Faculty of Computer Science, HSE,
Skoltech, data science

Zybin Mikhail Aleksandrovich

On correspondence between TAD boundaries and normalized methylation entropy blocks in human genome

1st year MS thesis

HSE scientific supervisor:
Candidate of Sciences, Assistant Professor
Alexey Naumov

Skoltech scientific supervisor:
Candidate of Sciences, Assistant Professor
Ekaterina Khrameeva

Moscow 2021

Abstract

The overall goal of our research is to predict TAD boundaries using methylation data. In this work, we present a report on a preliminary step towards this goal. Namely, we describe our attempt to reproduce the paper in which the authors generate methylation entropy data and propose that there is a correspondence between the entropy and TAD boundaries.

1 Introduction

1.1 Chromatine structure

Chromatin is a complex of DNA and proteins. In prokaryotes, chromatin is located in the nucleus, where it takes some specific 3d shape called chromatin conformation. This shape gives rise to DNA-DNA interactions, which are a relevant subject to a biological study. What's more, this 3d conformation can be thought of as consisting of parts called Topologically Associated Domains (TADs), where the molecule physically interacts with itself more frequently than with sequences outside the TAD [1].

DNA as a one-dimensional sequence has several important features such as openness and methylation level. When DNA folds into a 3D structure, some parts of DNA are located relatively inside the complex, while others are more close to the surface. This characteristic of being close to the surface is called openness, and the openness of DNA fragments affects the level of its expression. There are also various other types of data and experimental protocols of their gathering, such as RNA-seq and ATAC-seq, which will not be discussed in this report.

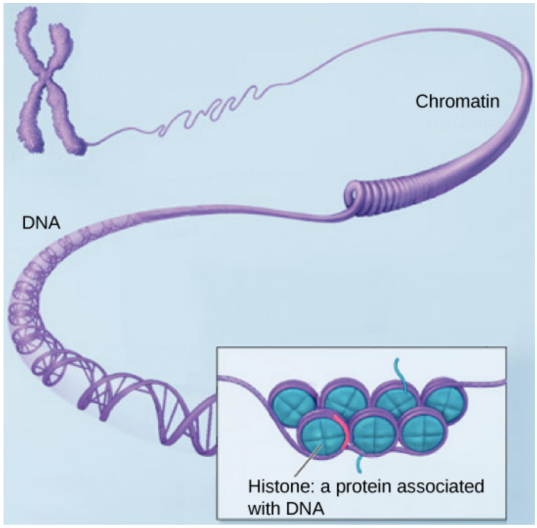


Figure 1: Schematic chromatin structure

A genetic "alphabet" consists of 4 "letters" - nucleotides Adenine (A), Thymine (T), Cytosine (C), and Guanine (G). A binds to T, C binds to G.

1.2 Methylation

Cytosine can undergo the process of so-called methylation and transform into 5-methylcytosine, which is different from cytosine in having a methyl group attached to the

fifth atom of the ring. DNA maintains the same sequence, but the expression of methylated genes can be altered. This is a very notable feature, because, although methylation does not modify genetic code, it can be inherited. Methylation plays role in the mechanisms of aging and cancer. DNA methylation is an important epigenetic modification that plays critical role in cellular differentiation, development, and disease.

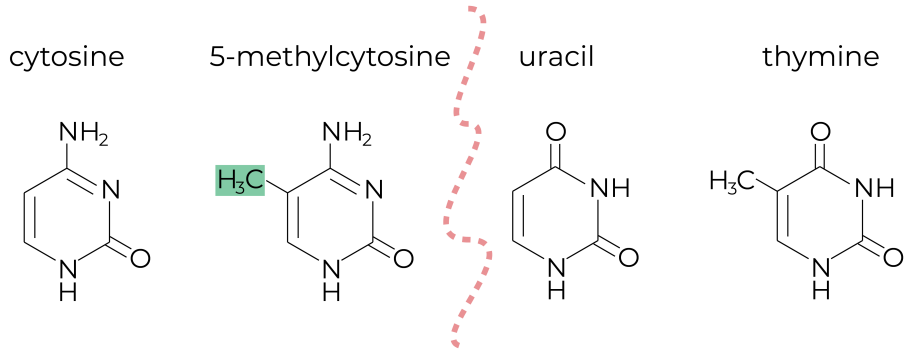


Figure 2: The chemical structures of cytosine, 5-methylcytosine, uracil and thymine. 5-methylcytosine is commonly referred to as DNA methylation.

