Enhancing HiC data resolution with convolutional neural networks

Behnam Rasoolian
Department of Software Engineering
Auburn University
Auburn, AL 36832
behnam@auburn.edu

Liangliang Xu
Department of Industrial Engineering
Auburn University
Auburn, AL 36832
lzx0014@auburn.edu

Zheng Zhang
Department of Software Engineering
Auburn University
Auburn, AL 36832
zzz0069@auburn.edu

1 Abstract

2 Introduction

The cell of a eukaryotic species forms a multi-granularity genome structure in order to compactly store a very long genomic DNA sequence in its small nucleus. A **nucelotide** is the building block of DNA. There are 4 types of nucleotides: C, G, A and T. Each pair of nucleotides in the DNA are called a **base**. A kilo-base is a group of 1000 bases. thousands of bases join together to form **gene loci**. A number of loci then fold into a large independent physical structure called **chromosome** (Wang et al. (2013)).

Study of spatial conformation of chromosomes is of high importance in the field of (computational) biology. Although all cell is a living being have the same sequence of genes, it is the 3D positioning of these genes in space that determines how the cell functions. Roughly said, if two genes are close to each other in space, they can interact with each other in order to create a certain protein that regulates a certain task. Thus, being able study this 3D configuration can help unravel mysteries of cell functioning. However, this spatial organization of chromosomes can not be observed through traditional microscopy. As an alternative, high-throughput chromosome conformation capture (Hi-C) has emerged as a powerful method for studying the 3D organization of chromosomes in space. The HiC method, which was developed by Lieberman-Aiden et al. (2009), captures interactions between chromosomal fragments. In this method, a chromosome is divided into very small equally sized sections called *loci* which is composed of 1K to 1M bases. this method then measures all pair-wise interaction frequencies across all chromosomes. In the past years, Hi-C method has lead to some exiting discoveries about the topology of chromosomes such as presence of *chromatin loops*. Hi-C data are usually provided as a $N \times N$ heatmap or *contact matrix* where N is the number of loci in the genome. Each cell in the heatmap indicates the number of *interactions* found between a pair of loci corresponding to the rows and columns. 'Resolution' of a Hi-C data is the size of the loci the genome is divided into. As mentioned above resolution can range from 1 kb to 1 Mb. *sequencing depth* is the most important factor that determines the resolution of data. A higher sequencing depth results in capturing interactions between smaller loci, thus improving the resolution of the data. the sequencing process is costly and linear increase of resolution requires quadratic increase of sequencing reads. thus, most of the Hi-C data available have low resolutions.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

Data	Short	Resolution	Used . . .
Normal B-cell(GM12878)	RAO	high	As response in training
Normal B-cell(GM06990)	MIT	low	As feature in training
B-acute lymphoblastic leukemia	ALL	low	To predict high resolution
Follicular lymphoma cell-line	RL	low	To predict high resolution
MHH-CALL-4	CALL4	low	To predict high resolution

Table 1: Data description

Therefore, it is required that a computational method be developed to improve the resolution of currently availabe Hi-C data and generate Hi-C contact matrices of higher contrast. Recently, deep learning especially Convolutional Neural Network has emerged as a successful method in several applications such as computational epigenomics. It has been successfully used to predict DNA methylation or gene expression patterns.

3 The Model

In this project, we are building upon HiCPlus, a model proposed in Zhang et al. (2018), which uses CNNs to predict a high resolution contact matrix from a down-sampled matrix.

In this project, however, we used HiCPlus model to enhance the contrast of our low resolution data by training on a high resolution data.

Our final purpose is to be able to compare the three leukemic cells with a normal cell in terms of spatial structure and find whether there is any difference in their 3D conformation or not. In our research, we have 5 Hi-C data sets, belonging to 4 cell lines. Two data sets are sequenced from a normal cell line, one with high resolution (GM12878) and the other with low resolution (GM06990) The other three data sets are sequenced from the same cell lines with with three different type of leukemia.

