Marie Zufferey
Department of Computational Biology
University of Lausanne
marie.zufferey.1@unil.ch

Data Science Project

Prediction of TAD boundaries from epigenetic data

Conceptual Design Report

May 2020

ABSTRACT

Recently developed molecular techniques allow to quantify the frequency of interactions between any two loci of the genome. Analysis of such data reveals the presence of regions of the chromatin with high level of internal contacts while being depleted of interactions with the adjacent regions, so-called topologically associating domains (TADs). These domains are demarcated by boundaries characterized by specific features such as an enrichment of housekeeping genes or tRNA and decorated with distinct histone marks. The aim of this project is to build a model that will allow to predict the probability that a genomic region is a TAD boundary based on its epigenetic signature. To this end, we will use publicly available chromatin interaction and epigenetic data from the GM12878 cell line and build a convolutional neural network model. In this document, we present an overview of the different steps that we will conduct, with a focus on the description of the input data. Eventually, some explanatory analyses are also presented.

TABLE OF CONTENTS

OBJECTIVE

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 allows to 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 global aim of this project is to build a neural network to predict the “boundary score” of a genomic bin (i.e. a fixed-size genomic interval) based on epigenetic data. In this conceptual design report, we pursue more specifically the objective of discussing methodological considerations and preliminary statistical analyses. In particular, we will describe steps that precede the development of the machine learning model, including data acquisition and preprocessing as well as an exploratory examination of the dataset.

METHODS

Data preparation: from Hi-C data to boundary scores

For the analysis of chromosome conformation capture data, the genome is traditionally partitioned into genomic bins (i.e. a fixed-size genomic interval). The size of the bin is arbitrarily fixed but mostly depends on the quality of the assay: smaller bin size can be used only if the sequencing is deep enough. In this project, we use Hi-C data binned at 25 kilobases (kb) from the GM12878 cell line (i.e. the first bin covers the first 25’000 base pairs of a given chromosome, the second bin from the 25’001-st to the 50’000-th base pairs, and so on). In most cases, these data are stored, for each chromosome separately, either in a dataframe of three columns where the value of the third column indicates the frequencies of interactions (“count value”) between the loci given in the first two columns (cf. Fig. 1 in DATA section); or they can be stored in more intuitive symmetric square matrix where the i,j-cell stores the interaction frequencies between the loci of the i-th row and j-th column (cf. Fig. 2 in DATA section). The data made publicly available were stored according to the former format and we had to convert them into the latter format for the following step (custom R script; see Supp. mat.). In addition, the downloaded data are already

normalized, therefore we do not discuss here the Hi-C data normalization procedure (see e.g. Imakaev et al. 2016).

In order to detect TADs from Hi-C data, several algorithmic approaches have been developed (Forcato et al. 2017). In this project, we use TopDom (Shin et al. 2016). In particular, this method has the advantage to assign a so-called "boundary score" to any bin of the genome that can be easily retrieved from one of the output files (the one with the ".binSignal" suffix; cf. Fig. 3 in DATA section). This value corresponds to the p-value of a Wilcoxon rank-sum test that assesses statistically the depletion of interactions between upstream and downstream intervals, hence putative boundary regions. And this p-value will serve as the "target variable" that we aim at predicting at the end with the neural network model.

Data preparation: from epigenetic data to signal values

As for the predictive variables, the epigenetic marks, we use publicly available ChIP-seq data for the GM12878 cell line from the ENCODE portal (Sloan et al. 2016; https://www.encodeproject.org). We selected data from 116 experiments ("broad peaks" data in BED format; cf. Fig. 4 in DATA section), corresponding to 56 distinct epigenetic marks (histone marks, transcription factors, binding proteins, etc.). For the marks for which several datasets were available, an additional preprocessing step consists in aggregating (summing up) the signal values in such a way that each column corresponds to a distinct mark (custom scripts; see Supp. mat.). Note that adding up the signal values from several experiments for a subset of marks is not problematic here as each of the signal values will be normalized across the bins for each mark separately before the machine learning process. Finally, the signal values are aggregated (summed up) to the corresponding 25 kb bins of each chromosome (e.g. for each mark, all the signal values that map to any of the first 25000 base pairs of a chromosome will be mapped to the first bin of this chromosome).

