Research on Hi-C data enhancement technology based on generative adversarial networks

Qian Bai\textsuperscript{a}, Zhe Cheng\textsuperscript{b}, Shipu Wang\textsuperscript{c}, Wei Zhou\textsuperscript{d}

School of software Yunnan University Kunming 650000, China

\textsuperscript{a}1299039053@qq.com, \textsuperscript{b}1506720972@qq.com, \textsuperscript{c}wangshipu@ynu.edu.cn, \textsuperscript{d}zwei@ynu.edu.cn

Abstract. Hi-C technology is one of the most popular tools for studying three-dimensional (3D) genome organization. Due to the high cost of sequencing, most Hi-C data have low resolution and cannot be used to connect distal regulatory elements to their target genes. To solve the problem that hi-c data of high resolution are not easy to obtain, this paper proposes a Hi-C enhancement method (HiCGAN) based on generative adversarial networks, taking the down-sampling interaction matrix which is highly similar to the original matrix as input, only 1/16 of the original sequencing reading can be used to generate the Hi-C interaction matrix of high resolution. In the experiment, Pearson correlation coefficient was used to measure the similarity between the generated high-resolution matrix and the real high-resolution hi-c matrix in numerical distribution. The apparent interaction pairs were analyzed by Fit-Hi-C, and calling ChromHMM annotates state of 12 kinds of chromatin. Experimental results show that HiCGAN models learned in one cell type can predict high-resolution Hi-C matrices for other cell types. This study proposes a computational framework (HiCGAN) for accurately predicting Hi-C data improving the resolution of Hi-C data.

1. Introduction (Heading 1)

The chromatin in the nucleus has a complex three-dimensional structure, which is essential to regulate genome function, DNA replication, transcription and other nuclear processes. Previous studies have shown that this form of genomic tissue is correlated with nuclear structure and highly dynamic among different cell states [18] [19] [20] (Quinodoz et al., 2018; Schmitt et al., 2016; uhler and shivashankar, 2017).

Hi-C [1] technology has been developed to display the three-dimensional (3D) conformation of genomic tissue. Hi-C technology is a technique to measure the frequency of all pairwise interactions in the whole genome. Compared with traditional chromosome conformation capture techniques (such as 3C [2] (Dekker, et al., 2002), 4C [3] (Zhao, et al., 2006) and 5C [4] (dostie, et al. 2006), etc, the main advantage of the Hi-C method is that it can capture potential contact points in the entire genome and lay the foundation for reconstructing the three-dimensional structure of the entire genome. Hi-C data has been applied to predict DNA methylation [10] (Wang, et al., 2016) and explore the relationship between Xist lncRNA and 3D genomic structure [5] (Engreitz, et al., 2013) and other fields. In addition, through a systematic analysis of Hi-C data, researchers discovered important conformational features of the genome, including: open and closed compartments [1] (Lieberman-Aiden, et al., 2009) and six
subcompartment of their successors [7] (Rao, et al., 2014). Topological association domains TADs [6] (Dixon, et al., 2012.) and Hi-C peaks indicating chromatin cycle [7] (Rao, et al., 2014). Although in the past few years, the experimental and computational methods for three-dimensional genome analysis have been developed rapidly, the current understanding of how genome organization affects cell function and disease is still limited. One of the main problems is that our understanding of genomic structure is still relatively weak on the kilobase pair, which hinders further analysis of more refined chromatin structure (such as subdomain or enhancer promoter interaction). The resolution of Hi-C data is usually defined as the smallest scale that can reliably identify local features. It also refers to the size of the container used to segment the genome when constructing the Hi-C contact matrix. The resolution of most available Hi-C data is relatively low, ranging from 25KB to 1MB.

On the one hand, high-resolution Hi-C data that requires large amount of sequencing reading and high cost of sequencing can only be obtained in several tissues or cell lines. On the other hand, the resolution of Hi-C data greatly affects the downstream analysis, such as identifying topological correlation domain and chromatin ring. Generally, the deeper the sequencing depth, the higher the resolution. High resolution Hi-C data can not only help us to recognize TADs more clearly, but also can find sub TADs domains on a more precise genome scale, which are considered as more cell types and inter species variables [21][22](Phillips cremins et al., 2013; Yu and Ren, 2017). Based on the above considerations, it is urgent to develop a method to improve data resolution by learning the mapping relationship between low resolution and high resolution data.

