**Marie Zufferey**

Department of Computational Biology

University of Lausanne

***Data Science Project***

**Prediction of TAD boundaries from epigenetics data**

***Conceptual Design Report***

**May 2020**

# ABSTRACT

10-20 lines

Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper.

# 

# 

# TABLE OF CONTENTS

...

# OBJECTIVE

0.5-1.0 page

Formulate goal and purpose of your project here.

In the recent years, several molecular assays have been developed with the aim of studying the tridimensional organization of the genome. In particular, chromosome conformation capture technology followed by high-throughput sequencing (Hi-C; Liebermann-Aiden 2009) is a technique that can quantify the frequencies of interactions between any two loci of the genome.

The analysis of such data reveals the existence of topologically-associating domains (TADs; Nora et al. 2012), which represent linear regions of the genome with higher frequencies of self-interactions, while being depleted of contacts with the adjacent ones.

The boundaries of TADs hence represent specific regions depleted of contacts where the switch between preferential upstream (resp. downstream) interactions occur. Notably, it has previously been shown that these regions of the genomes are enriched in specific features such as specific histone marks, CTCF binding sites, housekeeping or tRNA genes (e.g. Dixon et al. 2012).

The aim of this project is to develop a neural network model to predict the “boundary score” of a genomic bin (i.e. a fixed-size genomic interval) based on epigenetics data.

# METHODS

0.5-1.0 page

Which infrastructure, tools, software libraries, statistical methods etc will be used

*Data preparation (pre-processing)*

For the analysis of chromosome conformation capture data, the genome is traditionally partitioned into genomic bin (i.e. a fixed-size genomic interval).

For the chromatin interaction data, I will use publicly available datasets (already normalized).

In order to detect TADs from Hi-C data, several algorithmic approaches have been developed (Forcato 2017). In this project, I will use TopDom (Shin). In particular, this method is particularly useful for my analyses, as it assigns a so-called “boundary score” to any bin of the genome. This value corresponds to the p-value of a Wilcoxon rank-sum test that quantifies the shift in upstream and downstream interactions. This will be the “target variable” that I would like to predict with my neural network model.

TopDom (Shin et al. 2016) proceeds in three steps. First, it computes a score called binSignal that represents, for each bin, an average of the counts between upstream and downstream regions around it. This metric is expected to be low at TAD boundaries and high in mid-position of the domains. Next, it detects binSignal minima (i.e. candidate TAD boundaries) using a piecewise linear function. In the final step of the algorithm, the list of boundaries is refined by retaining those that represent regions with significantly depleted interactions between upstream and downstream intervals, compared to contact counts within these intervals, as assessed with a Wilcoxon rank sum test.

As for the predictive variables, I will used publicly available ChIP-seq data from the ENCODE portal (Sloan et al. 2016; <https://www.encodeproject.org>). I selected data for 56 epigenetic marks (histone marks, transcription factors, binding proteins, etc.).

*Statistical methods*

Before building machine learning models, some basic exploratory analyses will be performed (e.g. correlations, linear models, data description).

Predicting the scores for a sequence of bins.

I will build a convolutional neural network (CNN) model. The architecture of the network will probably need some tuning, but I plan to include several convolutional layers with several convolutional kernels and, optionally, one or several max-pooling layers. :

As the surrounding of the bin is assumed to be important for the prediction of the boundaries, the model will process as a kind of sliding window by taking into account the information for the *n* bins (n=100).

As each chromosome can be considered as a dataset, a subset of the chromosomes will be used for training, and the other for testing.

•

*Tools, softwares and libraries*

The basic exploratory analyses will be conducted using Python (XX modules).

For the chromatin interaction data, the program “dump” from the Juicebox toolbox (2016) is used to download normalized Hi-C data. The boundary score for each of the genomic bin is retrieved from TopDom outputs (Shin et al. 2016; distributed as an R package from Bengtsson and Shin 2018).

The ChIP-seq data are directly downloaded from the ENCODE portal from command line (“wget” command). The signal values are aggregated to genomic intervals (bins) using the programm “map” from the BEDTools suite (version 2.27.1; Quinlan and Hall 2010).

