Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Subject Section

The Hi-C Super-resolution problem: a survey and analysis of deep learning methods for enhancing experimental Hi-C data

Jiaqi Zhang [1,*], Tyler DeFroscia [2] and Michael Nisenzon [2,*]

[1]Department, Institution, City, Post Code, Country and
[2]Department, Institution, City, Post Code, Country.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

In recent years, Hi-C experiments have been widely used to analyze chromatin interactions. But many applications regarding using Hi-C data are facing the problem that available Hi-C data have low resolution, which will hurt the related analysis. To address this problem, deep learning models such as a convolutional neural network (CNN) and a generative adversarial network (GAN) are used to enhance data resolution due to their effectiveness in various image processing tasks. The estimations of these models indeed increase the data resolution, however, the training cost is a significant increase as well. Previous papers pay little attention to compare the computational resources used in a different model. Moreover, these models consider the Hi-C data as images and apply correlation and image-based metrics to evaluate similarities between their estimations and high-resolution counterparts. Therefore, the feasibility of these models in real applications is doubtful. In this paper, we implement comprehensive experiments to compare most of the ad hoc models on enhancing Hi-C resolution and utilize four biologically plausible similarity scores to measure the estimation. Based on the experimental results, we give guidance on how to choose from various methods to best fit the application requirement and available computational resources.

1 Introduction

Hi-C data allows researchers to examine the spatial chromatin structure of cells, providing an important viewpoint for analyzing DNA accessibility and its downstream impact. However, Hi-C data can be time consuming and expensive to collect, especially at high read depths. High read depth data provides a higher resolution view of the chromatin structure and is naturally more useful to experimental researchers. The problem of imputing, upscaling, or upsampling low read depth Hi-C data reliably to high read depth data, which we refer to as the *super resolution problem* is therefore of great interest to both computational and biological genomic researchers.

There have been multiple proposed solutions to this problem, ranging from standard matrix imputation techniques to cutting edge deep learning models. Research suggests that as these models become more complex and leverage new ideas in deep learning their ability to reconstruct high read depth Hi-C data increases. Models such as CNNs and even GANs are being used to upsample low read count data. With each new model comes analysis suggesting that it is the new state of the art in the super resolution problem, with the authors usually benchmarking their model against a previous one using either mean squared error, Pearson correlation, or another accuracy metric.

However, there has not been a concerted effort by the community to analyze these different methods in the context of their use for experimental researchers. In this paper, we seek to analyze the current state of the art in the super resolution problem, providing computational costs, a suite of accuracy metrics, and overall guidance to experimental researchers looking to apply these models in their work. As such, we aim to provide a full comparative analysis of the current state of the art, with an emphasis of practical application of the models in the field. We report accuracy on a wide variety of metrics designed to capture different aspects of the quality of the upsampled data, as well as computational costs of training and inference, and finally guidance for experimental researchers looking to apply a model to their work. We hope these data and analyses will provide the context necessary for experimental researchers to use Hi-C super resolution methods in practice.

Notations: This paper uses bold uppercase letters such as $\mathbf{A}$ and $\mathbf{B}$ to represent matrices. Let calligraphic letters, such as $\mathcal{M}$ and $\mathcal{D}$, denote functions or models.

2 Related Work

2.1 Super Resolution Problem

The goal of super resolution (SR) problems is to predict a high-resolution image from a low-resolution input image. SR is a classical problem in the field of computer vision. Specifically, given N low-resolution images $\{\mathbf{X}_{\text{LR}}^{(i)}\}_{i=1}^{N}$, the SR problems aim at estimating N images $\{\mathbf{X}_{\text{HR}}^{(i)}\}_{i=1}^{N}$ with higher resolution through a mapping function (model) $\mathcal{M}(\cdot)$ as

$$\mathbf{X}_{\text{HR}}^{(i)} = \mathcal{M}(\mathbf{X}_{\text{LR}}^{(i)}), \quad \text{for } i = 1, 2, \cdots, N. \tag{1}$$

Some state-of-the-art models, such as Bevilacqua *et al.* (2012); Chang *et al.* (2004); Timofte *et al.* (2013), are proposed to learn the mapping using from low-resolution images to their high-resolution counterparts based on neighboring regression or sparse coding.