Currently, most of the publicly available high-resolution Hi-C data comes from time-consuming Hi-C experiments [7] (Rao, et al., 2014), requiring millions of mammalian cells, and the cost of sequencing depth is high(Sequencing depth is the most crucial factor in determining the resolution of Hi-C data—the higher the depth, the higher the resolution). Therefore, how to convert a large amount of low-resolution HiC data into high-resolution data through an economical and effective computable method is a very significant and urgent problem. Previous studies have shown that low-resolution Hi-C contact matrices include repeatable patterns [6] (Dixon, et al., 2012; Rao, et al., 2014). Learning repeated pattern can further reveal patterns that are not obvious in low-resolution Hi-C data. Zhang et al. (Zhang, et al., 2018) developed a calculation method called HiCPlus [9] to improve the resolution of Hi-C data, which uses a three-layer convolutional neural network (ConvNet) to Learn the mapping relationships between low-resolution and high-resolution Hi-C contact matrices. Experimental results show that HiCPlus is superior to traditional machine learning-based interpolation methods (Zhang, et al., 2018), such as random forest and Gaussian smoothing. Correlation coefficients between HiCPlus-enhanced Hi-C data and experimental high-resolution Hi-C Even greater than the correlation coefficient between two experimental replicates. However, due to the network depth of its model, HiCPlus failed to fully consider the relationship between the global mode and local mode of the data in the Hi-C matrix. Therefore, this paper proposes a new Hi-C resolution enhancement method (HiCGAN) based on generative adversarial networks. HiCGAN consists of a generator model and a discriminator model. The generator uses dense residual blocks to densely connect and deepen the network, extract rich local feature information, and superimpose it with the global pattern transmitted by Skip Connection to capture the data distribution of the samples. Thus, high-resolution Hi-C data is generated; and the discriminator model confronts the generator during the training process. With the help of game-type optimization, both models have better performance.

2. Method

2.1. Hi-C data preprocessing and contact matrix generation

Step 1. Pre-processing the Hi-C matrix: Many currently available Hi-C dataset, especially Hi-C data in human tissue, are only available in a 40kb resolution matrix. This paper uses these data sets, starting from the BAM file, to generate an interaction matrix with a resolution of 10 Kb as high-resolution Hi-C data, then uses a random downsampling method (using only 1/16 of the original sequencing depth) to simulate the generation of low-resolution Hi-C matrices as low-resolution Hi-C data. Compared with
deeply sequenced Hi-C library, the ratio of noise to signal of the Hi-C data used in this paper increases due to depth actual reached is lower.

Step 2: Divide a Hi-C matrix into multiple equal-sized sub-regions (each sub-region has an index number corresponding to the index number of the high-resolution sample, and the step size of the segmentation is 25). Each sub-region (size 0.4 × 0.4 Mb²) is treated as each sample and contains 40 × 40 = 1600 pixels. This paper only studies and predicts chromatin interactions with a genomic distance of <2 mb between the two loci, because the average size of the TADs domain is <1 mb, and there are few meaningful interactions outside the TADs domain.

Step 3: Train the generator. During the training phase, you can learn the relationship between low-resolution samples (samples of the same size but insufficient sequence depth for sequencing) and high-resolution samples, then use the low-resolution samples to predict the high-resolution samples.

Step 4: According to the index generated by segmenting the Hi-C matrix in step 2, merge the predicted high-resolution sub-matrix into the chromosome-size Hi-C interaction matrix.

Fig. 1 Hi-C matrix pre-processing flow chart

2.2. Overview of the HICGAN framework
The overall framework of HiCGAN is shown in Figure 2. HiCGAN is inspired by game theory. HiCGAN does not use a single neural network, but consists of two competitive neural networks called generator network (Figure 2) and discriminator network (Figure 3). The generator takes low resolution Hi-C samples as input and tries to generate pseudo high resolution Hi-C (or super-resolution) samples. The discriminator is used as a classifier to distinguish real high-resolution Hi-C data from super-resolution Hi-C data. After adversary training, generators tend to generate real data under almost the same high-resolution Hi-C data distribution. For low-resolution and high-resolution Hi-C samples, the low-resolution data is usually obtained through downsampling 1 / 16 sequencing reads of the original high-resolution Hi-C experiment, and HiCGAN can conduct the adversary training process. Then, after training convergence, we only need generator network to improve the resolution of Hi-C data.

