

Enhancing Hi-C contact map resolution with neural network

November 4, 2020

1 Introduction

Recently, the high-throughput chromosome conformation capture(Hi-C) technique has become a powerful tool for studying the three-dimensional structure of chromosomes. Hi-C data is usually expressed as a $n \times n$ matrix. The resolution of Hi-C data is defined as the bin size of each cell of the matrix. Hi-C data at kilobase level are requisite for future genome 3D structure studies. Rao et al.(2014) generated Hi-C data with 1 kilobase resolution. However, millions of sequenced reads are required to archive this resolution with a huge amount of money and time consumption.

Zhang et al. presented a approach to enhance the resolution of Hi-C data called HiCPlus. Which generated low-resolution data by down-sampling the number of sequenced reads and then a neural network was used to create the mapping between high-resolution contact map and low-resolution contact map.

[1]

2 Methods

Let D be a set of paired ends reads of a Hi-C experiment. We make a low and high contact maps from D , denoted by M_ℓ and M_h . Let S_ℓ and S_h be the size of M_ℓ and M_h . Let R be the ratio of M_ℓ to M_h , which can be represented by $R = \frac{M_h}{M_\ell}$. And O be the number of overlapping pixels between adjacent sub-maps.

% Divide matrices M_ℓ and M_h

To M_ℓ :

```
for  $i = 1, 1+40 - O, 1+2 \times (40 - O), \dots$ 
  for  $j = 1, 1+40 - O, 1+2 \times (40 - O), \dots$ 
    IF  $i + 40 > M_\ell$  ||  $j + 40 > M_\ell$ , BREAK
    ELSE extract  $40 \times 40$  sub-maps whose lefttop coordinate
      is  $(i, j)$  from  $M_\ell$ .
```

Do the same process to M_h :

```
for  $i = 1, 1+(40-O) \times R, 1+2 \times (40-O) \times R, \dots$ 
  for  $j = 1, 1+(40-O) \times R, 1+2 \times (40-O) \times R, \dots$ 
    IF  $i + 40 \times R > M_h$  ||  $j + 40 \times R > M_h$ , BREAK
    ELSE extract  $(40 \times R) \times (40 \times R)$  sub-maps whose lefttop
      coordinate is  $(i, j)$  from  $M_h$ .
```

% training part:

Let C_ℓ and C_h be a collection of the resulting sub-maps.
 Train a neural network using C_ℓ and C_h . Mean square error is used as
 loss function in the training process.

We can use (f,n) to represent the parameters of each layer. Parameter f means the size of the
 ##Layer1(Pattern extraction)
 ##Layer2(Low-res mapping to high-res)
 ##Layer3(Predicted contact maps generation)

% test part:

Use other chromosome.

Do the same dividing process like M_ℓ and M_h

Calculate the Pearson's correlation between the output and M_h .

Step 1 Data preparation and processing

Since this experiment is to validate the algorithm for mapping low-resolution data to high-resolution data, high-resolution data are required.

In order to compare to some state-of-the-art approaches (HiCPlus and HiCNN), we use data sets (such as GM12878 from GSE63525) which are also used in other approaches. We start from generating a 10kb resolution contact map using Hi-C Pro. Then we perform down-sampling on high-resolution data. We use BAM files to generate low-resolution contact maps by changing the bin size bigger. We generate three contact maps with bin sizes are 20kb, 30kb and 40kb, respectively. We use chromosome 1-8 as training sets, and chromosome 17 as test set.

Step 2 Learning by Neural network

We separate the low-resolution contact map into many 40×40 submatrices. Those submatrices are used as inputs.

2.1 Layer Structure

We consider the

References

- [1] MIZUSHIMA, N., AND KOMATSU, M. Autophagy: renovation of cells and tissues. *Cell* *147* (2011), 728–741.