2.2 Deep Learning Models for the Super Resolution Problem

Because of its rapid development, deep learning models are widely used in many fields, such as image processing Maier *et al.* (2019); Razzak *et al.* (2018); Yamashita *et al.* (2018). Based on the powerful ability of convolutional neural networks (CNN) to process image data, Dong *et al.* (2014) proposed Super Resolution CNN (SRCNN) to predict high-resolution images with the help of end-to-end CNN, which significantly outperforms previous works. SRCNN minimizes the mean squared error (MSE)

$$\arg\min_{\mathcal{M}} \frac{1}{N} \sum_{i=1}^{N} \|\mathcal{M}(\mathbf{X}_{\text{LR}}^{(i)}) - \mathbf{X}_{\text{LR}}^{(i)}\|_2^2 \tag{2}$$

for reconstructions. However, it is proved that the pixel-wise MSE leads to blurry images and is not a good loss function for SR problems. Because a blurry filter is likely to have the lowest loss and the easiest solution to converge. The blurry effect makes MSE-based models fail in restoring fine texture details. In the Hi-C matrices, these sharp details correspond to loops and boundaries.

To address such a problem and obtain photo-realistic images, the generative adversarial network (GAN) Goodfellow *et al.* (2014) is utilized to approximate the distribution of realistic images, instead of just learning the pixel-wise projection from low-resolution to high-resolution data. For example, Goodfellow *et al.* (2014) proposed Super Resolution GAN (SRGAN) enhancing image resolutions based on a GAN. SRGAN uses a generator network $\mathcal{M}(\cdot)$ to generate estimations of high-resolution data and employs a discriminator network $\mathcal{D}(\cdot)$ for distinguishing generator estimations from true high-resolution samples. Two networks compete against each other to obtain a good distribution approximation of high-resolution images through minimizing reconstruction errors of the generator network together with maximizing the log-likelihood of the discriminator network as

$$\arg\min_{\mathcal{M}} \max_{\mathcal{D}} \mathbb{E}_{\mathbf{X}_{\text{HR}}} \left[\log \mathcal{D}(\mathbf{X}_{\text{HR}})\right] + \mathbb{E}_{\mathbf{X}_{\text{LR}}} \left[\log(1 - \mathcal{D}(\mathcal{M}(\mathbf{X}_{\text{LR}})))\right]. \tag{3}$$

Benefiting from its good performance in generating more visually appealing results, GAN is widely used for SR problems.

After experiencing the benefits of deep learning models, various deep learning techniques and architectures are introduced in this field to make further improvements. Many models (Kim *et al.* (2016); Lim *et al.* (2017); Zhang *et al.* (2018c); He *et al.* (2016); Zhang *et al.* (2018b)) utilize residual architecture to obtain better stabilization for very deep networks. Using skip connections in networks, residual architecture reuses outputs of previous layers to avoid gradient vanishing problems. In another line of works, researchers propose perceptual-based models to improve the visual quality of estimations and obtain photo-realism. For instance, Johnson *et al.* (2016) minimizes the loss in the feature space rather than the pixel space, and Wang *et al.* (2018) integrates semantic prior in an image for improving the recovered textures. These deep learning models have shown promising good performance in SR problems.

2.3 Hi-C Super Resolution Problem

As one of the genome-wide technology, Hi-C data measures the contacts between every pair of genome regions and represent it as a symmetric matrix called contact matrix. The resolution of the contact matrix depends on the size of genome regions bins and high-resolution Hi-C data leads to effectiveness in related analysis tasks. However, available Hi-C data are relatively low-resolution that cannot meet the needs of accurate analysis. To address this problem, if considering the Hi-C contact matrix as image-like data, the above-mentioned deep learning models can also be used for improving Hi-C data resolution in a very similar way.

2.4 Deep Learning Models for Hi-C Super Resolution Problem

The current state of the art focuses on two main models: CNNs and GANs. The HiCPlus proposed by Zhang *et al.* (2018a) is a CNN that interprets the low-resolution HiC matrices as images, applying three layers of convolutions of varying kernel sizes to first extra feature maps from the low-resolution matrices and then using a dense neural network to impute a high-resolution matrix. Moreover, HiCNN (Liu and Wang (2019a)) is an even deeper CNN with a residual block and HiCNN2 (Liu and Wang (2019b)) is an extension of HiCNN considering ensembles of more than one HiCNN CNNs to impute high-resolution HiC data. SRHiC (Li and Dai (2020)) is a recent model that adds a second residual block to its CNN architecture.

Another main research direction is the use of GANs. The first GAN used in the Hi-C super resolution task was HiCGAN (Liu *et al.* (2019)), a GAN consisting of a CNN discriminator and a dual-stream residual generator that learns to generate high-resolution data from low-resolution input matrices via adversarial training. Another GAN model, DeepHiC (Hong *et al.* (2020)), takes a similar approach but uses a different loss function and a deeper generator. Basically, GAN-based models outperform other models on recovering high-resolution data. However, GAN is well-known to be difficult to converge, which limits its usage in some real-world scenarios.