Fig. 2 HICGAN Model frame diagram
2.3. Generator

In order to improve the SR performance of the model, the generator model shown in the figure below is designed. The network structure of the generator is based on the GAN network architecture. It improves the first ESRGAN [16] algorithm in the PIRM2018-SR challenge, abandons common deconvolution operations, and broadens the network width and reduces the network depth (65-level convolution operation) to improve the performance of the algorithm. The Hi-C matrix contains repeated local patterns, and the interaction intensity of each point is not independent of its local adjacent region. Therefore, an important parameter in the model is the size of the adjacent region. The convolution of 13×13 is used to check the input matrix (40×40×1) for feature extraction and representation; then apply a 1×1 convolution and increase the number of channels to 128 to reduce the computational complexity and prepare for subsequent dense residual blocks. Deepen the network structure with 10 Dense Blocks to obtain a larger receptive field and complete the non-linear mapping of the input data. Each block is concatenated by 5 convolutional layers with a number of 128 channels and a convolution kernel size of 3×3, and is activated using LRelu. The dense residual block (Dense Block) can not only improve the structure of the model, make it larger and easier to train, but also densely connect the rich local feature information of the Hi-C data extracted by the convolution layer, allowing the previous state is directly connected to all layers of the current state, thus forming a continuous memory mechanism to reduce the distance effect (There is distance attenuation in the hi-c interaction matrix, and the further off the diagonal, the less frequency of pair interactions in the genome, and the less biological information it contains). Then use a 3×3 convolutional layer in series to make up for the information loss caused by channel reduction and superimpose it with the global information obtained by skip connection. The final result is convolved to obtain a high-resolution Hi-C matrix.

Fig. 3 Schematic diagram of generator model

2.4. Relative Discriminator

The discriminator network structure is shown in the figure below, and the specific parameters of the network (input channel * output channel * convolution kernel size * step size) are marked in the figure. The data generated by the generator and the real high-resolution data are input into the discriminator, and the detailed features of the data are extracted through the 6-layer convolution operation. The number of Feature Channels is increased every two layers, and the number of Feature Channels in the first layer is 64, and the number of Feature Channels in every two layers is increased by two times until it reaches 256 Channels in the sixth layer. Then, linear transformation data dimension is used, and activation function is used to activate it. Finally, identification results are output through linear transformation again. The results represent the probability that the generated data is more realistic than the real data.
2.5. loss function

In the neural network algorithm based on convolution, any model is optimized based on a preset loss function. The design of the loss function directly affects the overall performance of the model.

In this paper, the HICGAN algorithm loss function is designed as follows: (1)–(7)

\[ L_{\text{HICGAN}} = \{ \min(L_D), \min(L_G) \} \] (1)

\[ L_D = -E_{x_r}[ \log(D_{Ra}(x_r, x_f))] + E_{x_f}[ \log(1 - D_{Ra}(x_f, x_r))] \] (2)

\[ D_{Ra}(x_r, x_f) = \text{sigmoid}(x_r) - \text{IE}[x_f] \] (3)

\[ L_G = L_{\text{MSE}} + L_g + 10^{-7} \times L_{\text{VGG}} \] (4)

\[ L_{\text{MSE}} = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{1}{H \times W} \sum_{j=1}^{H} \sum_{k=1}^{W} (I_{HR}(i,j) - G(I_{LR}(i,j)))^2 \right) \] (5)

\[ L_g = -E_{x_r}[ \log(1 - D_{Ra}(x_r, x_f))] - E_{x_f}[ \log(D_{Ra}(x_f, x_r))] \] (6)

\[ L_{\text{VGG}} = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{1}{H \times W} \sum_{j=1}^{H} \sum_{k=1}^{W} (V(G(I_{LR})) - V(I_{HR}))^2 \right) \] (7)

\[ L_{\text{g}} \] is made up of three parts: \( L_{\text{MSE}} \) calculates the loss of pixels (\( n = 10, H, W \) represents the height and width of the input matrix); \( L_g \) represents the confrontation loss with the discriminator; \( L_{\text{VGG}} \) represents the feature loss of Hi-C data; \( V(G(I_{LR})) \) is the output of that the low-resolution Hi-C data is first reconstructed by the generative network and then input to the trained vgg19. \( V(I_{HR}) \) is the output result that the high-resolution Hi-C data input to the already trained vgg19.