As mentioned above, four of the data sets we have are sequenced with low depth, resulting in relatively low resolution. We also have access to a high-resolution data with much higher resolution which is sequenced from exactly the same cell as the normal low-resolution data that we have. Therefore, we used the HiCPlus model in Zhang et al. (2018) to enhace the contrast of our data. We trained the model by using the low- and high-resolution data of normal cells and then applied it to the other three cells in order to improve their contrast. A summary of the data and how they are used can be found in table 1.

3.1 Overview of HiCPlus framework

The inputs to the model are a low-resolution and a high-resolution date from the same cell line. In our project we used GM06990 for low-rosolution and GM12878 for high-resolution data. The two data are sequenced from the same cell lines with the difference that the former data cavers 979.4M bases while the latter covers 85.1G bases, that is, the resolution of GM12878 data is roughly 87 times higher than the GM06990 data. As to malignant cells, we re-used Leukemic Hi-C libraries created in Wang et al. (2013) These libraries were sequenced for cases of primary human B-acute lymphoblastic leukemia (B-ALL or ALL), the MHH-CALL-4 B-ALL cell line (CALL4), and the follicular lymphoma cell-line (RL). Just as Wang et al. (2013), we used normal B-cell line (GM068990) from Lieberman-Aiden et al. (2009) for our comparisons. We then fit the ConvNet model using values at each position in the high-resolution matrix as the response variable and using its neighbouring points from the low-resolution matrix as the predictors. The authors of Zhang et al. (2018) propose a neighborhood of size 40×40 as the neighborhoold that yields best results. Thus in order the prepare the data, we

Number	Name	Filter size	Filter Numbers	Strides	Output Shape
0	input	-	-	-	$1 \times 40 \times 40$
1	conv2d1	9	8	1	$8 \times 32 \times 32$
2	conv2d2	1	8	1	$8 \times 32 \times 32$
3	conv2d3	5	1	1	$1 \times 28 \times 28$
4	output_layer	-	-	-	1×784

Table 2: Desciption the CNN layers used in our project. The model is composed of three convolutional layers. The input is of shape $1 \times 40 \times 40$ and the output has a shape of 1×784. There are not deeply connected layers in the model.

first divided both low- and high-resolution contact matrices into patches of size 40×40. The model consists of 3 convolutional layers. The design of the model is described in table 2 and illustrated in figure 1.

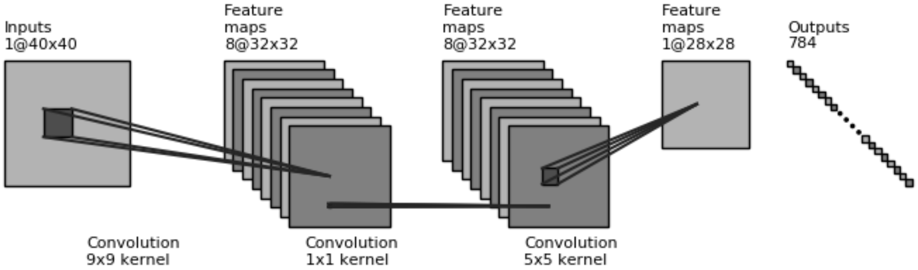


Figure 1: Illustration of the CNN layers used in our project. The model is composed of three convolutional layers. The input is of shape $1 \times 40 \times 40$ and the output has a shape of 1×784. There are not deeply connected layers in the model.

3.2 Loss Function

We used mean square of differences as the loss function. As can be seen in table 2 and 1, the output of the model hase a shape of 1×784. In order to calculate loss function, the model picks the middle 28 rows and columns of the corresponding high-resolution patch and flattens it. It then calculates the mean square of differeneces between the output of the model and the high-resolution sub-patch. The loss function is formulated as follows:

$$\mathbb{L} = \frac{1}{784} \sum_{i=1}^{784} (\hat{y}_i - y_i)^2 \tag{1}$$

where $\hat{y}$ denotes the output of the model and y denotes the actual high-resolution sub-patch.

