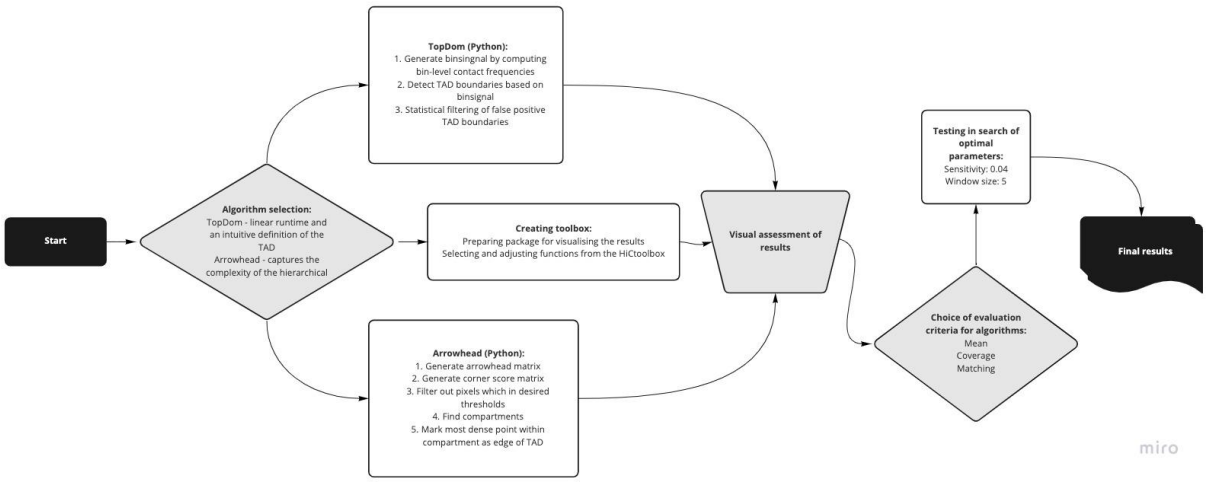
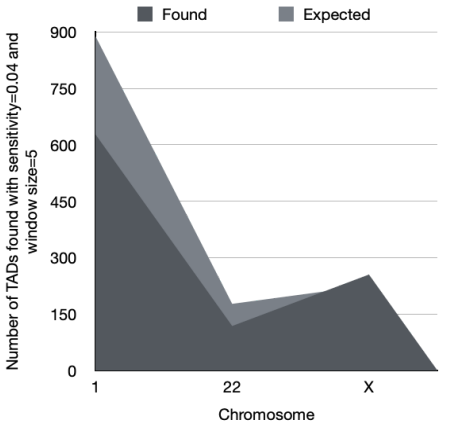
Prediction of TADs based on a redesigned TopDom algorithm

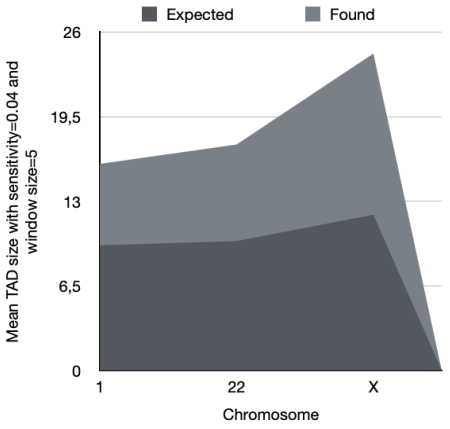
Sebastian Kot, Ignacy Makowski, Krzysztof Zdąbłasz, Leszek Troc