In summary, the purpose of designing generative loss and discriminant loss is to make the generative network and discriminative network continuously optimize themselves during the training process until they reach a balance, that is, neither can become better;
3. Evaluation of HicGAN algorithm Framework

3.1. Enhancement of Hi-C interaction matrix for different cell types
In order to quantitatively evaluate the performance of HicGAN, the Pearson correlation coefficient between the experimental high-resolution matrix and the HicGAN enhancement matrix at each genome distance was calculated. We used the generative model trained on GM12878 with down sampling ratio 1/16 to enhance the low-resolution K562 data. As a result, the Pearson correlation coefficient between real Hi-C data and Hi-C data generated by HicGAN on each chromosome is better than HiCPlus. The figure below shows the results for chromosomes 15 and 17. It is shown that the local patterns / features captured by HicGAN from the low-resolution Hi-C matrix can be used to enhance the Hi-C matrix of other cell types.

![Fig. 5 Comparison of Pearson correlation coefficient of K562 chr15 and K562 chr17](image)

3.2. HicGAN recognizes chromatin interactions and states
Previous experiments have shown that HicGAN can improve the resolution of under sequenced Hi-C samples. Then we study whether HicGAN can help to recognize chromatin contact or chromatin loops. Therefore, we use fit-hi-c software (ay et al., 2014) as the caller of chromatin loop to identify significant chromatin loops with high-resolution Hi-C data and predicted Hi-C data. we first trained a HicGAN model on chromosome 1-10 of GM12878 cell type, and then used the trained HicGAN model to predict the Hi-C samples of chromosome K562. Since not all chromatin interactions are equally important, what we are really interested in is the rich interaction for regulatory elements such as promoters and enhancers. Therefore, we apply fit-hi-c tool to the prediction of Hi-C data in K562 cell type and the actual high-resolution Hi-C data, respectively, calling significant chromatin rings with strict thresholds (q value <1e-06). Then, we filtered the so-called chromatin loop, and only kept the significant chromatin loop at a genomic distance of 50 to 200 KB. The figure below shows the measurement results of chromatin loops chromosome 18 of k652. Experiments show that fit-Hi-C detected 653, 3375, 4166, and 3944 significant interaction pairs in low-resolution, HiCPlus enhanced data, the results predicted in this paper, and experimental high-resolution Hi-C contact matrices. There are 223, 1269, and 1898 common interaction pairs detected in the low-resolution, HiCPlus and the Hi-C contact matrices predicted in this paper and the true high-resolution Hi-C. This shows that the chromatin loops enriched in the low-resolution Hi-C data are only a small part of the chromatin loops in the high-resolution Hi-C data. However, using HicGAN model, we can significantly increase the number of chromatin enriching loops, so as to achieve the same level as high-resolution Hi-C data.
Fig. 6 Statistical analysis of significant interactions based on Fit-Hi-C on chromosome 18 of K562;

We evaluated the HiCGAN model from the aspect of chromatin state. Gm12878 definition of chromatin state can be downloaded from http://rohsdb.cmb.usc.edu/GBshape/cgi-bin/hgFileUi?db=hg19&g=wgEncodeAwgSegmentation. It is generated by ChromHMM (Ernst and Kellis, 2012) software based on encoded data. The details of chromatin state and the relationship between chromatin state and important interaction matrix can be found in the introduction. Using fit-hi-c, we can detect important chromatin interactions based on real high-resolution Hi-C data. The chromatin fragments of the important interaction are gathered in a pool. ChromHMM is then executed in the chromatin fragment pool, calling 12 chromatin states. In this way, the interaction detected on the real high-resolution Hi-C data produces enrichment profiles of 12 chromatin states. In this paper, the low resolution (down sampling 1 / 16), hicplus-enhanced and HiCGAN-enhanced Hi-C data are processed in the same way, and the enrichment mode related to Hi-C data is obtained. The results show that the enrichment pattern of Hi-C data enhanced by HiCGAN is more similar to that of real high-resolution Hi-C data; the following figure shows the enrichment pattern related to chromosome 15.