Statistical methods

Before focusing on machine learning models, some basic descriptive and exploratory analyses are performed (e.g. distribution, correlation, linear model). We will then build a convolutional neural network model to predict scores for a contiguous sequence of bins from the corresponding epigenetic signals. The architecture of the network will probably need some tuning, but we plan to include several convolutional layers with several convolutional kernels and, optionally, one or several max-pooling layers. As each chromosome can be considered as a dataset, a subset of the chromosomes will be used for training, and the other for testing. The details of the neural network model will be further discussed in Module 3.

Tools, softwares and libraries

For the chromatin interaction data, the program “dump” from the Juicer toolbox (Durand et al. 2016) is used to download normalized Hi-C data. The boundary score for each of the genomic bins 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 (“curl” 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).

Preparation of the data was conducted from the terminal and on R (R Core Team 2018) using custom scripts (cf. Supp. mat.). The basic exploratory analyses presented below were performed using Python and the following modules: pandas (McKinney 2010), re, numpy (van der Walt et al. 2011), scipy.stats (Pedregosa et al. 2011), statsmodels.api (Seabold and Perktold 2010). The matplotlib (Hunter 2007) and seaborn (Waskom et al. 2014) libraries were used for plotting. Finally, the neural network model will also be built in Python using the tensorflow library (Abadi et al. 2015) and executed on the Colaboratory platform (https://colab.research.google.com).

Infrastructures

A commercially available laptop was used as data storage and analysis infrastructure (Lenovo Thinkpad T560, Intel® Core™ i7-6600U, 2.60GHz × 4 Cores, 8 GB RAM, 256 GB hard disk, with a Linux Ubuntu 16.04 LTS operating system).

DATA

Chromatin interaction data and boundary scores

The Hi-C data are downloaded using “dump” (Fig. 1) and, converted in a format accepted by TopDom (Fig. 2). We retrieve the “.binSignal” file from TopDom outputs (Fig. 3) and extract the boundary scores. The final data frame here is of dimension (n_bin x 1), where n_bin is the total number of bins of a given chromosome. These steps have to be repeated for each of the chromosomes.

binA	binB	count
9400000	9425000	5400.409
9425000	9425000	22256.86
9400000	9450000	2780.6067
9425000	9450000	4916.144
9450000	9450000	18140.729

Figure 1: first rows of the Hi-C data as downloaded using the Juicer tool for chromosome 21. The file stores the count value (third column) between a pair of genomic bins (whose positions are indicated in the first two columns).

chr21	0	20000	0	0
chr21	20000	40000	0	0
chr21	40000	60000	0	0
chr21	60000	80000	0	0
chr21	80000	100000	0	0

Figure 2: first rows of the Hi-C data formatted for TopDom for chromosome 21. This corresponds to a three-column data frame (first column: chromosome, second column: bin start, third column: bin end) appended to a symmetric numeric matrix of pairwise contact counts.

id	chr	from.coord	to.coord	local.ext	mean.cf	pvalue
1	chr21	0	20000	-0.5	0	1
2	chr21	20000	40000	-0.5	0	1
3	chr21	40000	60000	-0.5	0	1
4	chr21	60000	80000	-0.5	0	1
5	chr21	80000	100000	-0.5	0	1

Figure 3: first rows of the ".binSignal" output from TopDom from which the boundary score is retrieved (seventh column). The three first columns give the genomic positions.