CpG site is a region of DNA where a cytosine nucleotide is followed by a guanine nucleotide. There are 3 types of cytosine methylation: CpG, CHG, CHH (H denotes either A, T, or G).

Several methods of methylation profiling exist. We will explain the Whole Genome Bisulfite Sequencing method because the data obtained with this method is used in our research. WGBS family of experimental protocols consists of two steps [2] [3]. Firstly, after denaturation, the DNA is treated with sodium bisulfite, which acts on the DNA molecule by transforming unmethylated cytosine into uracil, and leaving methylated cytosine intact. Secondly, the PCR amplification step transforms uracil into thymine and methylated cytosine into cytosine. As a result, the reads have cytosines left only in those places where it was originally methylated. Advancements in NGS technologies gave rise to BS-seq [4] [5], a protocol with higher quality than the previous versions.

2 Theory

2.1 Overview

This research is largely based on the papers [6] [7] [8] and the software implementation of them. In these papers, the authors analyze the methylation data obtained from the whole-genome bisulfite sequencing (WGBS) experimental protocol. They apply the principles of information theory (Shannon entropy) and statistical physics (1d Ising Model) and manage to predict the TAD boundaries. The data they use come from human cells from several healthy organs and organs with cancer. There are 17 cell types in total, and for each type 11 files are provided.

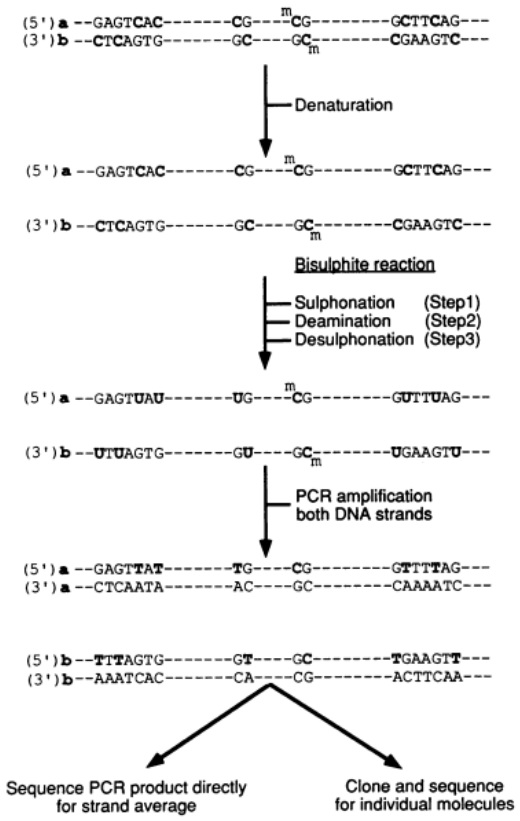


Figure 3: Bisulphite genomic sequencing procedure

2.2 1D Ising model

1D Ising model consists of a line of N points, each of which interacts only with its two (or one) neighbors [9]. Originally it is used to describe ferromagnetism, but it can also be applied in many other areas, including modeling of DNA methylation. Each point i has a characteristic σ_i associated with it, which is equal to +1 or -1. The energy of the configuration $\sigma = \{\sigma_i\}_{i=1}^N$ is defined as

$$H(\sigma) = -\mu \sum h_n \sigma_n - \sum_{n=2}^{N} J_{n-1,n} \sigma_{n-1} \sigma_n \tag{1}$$

As one can see, this formula incorporates the terms that correspond to individual particles and the terms that correspond to the interaction between them. The probability of the system exhibiting the state σ is defined by the formula

$$P(\sigma) = \frac{\mathrm{e}^{-AH(\sigma)}}{\sum_\sigma \mathrm{e}^{-AH(\sigma)}}, \tag{2}$$

where A is the temperature parameter, and the denominator is the sum over all possible σ of length N.

In the paper, the authors define for each genomic unit a vector $x = \{x_i\}_{i=1}^N$, where $x_i = 0$ if ith CpG unit is unmethylated, and 1 otherwise. They take $A = 1, \mu = 1, h_n = \alpha + \beta\rho_n, J_{n-1,n} = \frac{\gamma}{d_n}$, ρ_n is CpG density, d_n is the distance between CpG sites $n-1$ and n.

