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

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 genese 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 powerfull method for studying the 3D organization of chromosomes in space. In this method, a chromosome is divided into very small equally sized sections called *loci* which is composed of 1K to 1000K genes. 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 samller loci, thus improving the resolution of the data. the sequencing process is costly and linewar increase of resolution requires quadratic increase of sequencing reads. thus, most of the Hi-C data availabe have low resolutions.

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.

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

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 1: My caption

3 The Model

In Zhang et al. (2018), a model was proposes as HiCPlus, that uses CNNs to predict a high resolution contact matrix from a down-sampled matrix. In this project, we have used HiCPlus model to enhance the contrast for our own data.

In our research, we have Hi-C data of 4 cell lines. One of which is sequenced from a normal cell line and the other three sequenced from cells afflicted with three different malignancies. Our purpose is to compare them in terms of spatial structure and find whether there is any difference in their 3D conformation or not. All 4 data that we have are sequenced with low depth, resulting in relatively low resolution. Therefore, we used the HiCPlus model in Zhang et al. (2018) to enhace the contrast of our data.

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. 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 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 1 and illustrated in figure 1.

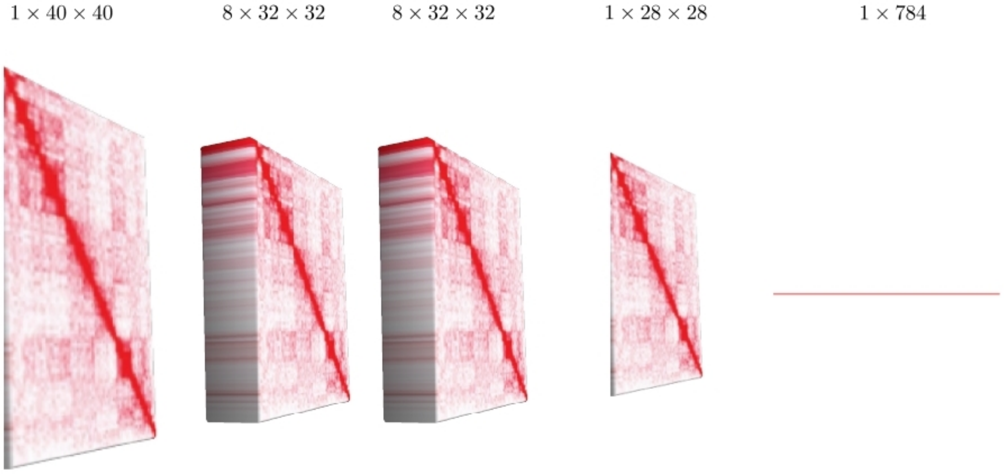

Figure 1

3.1.1 Loss Function

We used mean squre of differences as the loss function. As can be seen in table 1 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 colums of the corresponing high-resolution patch and flattens it. It then calculates the mean square of differneces between the output of the model and the high-resolution sub-patch. The loss function is formulated as follows:

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

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

4 Strengths and Weaknesses

5 Future Work

6 Questions & Answers

References

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