Epigenetic data and signal values

The epigenetic data are downloaded in BED format from ENCODE (Fig. 4). The signal values are next extracted, mapped to the genomic bins (aggregated by summing). Finally, if several datasets are available for a same mark, they are summed up. As there are here 56 distinct epigenetic marks, the final table (one per chromosome) is of dimension (n_bin x 56).

chrom	chromStart	chromEnd	name	score	strand	signalValue	pValue	qValue
chr21	9696016	9696252	.	871	.	15.717358	12.6	-1
chr21	9696100	9697137	.	366	.	3.887988	1.8	-1
chr21	9696101	9696350	.	836	.	14.896773	12.1	-1
chr21	9820498	9934664	.	269	.	1.614630	13.5	-1
chr21	9824586	9828540	.	1000	.	22.229185	14.7	-1

Figure 4: first rows of the raw data downloaded from ENCODE (BED format). It is the seventh column ("signalValue") that is extracted for upstream analyses.

Formatted dataset

For each chromosome, the data frames of boundary scores and signal values are merged (the jointure is done based on the genomic bin). The resulting dataset is then a (n_bin x (1+56)) data frame containing numeric values (Fig. 5).

binScore	CTCF	EZH2	H2AFZ	H3K4me1
0.274411082176029	214.440683	66.442082	23.843815	22.832315
0.0342406882240031	46.493788	0	1.61463	0
0.0133944859241037	44.122768	3.624789	1.61463	11.022849
0.10407917661042	59.999873	2.820925	1.61463	2.124582
0.365480704274996	166.417762	12.30813	1.61463	2.124582

Figure 5: first rows of the formatted data for chromosome 21, that serves as input for the rest of the analyses.

We provide here below a short overview of the distribution of the variables that we will use in our analyses (Fig. 6). For sake of convenience, we selected data for chromosome 21 and a subset of epigenetic marks only. Notably, we notice the presence of a high number of zero values for the epigenetic marks (Fig. 7). This calls for caution, and we will need to treat them carefully in the upcoming statistical models, and decide whether to remove them or not. Moreover, the distribution of the response variable (boundary score) seems to have a bimodal shape. Finally. we

note that, after the removal the zero values and a log10-transformation, the signal values are almost normally distributed for the H3K4me1 mark and bimodal for the H2AFZ mark (Fig. 8).

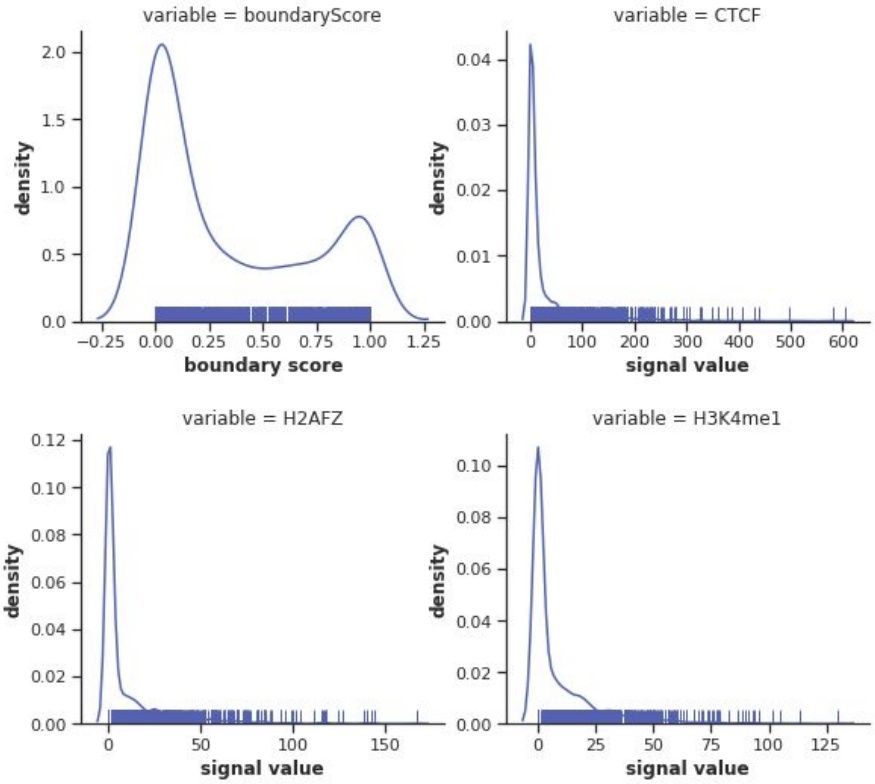

Figure 6: density plot showing the distribution of the dependent variable ('boundaryScore') and a subset of the independent variables (CTCF, H2AFZ and H3K4me1).

