Abstract

In this report we state a problem of Hi-C data prediction using ML methods, give necessary definitions, propose possible solution methods and next steps.

1 Introduction

Let us start with defining several relevant terms.

Chromatin is a complex of DNA and proteins. In prokaryotes, chromatin is located in 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.

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 fragment obviously 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.

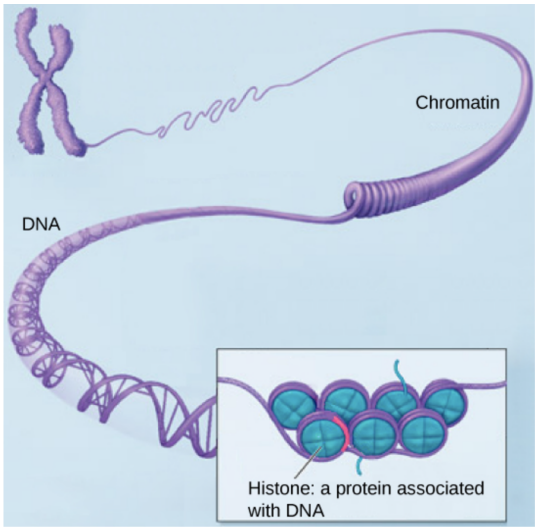

Figure 1: Schematic chromatin structure

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. CpG site is a region of DNA where a cytosine nucleotide is followed by a guanine nucleotide.

In mammals, the majority of CpG pairs are chemically modified by the covalent attachment of a methyl group to the C5 position of the cytosine ring. This situation is called methylation, and it is an important property of DNA sequence. This modified residue is distributed throughout the majority of the genome. The genome is punctuated however by non-methylated DNA sequences called CpG islands (CGIs) which have an elevated G + C content and little CpG suppression.

Hi-C method is a state-of-the-art method of determining the 3d chromatin structure. Hi-C matrix is a huge symmetric matrix, rows and columns of which correspond to

nucleotide bins, and values of which represent the probability of interaction of the corresponding bins. Hi-C data is obtained experimentally, however, these experiments tend to be very expensive and hard to perform. One can use the Machine Learning techniques based on 1d features of DNA sequence to predict the Hi-C matrix.

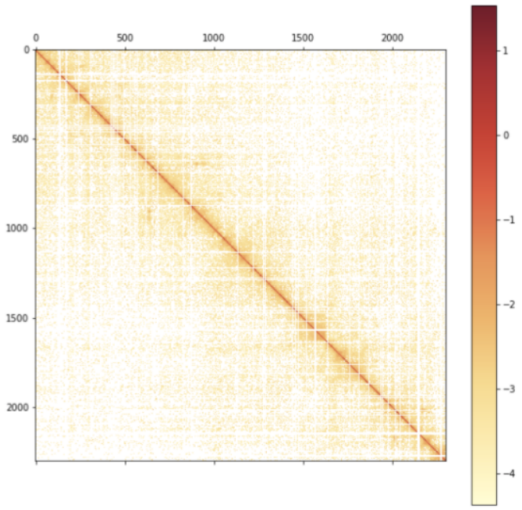

Figure 2: Example Hi-C matrix

Deep Learning is a rapidly developing field of Machine Learning which uses the so-called neural networks, parametric statistical models with huge variety of structures. In this research we are planning to concider the convolutional neural networks - a type of networks that are used in the problems of computer vision. The vanilla CNN consists of several convolutional layers followed by feed-forward linear layers. We will not go into much detail here, but, basically, a convolutional layer simply performs a cross-correlarion between two matrices plus bias term (see pictures for more clarity).

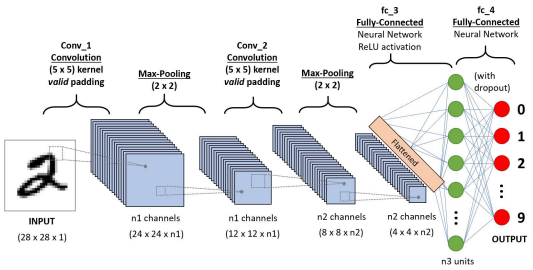

Figure 3: Example CNN architecture

Vanilla CNN takes an image as an input and produces a vector. In this research, we would need to make an opposite thing - take the sequence of data related to DNA and output an image.

