Enhancing Hi-C contact map resolution with neural network

Wang Minggu And Maruyama Osamu

November 18, 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 an $n \times n$ 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. HiPlus 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 generate 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 S_ℓ to S_h, which can be represented by $R = \frac{S_\ell}{S_h}$. And P be the number of overlapping pixels between adjacent sub-maps. Let K be the size of low-resolution sub-maps.

```
% Divide matrices $M_\ell$ and $M_h$
To $M_\ell$ :

for $i$ = 1, 1+$K-P$, 1+2$\times(K-P)$, ...
    for $j$ = 1, 1+$K-P$, 1+2$\times(K-P)$, ...
        IF $i+K$ > $M_\ell$ || $j+K$ > $M_\ell$, BREAK
        ELSE extract K $\times$ K sub-maps whose left-top coordinate
         is $(i,j)$ from $M_\ell$.
```

```
Do the same process to $M_h$ :

for $i$ = 1, 1+$(K-P)$ $\times$ $R$, 1+2$\times$$(K-P)$ $\times$ $R$, ...
    for $j$ = 1, 1+$(K-P)$ $\times$ $R$, 1+2$\times$$(K-P)$ $\times$ $R$, ...
        IF $i+K$ $\times$ $R$ > $M_h$ || $j+K$ $\times$ $R$ > $M_h$, BREAK
        ELSE extract $(K$ $\times$ $R)$ $\times$ $(K$ $\times$ $R)$ sub-maps whose left-top
        coordinate is $(i,j)$ from $M_h$.
```

Let C_ℓ and C_h be collections of the resulting low-resolution and high-resolution sub-maps. Let $\hat{C_h}$ be the collection of the high-resolution sub-matrices generated by neural network. We train a neural network using C_ℓ and C_h. The mean square error(MSE) is used as loss function in the training process.

$$MSE[C_\ell, C_h] = \frac{1}{K\times K}\sum_{i=1}^{K\times R}\sum_{j=1}^{K\times R}(\hat{C}_{h_{i,j}} - C_{h_{i,j}})^2$$

Where $C_{\ell_{i,j}}$ and $C_{h_{i,j}}$ represent left-top coordinate (i,j) in C_ℓ and C_h respectively. We can use (f,n) to represent the parameters of each layer. Parameter f means the size of the filter and n means the number of filter.
1st Layer (Pattern extraction) Base on every $K \times K$ sub-matrix, using $f \times f$(13×13 in HiCPlus) filters to extract patterns of each sub-contact-map. Which can represented by following formula:

$$F_1(X) = ReLU(w_1 * X + b_1)$$

where $*$ represent the convolution process. w_1 represents $n \times f \times f$ filters. b_1 is the bias.
2nd Layer (Low-res mapping to high-res)
3rd Layer (Predicted contact maps generation)
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 $K \times K$ sub-matrices. Those sub-matrices 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.