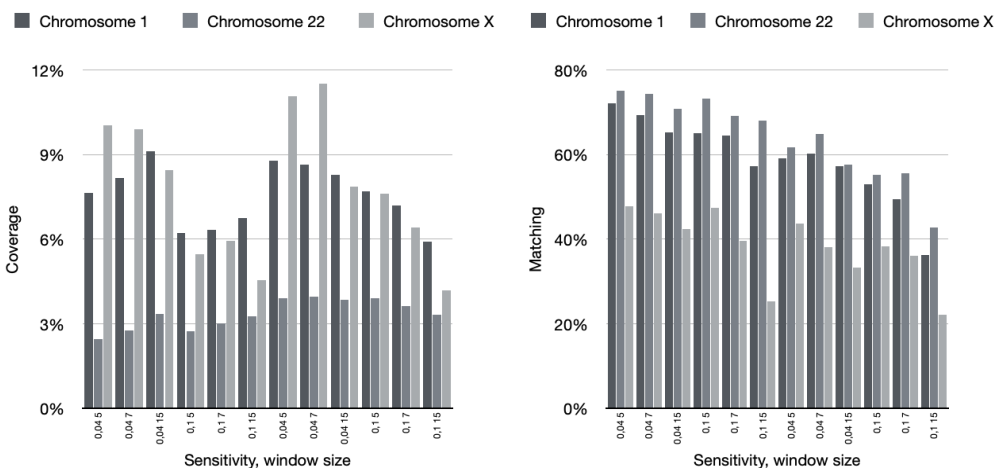
Workflow figure



Result figures







Methods

Our initial goal was simple. We decided to analyze two different algorithms in more depth, possibly modify them and for the final step, merge their results. The algorithms we chose are: Arrowhead, which captures the complexity of the hierarchical structure of Topologically Associating Domains (TADs), and TopDom, which is characterized by a linear runtime and intuitive definition of the TAD boundary.

However, when implementing the Arrowhead, we have encountered complex issues that disrupted our plans. Despite much tweaking and experimenting, the results achieved by our implementation were highly unsatisfactory regardless of the metric used for assessment. Therefore, we were forced to omit this part and focus on the TopDom. Arrowhead can still be run using Command Line Interface, but it is not recommended.

TopDom

The original TopDom consists of 3 stages:

1. Generate *binsingnal* by computing bin-level contact frequencies
2. Detect TAD boundaries based on *binsignal*
3. Statistical filtering of false positive TAD boundaries

TopDom depends on one parameter which is a window size. It is simply a side of a square that is generated for each bin on the diagonal of the Hi-C matrix. For each of these squares, the average contact frequency is calculated. Exact description can be found in the original paper.

We decided to redefine the second step and developed an entirely new approach compared to the original algorithm. Our goal was to give the users the ability to decide how precisely TADs are detected. Thus, we introduced the *sensitivity* parameter. It can be described as a value between 0 and 1, that is as a threshold, below which fluctuations in binsignal are recognized as noise in data. And as the algorithm is based on minima detection, hypothetical TADs that are smaller and blurred will also be ignored. With this option, the user can decide whether he wants to detect more clearly marked TAD boundaries or focus on higher precision.

Results

- Found less TADs than expected (mostly 2 to 5 times)
- Mean length of TADs was higher than expected (about 2 to 10 times)
- Best results were achieved with window size = 5 and *sensitivity* = 0.04
- The number of TADs decreases when window size and *sensitivity* increase
- The mean length of TADs increases when window size and *sensitivity* increase

- The mean length of partner team's TADs is consistently about 50% higher than ours

Discussion

This project increased our understanding about the Topologically Associated Domains, particularly what they are, why they can be useful and how to go about detecting them in the Hi-C data. Our first challenge was fully understanding the task at hand since TADs are not precisely defined in literature. Information on this subject is relatively new and some time was needed to familiarize ourselves with the topic.

As for the results, it is quite clear that the number of TADs we found is generally smaller than expected. This may be caused by the fact that TopDom is not able to detect nested TADs as some other algorithms. However, one more important factor needs to be considered when assessing the results. The number of found domains is highly correlated with the data resolution, regardless of the algorithm. We could not determine the resolution at which the golden data were generated. Hence, a certain amount of distrust needs to be maintained.

The modified version of TopDom seems to perform well and for certain combinations of parameters is able to return reliable results. One of its biggest advantages is linear execution time, which makes modifications more viable. Of course, there is room for improvement. For example, our implementation could benefit from multiple binsignal flattening with different values of *sensitivity* parameter, which would allow recognizing nested TADs in the analyzed data, maintaining linear complexity.

To summarize, thanks to this project, we were able to improve our interdisciplinary knowledge and dive into the most recent bioinformatics problems. Despite not fulfilling all of our initial goals, we are satisfied with the results, and see many possible ways to improve our work in the future.