For a given epigenetic mark, several datasets can be available. Thus, an additional pre-processing step consists in aggregating the signal values in such a way that each column corresponds to a mark. For this purpose, we used a custom R script and the “rowSums” from the *base* package (R citation). Note that summing the signal values from several experiments for a subset of marks is not problematic here as each of the signal values are then normalized across the bins for each mark separately before the machine learning process.

).

The BEDTools suite

I USED ONLY THE BROAD PEAKS DATA !!!

*Infrastructures*

d

# DATA

Which data will be used (with references) 0.5-1.0 page

A couple of plots, maybe the histograms of the columns

A couple table row as an example maybe

Security issues etc (see data management plan, you may attach a SNSF data management plan for your data)

# METADATA

What metadata is required for reproducing your analysis.

Where do you store the metadata, how can people access it.

A README file with the list of the URLs of the downloaded data as well as the scripts used for pre-processing the input data are deposited on a GitHub repository (<https://github.com/marzuf/CAS_ADS/tree/master/CAS_2020_M1>). This folder also contains Jupyter notebook to reproduce the figures presented in the current report.

# DATA QUALITY

What are the quality requirements you have to reach your analysis goal (precision ...)?

Are they met? If not, do you expect a significant impact on your results,

Any measures to improve the quality?

The GM12878 Hi-C data are to date the highest quality (human) Hi-C data available. As for the ENCODE database, it is well-established in the field of biology. If technical/experimental biases or failures can never be excluded, we will assume that the quality of the data is satisfactory. Also, to assess the robustness of our results, we can consider to reproduce our analyses with the Hi-C data from other cell line(s), and/or with different subset(s) of the epigenetics data.

# DATA FLOW

# <FIGURE>

# 

# DATA MODELS

The pre-processing steps are not shown in the data models (cf. METHODS section). These latter include on the one hand preparing the Hi-C data and running TopDom to obtain the boundary scores. On the other hand, ChIP-seq data are processed to obtain signal values for each of the epigenetics mark.

*Conceptual data model*

<FIGURE>

*Logical data model*

<FIGURE>

*Physical data model*

<FIGURE>

The raw Hi-C data files can be as heavy as 1.5 Gb (per chromosome, depending on chromosome length), and the output of TopDom up to ca. 1 Gb by chromosome.

The order of magnitude of the size of the BED files downloaded from ENCODE is of several hundreds of kilobases. For the analyses presented here, 116 of such files were downloaded.

The neural network models will be run using Colaboratory notebooks on the Google cloud platform (https://colab.research.google.com) and therefore do not require expensive computational infrastructures.

# RISKS

What can go wrong?

When this and that goes wrong, what measures do you have?

What will be the impact?

The raw Hi-C and ChIP-seq files represent a considerable amount of data. Therefore, we decided to store only the final processed data. We expect that the raw data will “always” remain available on the original websites, but we cannot guarantee it.

# PRELIMINARY STUDIES

From module 2

# CONCLUSIONS

This document provides the basis

# REFERENCES

Bengtsson H. and Shin H. 2018. TopDom: An Efficient and Deterministic Method for Identifying Topological Domains in Genomes. R package version 0.5.0. <https://github.com/HenrikBengtsson/TopDom>

Dixon J. R. et al. 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*. Vol. 485, pp. 376–380.

Forcato M. et al. 2017. Comparison of computational methods for Hi-C data analysis. *Nature Methods*. Vol. 14, pp. 679–685.

Quinlan A. R. and Hall. I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. Vol. 26(6), pp. 841–842.

Shin H. et al. 2016. TopDom: an efficient and deterministic method for identifying topological domains in genomes. *Nucleic Acids Research*. Vol. 44(7), p. E70.

Sloan C. A. et al. 2016. ENCODE data at the ENCODE portal. *Nucleic Acids Research*. Vol. 44(D1), pp. D726–D732.

# Supplementary materials

The scripts used for data pre-processing and for the analyses presented in this document are available on GitHub (<https://github.com/marzuf/CAS_ADS/tree/master/CAS_2020_M1>).