The overall formula is

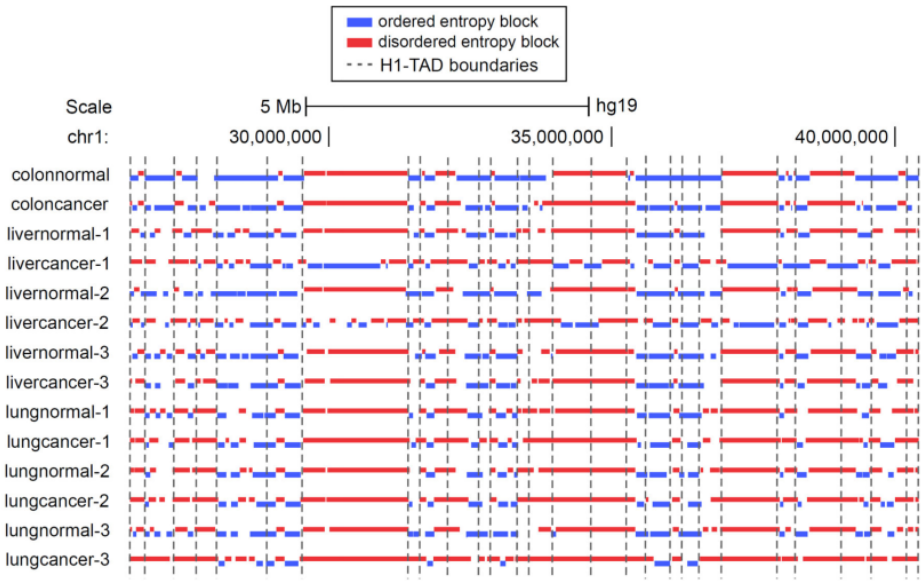


Figure 4: Example of correspondense between TAD bounaries and boundaries between ordered and disordered regions of DNA

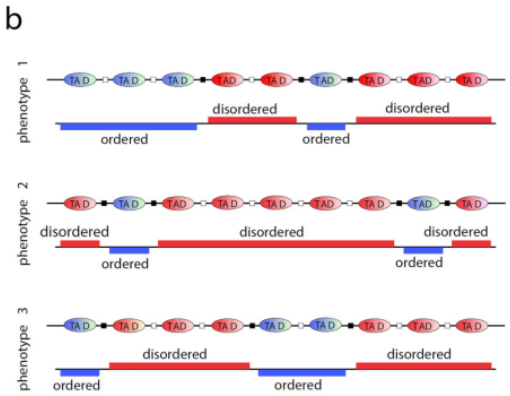


Figure 5: Another example of correspondense between TAD bounaries and boundaries between ordered and disordered regions of DNA

$$U(x) = -\sum_{n=1}^{N}(\alpha + \beta\rho_n)(2x_n - 1) - \sum_{n=2}^{N}\frac{\gamma}{d_n}(2x_n - 1)(2x_{n-1} - 1) \tag{3}$$

Let us define methylation level ℓ as the fraction of the CpG-sites that are methylated. The probability distribution of ℓ is thus given by the formula

$$P(\ell) = \sum_{x \in Q(N\ell)} P(x) \tag{4}$$

$$H = -\sum_{\ell} P(\ell)\log_2 P(\ell) \tag{5}$$

Note that it always holds that $H \in (0, \log_2(n+1)$. Therefore, the normalized methylation entropy of the genomic unit is defined as

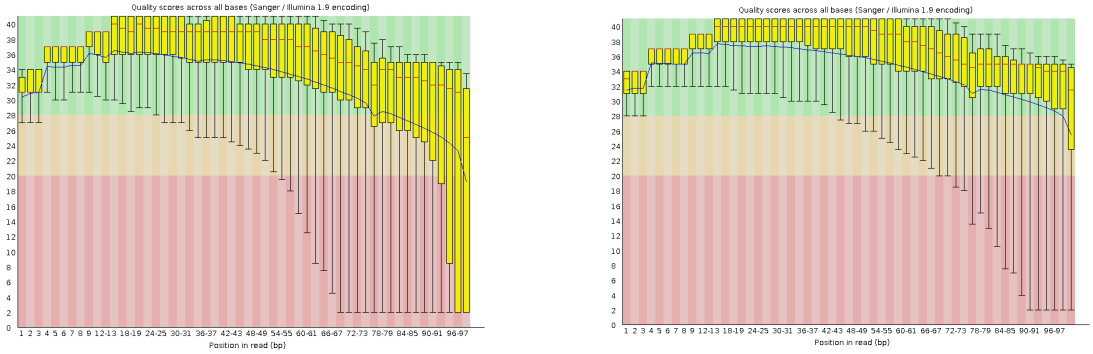
$$h = \frac{H}{\log_2(n+1)} \tag{6}$$

2.3 Entropy blocks computation

Entropy blocks are computed as follows:
1. The genome is split into genetic units of size 150 bp.
2. Normalized Shannon methylation entropy h is computed for each unit.
3. Units are classified like this:
$0 < h \leq 0.44$ ordered
$0.44 < h \leq 0.92$ weakly ordered
$0.92 < h \leq 1$ disordered
4. Use the sliding window of size 500 units (75 kb) and call the window ordered or disordered if at least 75% of the units are classified accordingly.
5. Unite the same-classified windows.