2 Potential energy landscapes

Let us mention an important paper [1] on the subject. In this paper, the authors analyze the methylation data obtained from whole-genome bisulfite sequencing (WGBS) experimental protocol. They apply the principles of information theory (Shannon entropy)

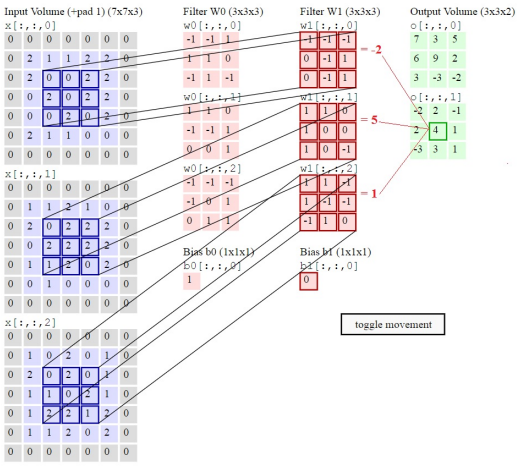

Figure 4: Possible convolutional layer visualization

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. Unfortunately, the file types and their exact usage remain unclear to us, and figuring this out is an important step in this research. Details of the Ising Model and Shannon entropy calculation are provided in a vague form, and another important aim is to understand these as well.

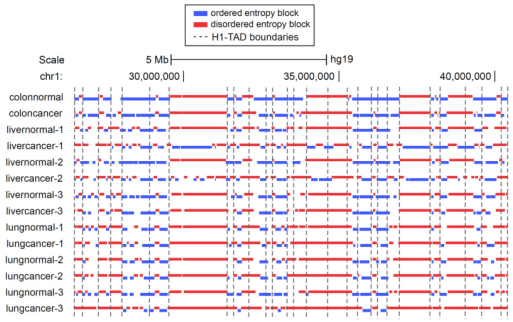

Figure 5: Example of correspondense between TAD bounaries and boundaries between ordered and disordered regions of DNA

Entropy blocks are computed as follows:

1. The genome is split into same-size genetic units.

2. Normalized Shannon methylation enthropy h is computed for each unit (we do not provide a formula here).

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.

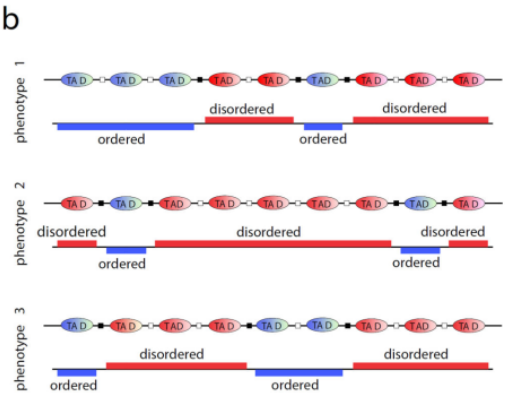

Figure 6: Another example of correspondense between TAD bounaries and boundaries between ordered and disordered regions of DNA

3 Next Steps

In this reseach we would like to create a model based on the data from Danio rerio (Zebrafish) species and use the model on the data coming from Stickleback fish. Here is the list of things that are necessary to do:

0. Create a comprehensive survey of related work, such as *DNA Methylation in Zebrafish, Base-resolution DNA methylation landscape of zebrafish brain and liver,* and *preciseTAD: A machine learning framework for precise 3D domain boundary prediction at base-level resolution*, and, additionally, provide better explanations of aforementioned unclear places.

1. Reproduce the procedure of entropy computation from the aforementioned paper [1] to make sure that the procedure works correctly and the results are the same as in the paper. Alternatively, install the software written by the authors and learn to use it.

2. Recreate the algorithm for Shannon enthropy computation and launch it on raw methylation data for D. rerio.

3. Predict the TAD boundaries of D. rerio using Shannon enthropy.

4. Obtain the gene methylation data for the stickleback and predict the TAD boundaries for it.

5. Use RNA-seq and ATAC-seq data for validation of the results (these data correlate with TAD boundaries reasonably well).

6. Extend the work [3] and incorporate the methylation data and possibly other types of 1d DNA sequence properties in order to predict Hi-C matrices better. This step includes preparing the data, creating the training dataset and Deep Learning model, tuning the hyperparameters.

4 Conclusion

In this report we have outlined important points about our research on predicting Hi-C data. We have given an introduction to the topic, possible approach to the formulated problem and the plan of next actions.

References

[1] Jenkinson, G., Pujadas, E., Goutsias, J., and Feinberg, A.P. (2017), *Potential energy landscapes identify the information-theoretic nature of the epigenome*

[2] Illingworth, Robert S. and Bird, Adrian P. *CpG islands – 'A rough guide'*

[3] Alishev N. A. *Генерация Hi-C карт из последовательностей ДНК с помощью глубокого машинного обучения* (written in Russian)

[4] DeepTACT: predicting 3D chromatin contacts via bootstrapping deep learning Wenran Li, Wing Hung Wong, Rui Jiang Nucleic Acids Research, Volume 47, Issue 10, 04 June 2019, Page e60, https://doi.org/10.1093/nar/gkz167