Detection of topologically associating domains with modified TopDom method

Julia Rymuza, Ewa Kizling, Anna Semik, Konrad Łukaszyk

January 2022

1 Method

Our approach is based on a well-known TAD detection tool TopDom. This method has been proven to detect TADs very efficiently. TopDom demonstrated robust and consistent TAD partitions, independent of the normalization and bin size. TopDom is originally written in R programming language. As part of

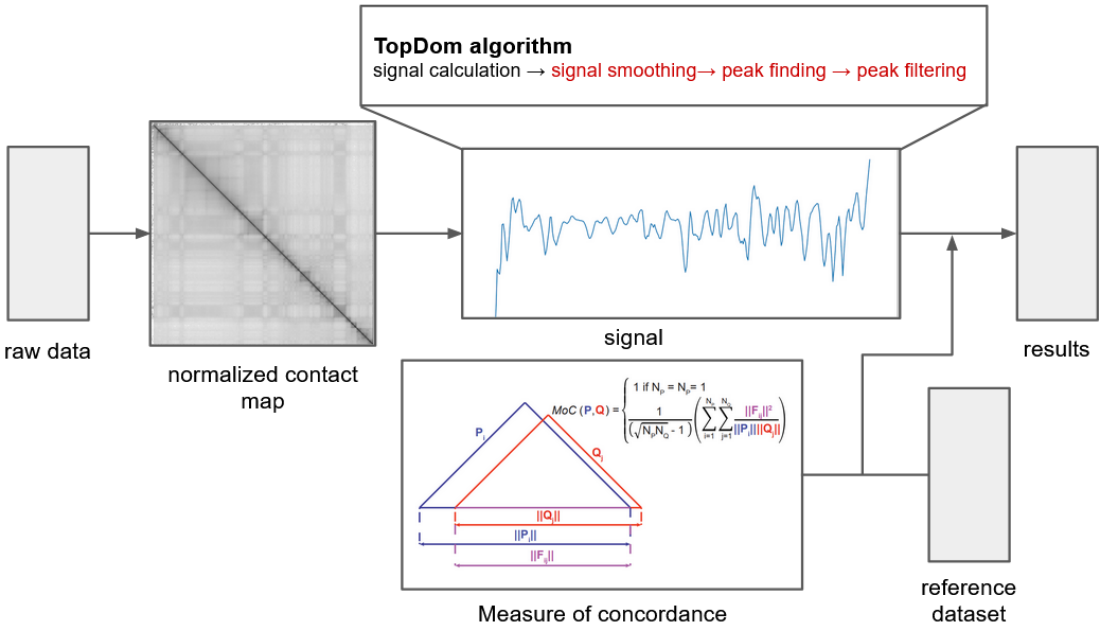

Figure 1: Proposed TADs detection pipeline.After normalization of raw data, algorithm calculates signal,smooths it, finds local minima and calculates TAD's boundaries. Last step consist of comparing results with reference dataset.

our project we reimplemented it in Python. TopDom algorithm consists of three steps:

1. generating binSignal
   TopDom is based on assumption that the contact frequencies between TADs is lower then inside one TAD. Since this step is essential for the TopDom algorithm we decided not to change it and in our approach we calculate the binSignal the same way.

2. detecting TAD boundaries
   TopDom expects to find TAD boundaries in binSignal local minima. It takes into consideration existence of noise, smooths the function and based on that detects boundaries.

3. statistical filtering
   after detecting TAD boundaries TopDom performs statistical filtering of false positives.The idea behind it is that interactions between two different TADs should be less frequent than between different upstream/downstream neighbor bins.

To improve TopDom's performance we introduced several modifications to TAD boundaries detection and statistical filtering steps.

TAD boundaries detection improvements
We implemented new methods for smoothing binSignal and finding local minima like: smoothing by Savitzky–Golay filter, smoothing by quadratic spline interpolation, finding minima by wavelet transformation, finding extrema by comparison of neighboring values, finding relative extrema of data.

Statistical filtering improvements
Beside statistical filtering introduced in TopDom we added possibility to filter peaks with scipy function peak_prominences. Only minima that are stand out from surrounding more than given threshold are kept

2 Results

To test our modifications we tried different combinations of functions to process data. We produced three different metrics to compare results of our approach with different parameters and golden standard : overlap score function from TopDom, which is an asymmetrical measure (so we use it bidirectionaly) and the Measure of Concordance (MoC).

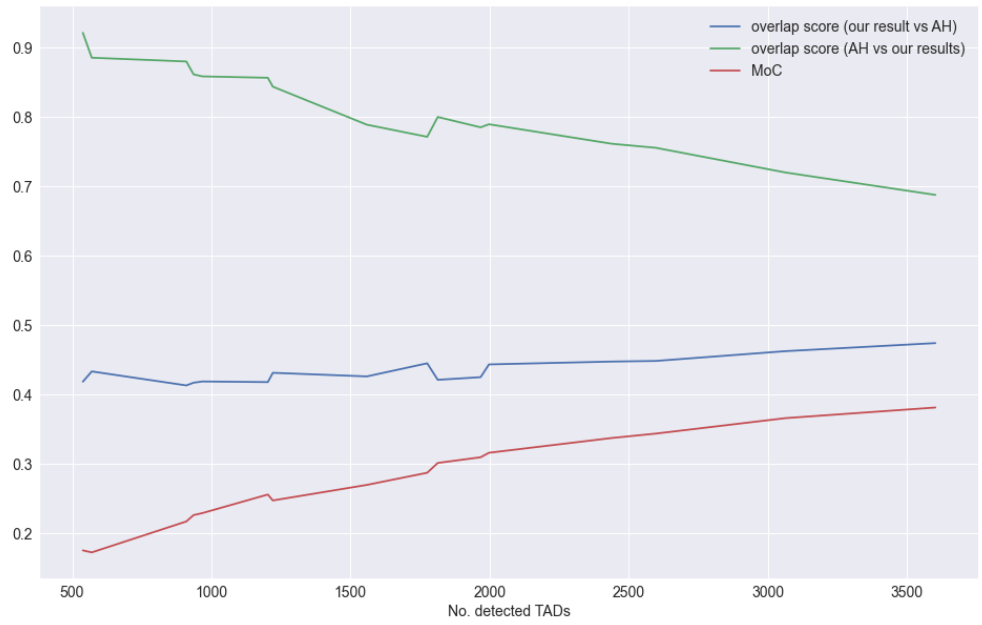

Figure 2: For combinations of functions producing more TADs the overlap between Arrowhead and them becomes higher. Moreover when one overlap metric increases the another one decreases.

Comparison to reference data
The results of overlapscore varies greatly (usually in range from 0.40 to 0.93) between direction of comparison, as it is not a symmetrical measure. The MoC measure implemented specially because of its symmetry returns even lower values (always less than 0.50). The reasons for this seem to be the differences both in TADs sizes and abundance between our results and reference datasets. It must be due to the shape of signal curve, we based our method on. TADs such short as those produced by Arrowhead cannot be obtained with our approach, even with all minimas considered as TAD boundaries .