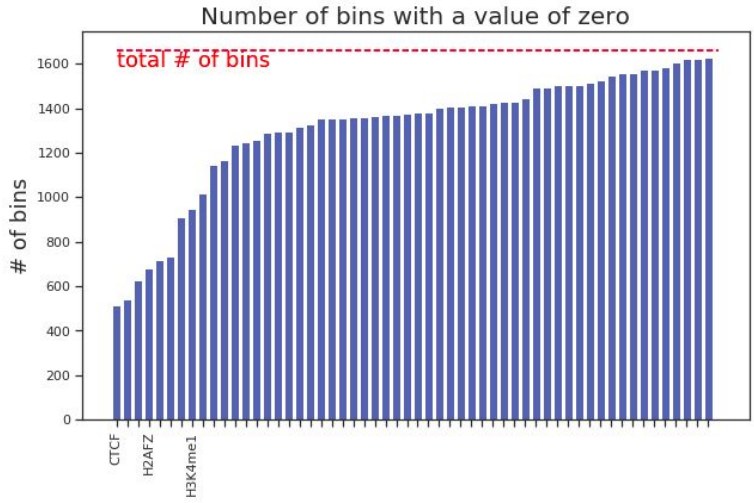

Figure 7: barplot showing the number of zeros in the signal values (y-axis) for all the 56 epigenetic marks (x-axis). Explicitly indicated on the x-axis are the subset of the marks for which we show the results of the explanatory analyses in a subsequent section.

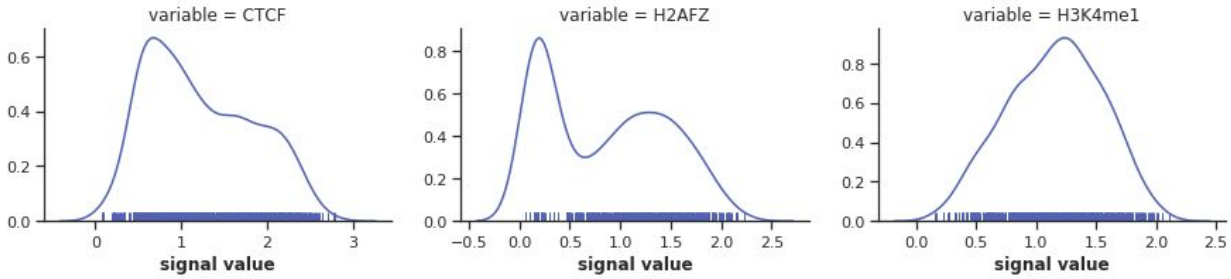

Figure 8: density plot showing the distribution of a subset of the independent variables (CTCF, H2AFZ and H3K4me1) after removal of zero values and log10-transformation.

METADATA

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

DATA QUALITY

The GM12878 Hi-C data from Rao et al. (2014) are to date the highest quality (human) Hi-C data available. As for the ENCODE portal, it is a well-established database 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 epigenetic data.

DATA FLOW

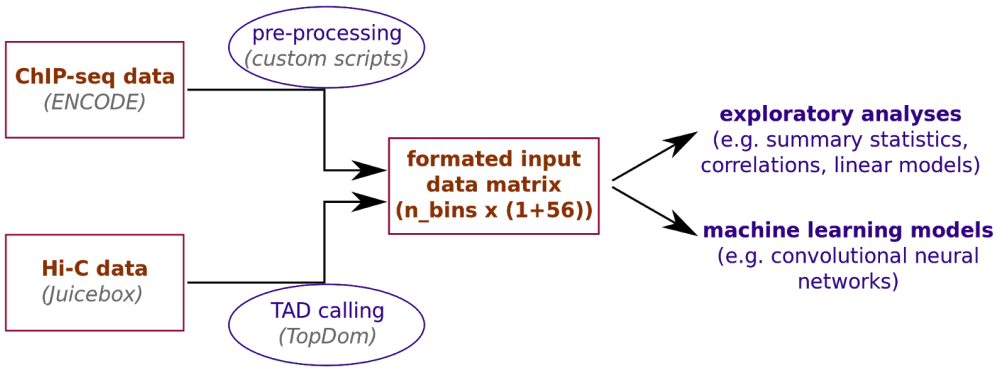

Figure 9: schematic data flow of this data science project.

