¡¡¡¡¡¡¡ HEAD
======= ¿¿¿¿¿¿¿ a50e8a4f5c156bda61eb1a4c986f0c5992f69fee

Enhancing Hi-C contact map resolution with neural network

November 2, 2020

1 Introduction

Recently, the high-throughput chromoseome comformation 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.

[**Mizushima2011Autophagy**]

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 be the size of M_ℓ and M_h.

```
% training part:
for i = 1, 2, ..., $S-39$
    for j = 1, 2, ..., $S-39$
        extract sub-maps whose lefttop coordinate is (i,j) from $M_\ell$ and $M_h$.

Let $C$ be a collection of the resulting sub-maps.
Train a neural network using $C$.

% test part:
Use other chomosome.
```

```
for i = 1, 2, ..., $S-39$
    for j = 1, 2, ..., $S-39$
        extract sub-maps whose lefttop coordinate is (i,j) from $M_\ell$ 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

ref