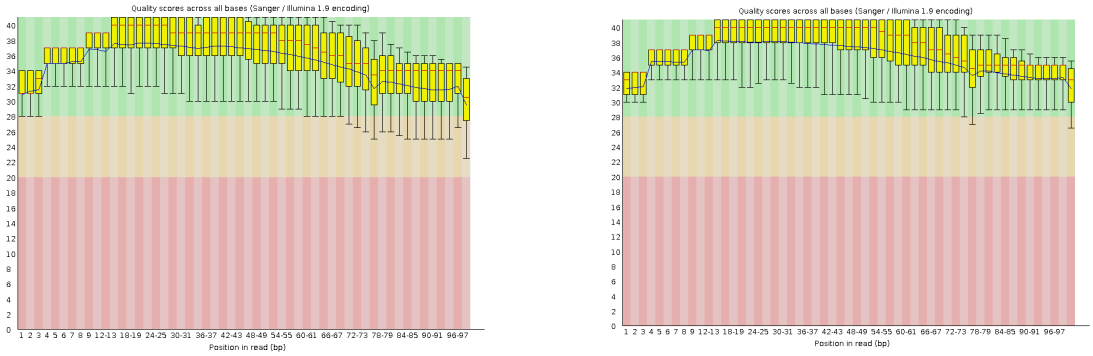
3 Methods and results

We have downloaded the paired-end WGBS data for colon normal cells from SRA database that were obtained in this study [10]. We extracted the fastq-files using fastq-dump, assessed the quality of the reads using FastQC [**?**], trimmed the reads using Trimmomatic [11], and aligned them with the genome using Bismark [12].



(a) First reads (b) Second reads

Figure 6: Quality of untrimmed reads



(a) First reads (b) Second reads

Figure 7: Quality of trimmed reads

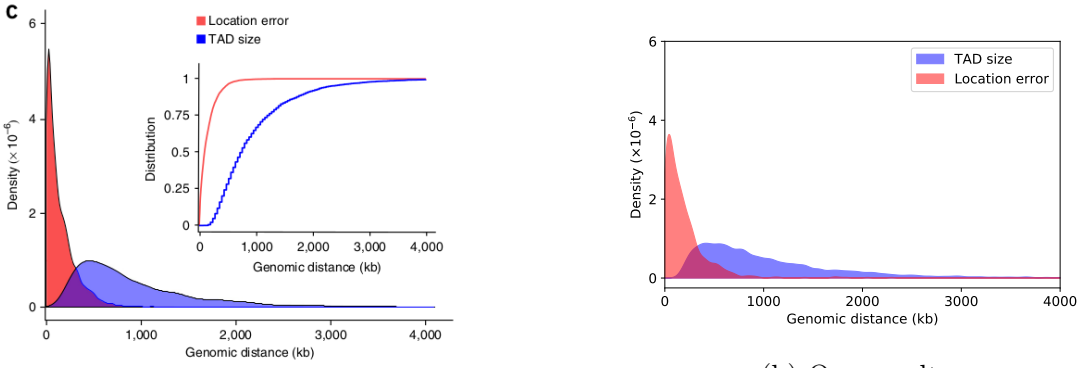
As a result, we obtained 5 bed-files:

- mean methylation levels
- normalized methylation entropy
- methylation-based classification (non-variable)
- methylation-based classification (variable)
- entropy-based classification

TAD boundaries, computed in the study [13], were obtained from the database. As suggested, we have taken all boundaries from all 6 files and obtained 8853 different regions. In the paper, the authors consider TAD boundary prediction to be correct if the distance between the boundary of entropy block and TAD boundary is less than the 1st quartile of TAD lengths.

We found 1st quartile of TAD lengths to be equal to 520kb, and 93.4% of entropy block boundaries are within this distance from the closest TAD boundary (which is similar to authors' 90%). These correct predictions cover 11.5% of all TAD boundaries (similar to authors' 6%).

In the paper, the experiment on combining block boundaries from 17 different tissue cells was performed, which increased the precision up to 95% and recall up to 62%. We did not do that in this work.