DATA MODELS

The preprocessing steps (prepare the Hi-C data and run TopDom to obtain the boundary scores, derive signal values from ChIP-seq data for the epigenetic marks) are not shown in the data models (cf. METHODS and DATA sections).

Conceptual data model

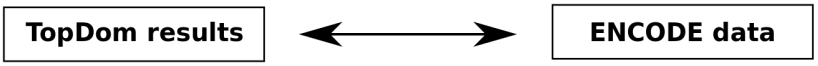

Figure 10: conceptual data model of this data science project.

Logical data model

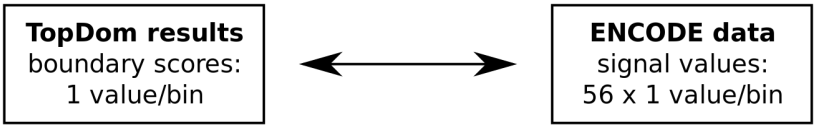

Figure 11: logical data model of this data science project.

Physical data model

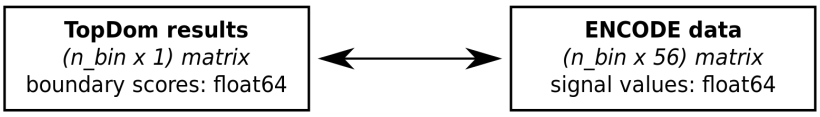

Figure 12: physical data model of this data science project.

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 kilobytes. 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

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

PRELIMINARY STUDIES

For exemplary purposes, we show here the results of some preliminary analyses obtained for chromosome 21 and three epigenetic marks of different types: CTCF - a binding protein, H3K4me1 - a histone mark and H2AFZ - a variant histone. First, we investigate the relationships among pairs of variables with correlation analyses (Fig. 13). For these analyses, we discard the genomic bins for which the signal values for all the subset of epigenetic values were equal to zero, and log10-transform the signal values. Overall, we observe limited correlation.

