, OnTAD as well as TADBit, and scoring them based on a set of data provided to us as ground truth (as lists of TADs called by Arrowhead on 5kb data). After scoring them, we then obtained a list of consensus TADs, which we believe to be a strong middle ground obtained from the combination of multiple recognized TAD calling methods.

Data

Ground truth data (Arrowhead 5kb) was provided to us by the lecturers.

Testing and training data (intrachromosomal contact matrices at 25 and 100kb resolution) were also provided to us and came from Rao et al. [1]. The cell lines we worked on were: GM12878, HMEC, HUVEC, IMR90, and NHEK.

CTCF data was downloaded as .BED files from ENCODE[2].

It is worth noting that we also split our data into a training and testing set, and kept a fixed seed so as to have our split be reproducible easily.

Methods

The first step taken was the re-implementation or downloading of existing relevant TAD calling methods. Topdom was chosen as it was known as a strong reference method in the field and was based on Python. TADTree was chosen for its ability to call sub-TADs (nested TADs) and Python based code. OnTAD was chosen for its ease of calling and call speed, as well as its similarity to Topdom while showing stronger results. Finally, TADBit was chosen for its Bioconda and naturally Python integration. Other considered contenders were TADTool, HiCTool, and TADCompare, amongst many others.

Once the methods were chosen, we calculated two metrics known as M_{gt} and M_{pred} which are both ratios depending on the number of TADs identified by a method versus by the ground truth, and took their average value in order to find an adapted accuracy A_m value. We also made use of CTCF data in order to determine if a called TAD was CTCF-enriched and extracted a C_m for every method.

The calculated values above would then serve to attribute a weight to each of our 4 methods, following the formula: $W_m = A_m * C_m$, which gave the methods a weight at each resolution that we re-implemented them on, as follows:

Method	TADTree	Topdom	OnTAD	TADBit
25kb	N/A	0.33656	0.84340	N/A
100kb	0.86019	0.32533	N/A	0.27248

Finally, we applied the obtained weights to every coordinate of the test sample chromosomes, where a specific coordinate would be the sum of all its method's weights, with a penalising division applied to the coordinate the farther it was from a predicted TAD boundary. Thus, as Banana-TAD output, we had two lists of consensus TADs (25 and 100kb) for each chromosome.

Results

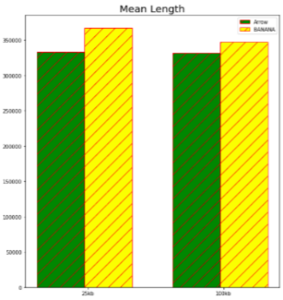

Figure 6: Plot ArrowHead vs Consensus on test set

Table 1: Comparison ArrowHead vs Consensus results on test set

	Mean M_{gt}	Mean M_{pred}	Quantity	Mean Length
25kb	0.749	0.605	20088	367230.4
100kb	0.795	0.568	22897	347552.1

Our results were very conclusive on the testing dataset. Furthermore, comparing with our partner teams WA1 and WA2 revealed that our method outperformed their metrics.

Discussion and challenges

We re-implemented and/or called all the methods into a Python abstract class. Due to certain difficulties in incorporating everything into one clean pipeline as well as severely excessive runtime on methods like TADTree, we were unable to get the weights for all methods at both resolutions, which might have improved our consensus TADs even more.

References

[1] S. S. P. Rao *et al.*, 'A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping', *Cell*, vol. 159, no. 7, pp. 1665–1680, Dec. 2014, doi: 10.1016/j.cell.2014.11.021.

[2] 'Experiments – ENCODE'. https://www.encodeproject.org/experiments/ (accessed Jan. 25, 2022).