The above-mentioned models are empirically shown to be effective in enhancing Hi-C data resolution. Unfortunately, the authors of these papers don't directly compare their results with other counterparts. So it's impossible to call one the state of the art over the other more theoretically, although visually one can distinguish the best-imputed matrices in some cases. In this paper, we implement integrated experiments to comprehensively compare various models on upscaling the Hi-C data resolution.

3 Methods

Our experimental design focuses on first training each Hi-C super-resolution model on a training set of Hi-C data, and then comparing results on a held-out test set using a variety of metrics. We also record how long each model takes to train to convergence on standardized hardware.

picture(0,0)(-35,0)(1,0)30 (0,35)(0,-1)30 picture

picture(0,0)(35,0)(-1,0)30 (0,35)(0,-1)30 picture

The Hi-C dataset used for training and testing is the same as used in (Liu *et al.* (2019)). This dataset consists of Hi-C data across four cell types (GM12878, K562, IMR90, and NHEK) downloaded from the GEO database. Cell types with data from multiple experiments are pooled and aligned using the same technique as (Liu *et al.* (2019)). When preprocessing is required, such as to transform data into a matrix for input into a CNN, the preprocessing steps specified by the original authors of the model are followed. We used a 90%/10% test/train split for training and evaluating each model.

The following public implementations of the models were used:

- HiCPlus (`https://github.com/zhangyan32/HiCPlus`)
- hicGAN (`https://github.com/kimmo1019/hicGAN`)
- hiCNN (`https://github.com/liambai/HiCNN`)
- SRHiC (`https://github.com/hzlzldr/SRHiC`)
- DeepHiC (`https://github.com/omegahh/DeepHiC`)
- Matrix completion model as a baseline (`https://github.com/iskandr/fancyimpute`).

Each machine learning model was trained using best known hyperparameters and allowed to train until convergence on GPU. The time require to train until convergence was recorded and is noted as *training time* in the Results section. Non-machine learning models are assigned a training time of zero.

After training, each model is evaulated on the held-out test set using a suite of metrics. Given a ground truth high read count experiment, the model is used to upscale a low read count representation, and the output high read count data is compared with the ground truth using a suite of accuracy metrics. The amount of time required to run inference on the validation set is recorded as *inference time*. We evaluated the models using the following metrics:

- *Pearson correlation*. Given the predicted data and the ground truth data, we compute the Pearson correlation coefficient between the two experiments. A higher correlation indicates a more accurate upscaled representation.
- *Mean squared error*. Given ground truth data and a prediction, we compute the component-wise mean squared error. A lower error indicates a more accurate prediction.
- *Image-based metrics*. Following the common technique of representing Hi-C data matrices as images, we apply evaluation techniques commonly used in computer vision to gauge performance. Signal-to-noise ratio (SNR) is a metric that measures the amount of degradation in an image, and is defined as

$$\text{SNR} = \frac{\sum_{x=0}^{M-1}\sum_{y=0}^{N-1}\hat{f}(x,y)^2}{\sum_{x=0}^{M-1}\sum_{y=0}^{N-1}[f(x,y)-\hat{f}(x,y)]^2} \tag{4}$$

  where $\hat{f}(x,y)$ is the x,y entry of the imputed Hi-C matrix and f is the ground truth matrix. The Structure similarity index (SSIM) is a metric that captures the overall quality of an image, taking into account accuracy to the ground truth image as well as its overall resolution and quality.
- *Hi-C based metrics*. We also use a variety of pre-implemented metrics for comparing the results of Hi-C experiments, such as GenomeDISCO, Hi-C Spector, HiCRep, and QuASAR-Rep to determine fidelity of the upsampled matrix.

These accuracy results are recorded along with the *computation time* of the model, defined as the sum of training time and inference time.

4 Results

As the models haven't been trained and evaluated yet, we will simply report our current predictions. Computationally speaking, we expect GANs to take much longer than standard training time due to the need to construct adversarial samples and previously reported training times, although the actual inference time should be negligible. We now examine our predictions for correlations.

HiCPlus, the first convolution deep learning approach to super-resolution, proved to be comparable to existing methods in terms of Pearson correlation. Deeper CNN networks, such as HiCNN, led to substantial increases in Pearson correlation with the ground truth data, as well as decreases in mean squared error. We expect to replicate the increase in Pearson coefficient of GANs in DeepHiC over the existing CNN models, although the increase was only 5%, which isn't a very significant difference between state-of-the-art CNN and GAN models.