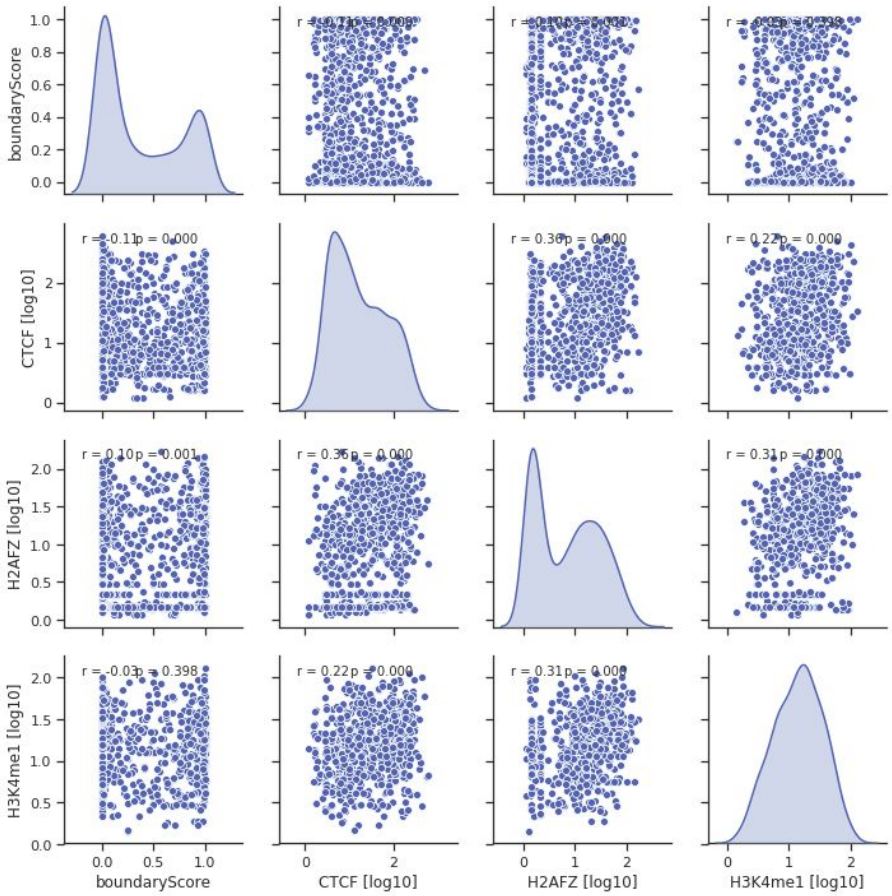

Figure 13: pairwise Pearson’s correlation and distribution (diagonal) of the target variable (boundaryScore) and a subset of the predictor variables (CTCF, H2AFZ and H3K4me1). The rows where all variables were equal to 0 have been discarded.

We then assess the explanatory potential of each variable independently with a simple linear model, with the boundary score as the response variable and a single epigenetic mark, one at a time, as the independent variable (Fig. 14). For these analyses, we discard the genomic bins for which the signal values was equal to zero before applying a log10-transformation to the signal values. Only the effects of CTCF and H2AFZ on the boundary scores was statistically significant ($p <= 0.001$).

OLS Regression Results

Dep. Variable:	boundaryScore	R-squared:	0.013
Model:	OLS	Adj. R-squared:	0.012
Method:	Least Squares	F-statistic:	14.86
Date:	Sun, 17 May 2020	Prob (F-statistic):	0.000122
Time:	13:48:08	Log-Likelihood:	-542.81
No. Observations:	1147	AIC:	1090.
Df Residuals:	1145	BIC:	1100.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const.	0.4576	0.025	17.997	0.000	0.408	0.507
CTCF	-0.0713	0.019	-3.855	0.000	-0.108	-0.035

OLS Regression Results

Dep. Variable:	boundaryScore	R-squared:	0.011
Model:	OLS	Adj. R-squared:	0.010
Method:	Least Squares	F-statistic:	10.46
Date:	Sun, 17 May 2020	Prob (F-statistic):	0.00126
Time:	13:48:09	Log-Likelihood:	-482.60
No. Observations:	982	AIC:	969.2
Df Residuals:	980	BIC:	979.0
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const.	0.3304	0.022	15.164	0.000	0.288	0.373
H2AFZ	0.0673	0.021	3.234	0.001	0.026	0.108

OLS Regression Results

Dep. Variable:	boundaryScore	R-squared:	0.001
Model:	OLS	Adj. R-squared:	-0.000
Method:	Least Squares	F-statistic:	0.7147
Date:	Sun, 17 May 2020	Prob (F-statistic):	0.398
Time:	13:48:10	Log-Likelihood:	-381.71
No. Observations:	716	AIC:	767.4
Df Residuals:	714	BIC:	776.6
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const.	0.4800	0.048	10.081	0.000	0.387	0.574
H3K4me1	-0.0330	0.039	-0.845	0.398	-0.110	0.044

Figure 14: summary of the linear models fitted for each of the epigenetic mark separately (top: CTCF, center: H2AFZ, bottom: H3K4me1).

We observe that, individually, these epigenetic marks have a poor predictive potential ($R^2 < 0.05$). Therefore, we next assess their combined effects with a multivariate linear regression model (Fig. 15). For these analyses, we discard the genomic bins for which the signal values for all the subset of epigenetic values were equal to zero. Before applying a log10-transformation of the signal

values, the remaining zeros were then replaced with a small offset (0.01; note that the lowest non-zero signal value is 1.17) to avoid infinite values. This leads to a slightly improved, but still limited, percent of explained variance. Interestingly, the effect of H3K4me1 is now statistically significant ($p <= 0.01$). Inversely, the effect of H2AFZ is in this case not statistically significant ($p > 0.01$). This might be caused by the correlation between H2AFZ and H3K4me1 signal values: when the multivariate model is built with only a pair of variables, the two variables are statistically significant in all cases, except when H2AFZ and H3K4me1 are considered together (not shown here; see also correlation in Fig. 13).