(a) Result from the paper

(b) Our result

Figure 8: Distribution of TAD sizes and errors

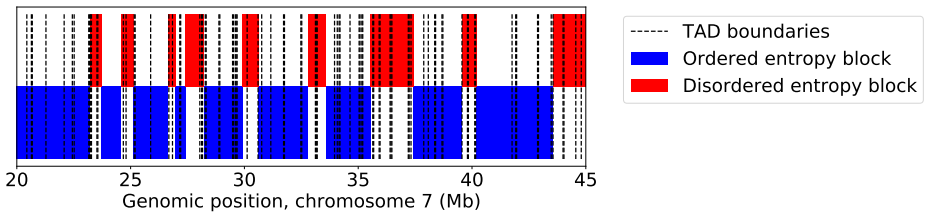


Figure 9: Example of visual similarity between entropy blocks and TAD boundaries

4 Discussion

The next steps in our research are as follows:

1. recreate TAD boundaries with higher resolution manually from Hi-C matrices by manipulating with insulation scores
2. use techniques from statistics and ML to create a model (probably, similar to this variant of convolutional neural network [14]) that would predict TAD boundaries using methylation entropy as a feature
3. fit the model on methylation entropy for D. rerio
4. use the model to predict TAD boundaries on another fish species, stickleback

5 Conclusion

We have verified that the available code works correctly and that the conclusion about correspondence between TAD boundaries and boundaries of methylation entropy blocks is correct. This was a preliminary step towards a goal of developing of ML algorithm that would predict TAD boundaries based on methylation data.

References

[1] B. Bonev and G. Cavalli, "Organization and function of the 3D genome," p. 19.

[2] M. Frommer, L. E. McDonald, D. S. Millar, C. M. Collis, F. Watt, G. W. Grigg, P. L. Molloy, and C. L. Paul, "A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands." *Proceedings of the National Academy of Sciences*, vol. 89, no. 5, pp. 1827–1831, Mar. 1992. [Online]. Available: http://www.pnas.org/cgi/doi/10.1073/pnas.89.5.1827

[3] J. Susan, J. Harrison, C. L. Paul, and M. Frommer, "High sensitivity mapping of methylated cytosines," *Nucleic Acids Research*, vol. 22, no. 15, pp. 2990–2997, 1994. [Online]. Available: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/22.15.2990

[4] S. J. Cokus, S. Feng, X. Zhang, Z. Chen, B. Merriman, C. D. Haudenschild, S. Pradhan, S. F. Nelson, M. Pellegrini, and S. E. Jacobsen, "Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning," *Nature*, vol. 452, no. 7184, pp. 215–219, Mar. 2008. [Online]. Available: http://www.nature.com/articles/nature06745

[5] L. Laurent, E. Wong, G. Li, T. Huynh, A. Tsirigos, C. T. Ong, H. M. Low, K. W. Kin Sung, I. Rigoutsos, J. Loring, and C.-L. Wei, "Dynamic changes in the human methylome during differentiation," *Genome Research*, vol. 20, no. 3, pp. 320–331, Mar. 2010. [Online]. Available: http://genome.cshlp.org/cgi/doi/10.1101/gr.101907.109

[6] G. Jenkinson, "Potential energy landscapes identify the information-theoretic nature of the epigenome," *Nature Genetics*, p. 14, 2017.

[7] ——, "An information-theoretic approach to the modeling and analysis of whole-genome bisulfite sequencing data," p. 23, 2018.

[8] J. Goutsias, "Ranking genomic features using an information-theoretic measure of epigenetic discordance," p. 17, 2019.

[9] S. Presse, K. Ghosh, J. Lee, and K. A. Dill, "Principles of maximum entropy and maximum caliber in statistical physics," *Rev. Mod. Phys.*, vol. 85, no. 3, p. 28, 2013.

[10] M. J. Ziller, "Charting a dynamic DNA methylation landscape of the human genome," p. 5.

[11] A. M. Bolger, M. Lohse, and B. Usadel, "Trimmomatic: a flexible trimmer for Illumina sequence data," p. 7.

[12] F. Krueger and S. R. Andrews, "Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications," p. 2.

[13] J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, J. S. Liu, and B. Ren, "Topological Domains in Mammalian Genomes Identified by Analysis of Chromatin Interactions," p. 13, 2012.

[14] G. Fudenberg, "Predicting 3D genome folding from DNA sequence with Akita," *Nature Methods*, p. 26.