Fig. 7 Chromatin status analysis based on ChromHMM marker.

4. Conclusion
In addition, we designed a series of experiments to verify the quality of Hi-C samples generated by HiCGAN model. The experimental results show that the Hi-C samples generated by HiCGAN are highly similar to the original high-resolution Hi-C data. HiCGAN has a strong learning ability, which effectively improves the resolution of low-resolution Hi-C data. More importantly, HiCGAN can help us to identify the significant chromatin interactions, chromatin loops and chromatin states and other
information in low-resolution Hi-C data that has lost some boundary information. A typical scenario of using HiCGAN is to apply it to low resolution Hi-C samples which cannot obtain corresponding high-resolution Hi-C data. Some meaningful interactions (such as promoter enhancer) and the position of Hi-C peak may not be detected in low resolution Hi-C data, but we may use HiCGAN model to recover these important information.

To sum up, HiCGAN proposes an end-to-end solution to improve the resolution of under sequenced Hi-C data. Our research not only provides a new method for computing high-resolution Hi-C data, but also helps to decipher the complex mechanism behind 3D genome organization.

Acknowledgments
This work was supported by the National Natural Science Foundation of China (61762089, 91631305, 61863036).

References
[1] Lieberman-Aiden, E., et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 2009;326(5950):289-293.
[2] Dekker,J, et al. Capturing chromosome conformation. Science 2002;295(5558):1306-1311.
[3] Zhao,Z., et al. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. Nat Genet 2006;38(11):1341-1347.
[4] Dostie,J., et al. Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome Research 2006;16(10):1299-1309.
[5] Engreitz J.M., et al. The Xist IncRNA exploits three-dimensional genome architecture to spread across the X chromosome. Science 2013; 341 (6147): 1237973.
[6] Dixon, J.R., et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 2012;485(7398):376-380.
[7] Rao, S.S., et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell 2014;159 (7): 1665-1680.
[8] Bonev, B., et al. Multiscale 3D genome rewiring during mouse neural development. Cell 2017;171(3):557-572. e524.
[9] Zhang, Y., et al. Enhancing Hi-C data resolution with deep convolutional neural network HiCPlus. Nature communications 2018; 9 (1): 750.
[10] Wang, Y., et al. Predicting DNA Methylation State of CpG Dinucleotide Using Genome Topological Features and Deep Networks. Scientific Reports 2016; 6: 19598.
[11] Dong, C., et al. Learning a deep convolutional network for image super-resolution. In, European Conference on Computer Vision. Springer; 2014. p. 184-199.
[12] Kim, J., Kwon Lee,J., and MuLee,K. Accurate image super-resolution using very deep convolutional networks. In, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016. p. 1646-1654.
[13] Kim, J., Kwon Lee,J., and MuLee,K. Deeply-recursive convolutional network for image super-resolution. In, Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. p. 1637-1645.
[14] Tai, Y., Yang, J. and Liu, X. Image super-resolution via deep recursive residual network. In, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017. p. 5.
[15] Tai, Y., et al. Memnet: A persistent memory network for image restoration. In Proceedings of the IEEE International Conference on Computer Vision. 2017. p. 4539-4547.
[16] Xintao Wang. ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks.
[17] Ay,F. et al. (2014) Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. Genome Res., 24, 999–1011.
[18] Quinodoz, S.A. et al. (2018) Higher-order inter-chromosomal hubs shape 3D genome organization in the nucleus. Cell, 174, 744–757.

[19] Schmitt, A.D. et al. (2016) Genome-wide mapping and analysis of chromosome architecture. Nat. Rev. Mol. Cell Biol., 17, 743.

[20] Uhler, C. and Shivashankar, G. (2017) Regulation of genome organization and gene expression by nuclear mechanotransduction. Nat. Rev. Mol. Cell Biol., 18, 717.

[21] Phillips-Cremins, J.E. et al. (2013) Architectural protein subclasses shape 3D organization of genomes during lineage commitment. Cell, 153, 1281–1295.

[22] Yu, M. and Ren, B. (2017) The three-dimensional organization of mammalian genomes. Annu. Rev. Cell Dev. Biol., 33, 265–289. Science, 1989.