We used simple gradient descent optimizer with learning rate of `1e-5` to train the model. The choise of optimizer did not make a great difference in the results. Our results from using Adam optimizier also were the same as SGD.

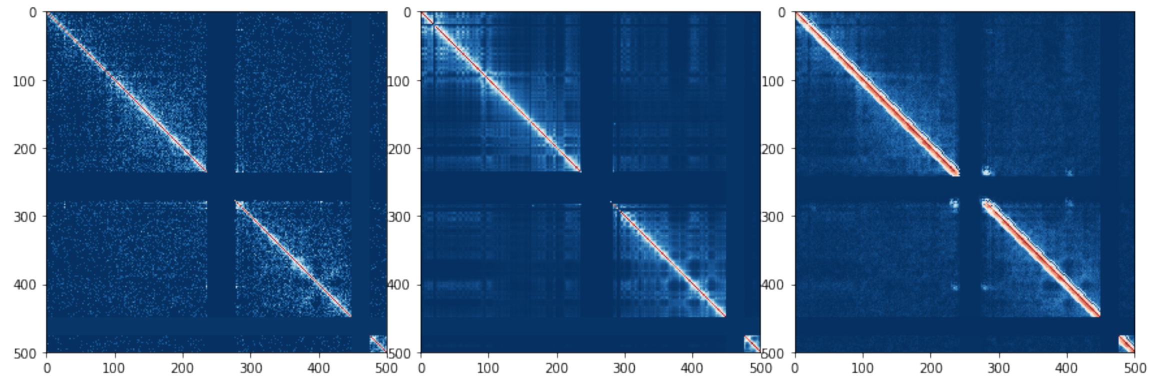


Figure 2: Results of training the data on Normal cells. There is a total of almost 6000 loci in a contact matrix. For the sake of demonstration, we only show the firt 500 loci. The left contact matrix is the original low-resolution data. The middle one is the high resolution data. and the rightmost picture is the result of training the model.

4 Results

Platforms: The original model was developed in `theano` framework. The package `lasagne` was used for model development. Although we made some experiments with tensorflow, conducted our final traning via the orginal model. Our results can be found in a git repository (refer to resources section).

We trained the model using low-resolution GM06990 and high-resolution GM12878 data, both of which are sequence from the same cell. We then used the other three data that we have (corresponding to three cancerous cells) as input to the model to improve contrast. The data that we uses is described in table 1