OLS Regression Results

Dep. Variable:	y	R-squared:	0.049
Model:	OLS	Adj. R-squared:	0.047
Method:	Least Squares	F-statistic:	23.24
Date:	Sun, 17 May 2020	Prob (F-statistic):	1.14e-14
Time:	11:36:29	Log-Likelihood:	-604.13
No. Observations:	1357	AIC:	1216.
Df Residuals:	1353	BIC:	1237.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.4172	0.013	32.709	0.000	0.392	0.442
CTCF [log10]	-0.0377	0.008	-4.480	0.000	-0.054	-0.021
H2AFZ [log10]	-0.0026	0.009	-0.296	0.768	-0.020	0.015
H3K4me1 [log10]	0.0566	0.008	7.071	0.000	0.041	0.072

Figure 15: summary of the multiple regression model including a subset of the epigenetic marks as predictor variables (CTCF, H2AFZ and H3K4me1) and the boundary score as target variable.

CONCLUSIONS

The analyses conducted for this report allowed us to familiarize with the data on hand and raised awareness of potential pitfalls. Consequently, this document lays the foundations for more thorough analyses. From the work presented here, we conclude that simple linear models do not lead to satisfactory results. Considering the distribution of the input data, applying different categories of linear models (e.g. multiple logistic regression as in Mourard and Cuvier 2016) might be more suitable. This will be achieved in a future study. Afterwhile, we will specifically focus on building a deep neural network. Indeed, recent work suggests that such models could be powerful to predict chromatin structure from epigenetic marks (Rozenwald et al. 2018). As adjacent genomic positions are not independent from one another, we also believe that including epigenetic signals from the surrounding bins, as a convolutional neural network allows, could improve the predictive power of the models.

REFERENCES

Abadi M. et al. 2015. *TensorFlow: Large-scale machine learning on heterogeneous systems*. Software available from tensorflow.org.

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.

Durand N. C. et al. 2016. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Systems*. Vol. 3(1).

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

Hunter J. D. 2007. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*. Vol. 9(9), pp. 90-95.

Imakaev M. et al. 2012. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature Methods*. Vol. 9(10), pp. 999-1003.

Lieberman-Aiden E. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*. Vol. 326(5950), pp. 289-293.

McKinney W. 2010. Data structures for statistical computing in Python. *Proceedings of the 9th Python in Science Conference*. Vol. 445, pp. 56-61.

Mourad R. and Olivier Cuvier. 2016. Computational identification of genomic features that influence 3D chromatin domain formation. *PLoS Computational Biology*. Vol. 12(5), p. E1004908.

Nora E. P. et al. 2012. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*. Vol. 485, pp. 381–385.

Pedregosa, F. et al. 2011. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*. Vol. 12, pp. 2825-2830.

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

R Core Team. 2018. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*. Vienna, Austria. https://www.R-project.org.

Rao S. S. P. et al. 2014. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*. Vol. 159(7).

Rozenwald M. et al. 2018. Prediction of 3D chromatin structure using recurrent neural networks. *IEEE International Conference on Bioinformatics and Biomedicine*.

Seabold S. and Perktold J. 2010. statsmodels: econometric and statistical modeling with Python. *Proceedings of the 9th Python in Science Conference*. 2010. Vol. 12(1), pp. 92-96.

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.

van der Walt S. et al. 2011. The NumPy array: a structure for efficient numerical computation. *Computing in Science and Engineering*. Vol. 13 (2), pp.22-30.

Waskom M. et al. 2014. seaborn: v0.5.0. https://doi.org/10.5281/zenodo.12710.

Supplementary materials

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