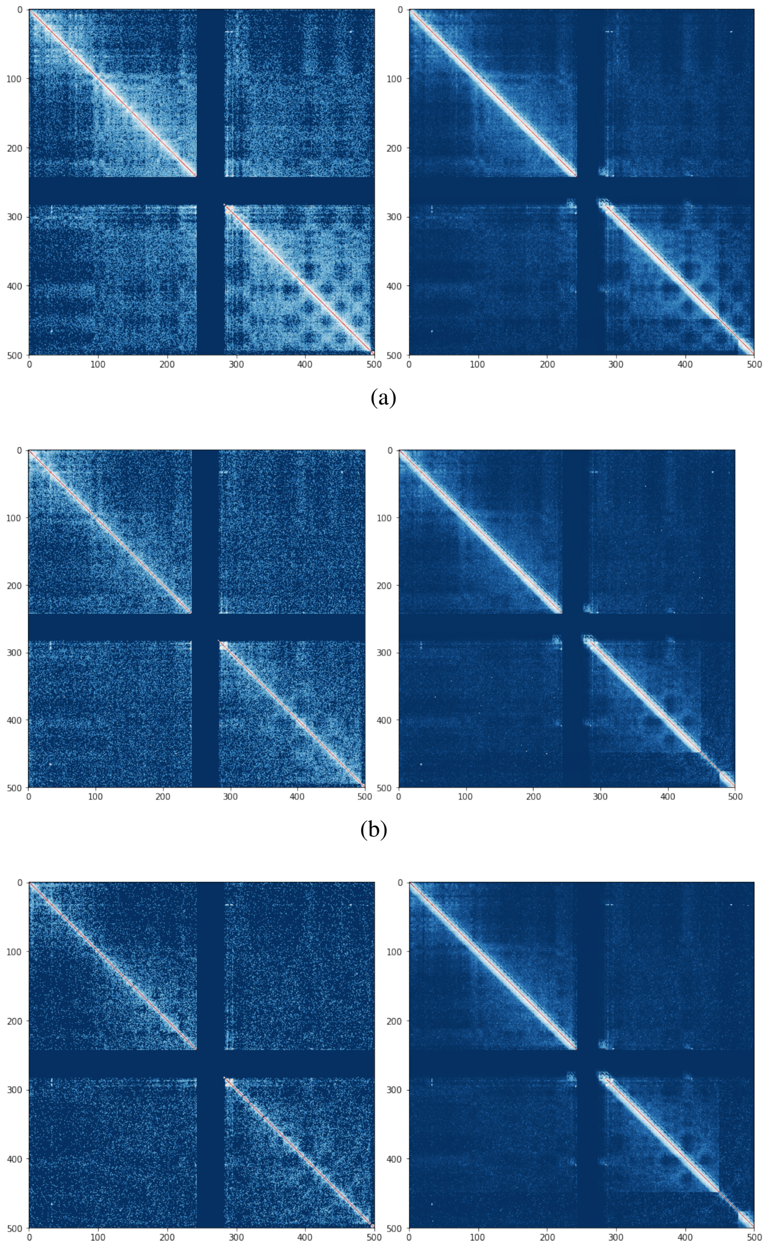
The model was ran for 30 epochs. The results of training can be found in figure 2. As can be seen the results are smoother and demonstrate higher contrast than the original model. Figure 3 show prediction results from the model for the three cancer cells. The results also demonstrat higher contrast compared to original contact matrices.

5 Strengths and Weaknesses and Future Work

This method can be considered as a noise reduction method for Hi-C contact matrices. Just as all currently available normalization and noise reduction approaches this method also relies on assumptions. For example Balance Network Deconvolution, proposed by Feizi et al. (2013) assumes that the observed graph G_{obs} is a summation of its direct graph G_{dir} and some indirect terms $G_{obs} = G_{dir} + G_{dir}^2 + G_{dir} + ...$. The major assumption of this approach is that CNNs can predict values in high-resolution matrix from the surrounding neighborhood in low-resolution matrix. The difference of this assumption from other approaches is that it is easier to put the assumption to statistical tests. Authors of Zhang et al. (2018) came to the conclusion that the assumption holds true for a down-sampled version of high-resolution data. Whether this is the case for an independent low-resolution data remains to be investigated in future works.

6 Resources

Source code for the project:
`https://github.com/rasoolianbehnam/watson/tree/master/HiCPlus`
Link to the low-resolution Hi-C datasets:
`http://sysbio.rnet.missouri.edu/T0510/tmp_download/link_to_download_genome_data/contact_file/`



(a)

(b)

(c)

Figure 3: Results of running the model on cancer cells. For the sake of demonstration, we only show the firt 500 loci. The left contact matrices are the original low-resolution data and the rightmost pictures are the result of model predictions.

Link to high-resolution Hi-C dataset:
`https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE63525`

References

Z. Wang, R. Cao, K. Taylor, A. Briley, C. Caldwell, and J. Cheng. The properties of genome conformation and spatial gene interaction and regulation networks of normal and malignant human cell types. *PloS one*, 8(3):e58793, 2013. 1, 2

E. Lieberman-Aiden, N. L. Van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science*, 326(5950):289–293, 2009. 1, 2

Y. Zhang, L. An, J. Xu, B. Zhang, W. J. Zheng, M. Hu, J. Tang, and F. Yue. Enhancing hi-c data resolution with deep convolutional neural network hicplus. *Nature communications*, 9(1):750, 2018. 2, 4

S. Feizi, D. Marbach, M. Médard, and M. Kellis. Network deconvolution as a general method to distinguish direct dependencies in networks. *Nature biotechnology*, 31(8):726–733, 2013. 4