Based on the greater weight of the mean-squared error as loss function in CNN-based models, we predict that training will lead to a smaller error from and a more accurate prediction of the ground truth as opposed to GANs. Notably, a mean squared error analysis was absent from previous comparisons of Hi-C GANs to existing CNN alternatives.

In Hi-C data, there is a much stronger tendency for nearby regions to overlap, leading to the existence of local patterns. We predict that the ability of convolutional deep learning models to capture these local patterns will minimize image and Hi-C based metrics from the ground truth as opposed to GANs.

5 Discussion

As more and more Hi-C experiment data becomes available, the Hi-C super resolution problem becomes more pertinent to researchers. As we have seen, current methods vary in their ability to produce accurate and high resolution upscaled low read count data. However, for experimental researchers with limited access to computing hardware, choosing a model isn't as simple as taking the highest accuracy performer on a given test set. Rather, a balance has to be struck between model complexity, computational cost, and accuracy and quality of results. We believe that our suite of evaluation metrics and computational costs will provide researchers with the data needed to make an informed choice of model.

As solutions to the super resolution problem become more robust, more researchers will rely on these models in their experiments. Experimental consistency and reproducibility requires that researchers use models that are sufficiently accurate while still generally accessible. Given its relatively high accuracy and low computational cost, we suggest researchers use CNN based upscaling methods, such as HiCPlus. These methods strike a balance between performance and accessibility while staying relatively interpretable.

While GAN based techniques are exciting and show the potential for very high performance, the relative difficulty of training a GAN from scratch and their high computational costs limit their use to researchers in the field. As computational hardware and training techniques evolve, GAN-based techniques will become more accessible: future work can focus on re-evaluation of these techniques as the field develops.

6 Proposed Timeline

We propose the following rough timeline:

- Oct 30-Nov 6: Implement models and pre-process experimental data as required for each model. Model implementation will use the publicly available implementations described in Methods. The goal is to have each model ready for training by the end of this stage.

picture(0,0)(-35,0)(1,0)30 (0,-35)(0,1)30 picture

picture(0,0)(35,0)(-1,0)30 (0,-35)(0,1)30 picture

- Nov 6-Nov 20: Train models. This is a very conservative timetable that assumes model training will be expensive and potentially difficult, and will potentially be shorter depending on our access to hardware and ability to parallelize the training process. We also plan to implement the evaluation metrics and evaluation pipeline during this training period. The goal is to have each model fully trained and ready for evaluation by the end of this time period.
- Nov 20-Nov 27: Run evaluation and analysis of models. This consists of running models through the evaluation pipeline, recording accuracy statistics, and generating graphs and other analyses for the final paper. By November 27, we hope to have all the data needed to fully write our final draft of the research paper.
- Nov 27-Dec 4: Incorporate results and analyses into final draft of research paper. This probably will not take a full week, so one can think of this period as a bit of "slack" in case difficult arises in any of the above time periods.

7 Conclusion

We analyzed existing methods to predict high resolution HiC matrices data from low resolution sources across multiple deep learning models. Currently, deep learning methods ares state-of-the-art, but at the moment, to the best of our knowledge, we are the first to systematically compare different approaches. Previous papers tend to focus on one metric over another, so we lack a nuanced understanding of how particular implementations CNNs and GANs compare on desired metrics. We will update with preliminary results along our timeline.

We could further extend our approach by testing specific properties of genome in Hi-C data across different models, such as chromatin interactions and topological associating domains (TADs), which we could do with AUC scores. Ultimately, visualization is intended to help detect useful patterns, allowing us to apply our results to infer new structures along the genome. Perhaps it's worth considering how to weigh which structures we wish to identify. We predict that given the ability of GANs to preserve finer details, it is likely that advances in computing power would further strengthen the case for GANs in the near future for feature detection. On the other hand, in the absence of such a specific focus, general matrix correlation along our suite of metrics would provide a useful baseline for making a choice between models.

Other tradeoffs worth considering would be the extent of upscaling we wish to achieve. The downsampling ratio from past CNN papers suggests that GANs can have comparable performance with lower resolution starting data, with downsampling of 1% in DeepHiC as opposed to 1/16 in HiCNN.

Acknowledgements

To be filled in later.

Funding

The authors have no outside sources of funding to disclose.

References

Bevilacqua, M., Roumy, A., Guillemot, C., and Alberi-Morel, M. L. (2012). Low-complexity single-image super-resolution based on nonnegative neighbor embedding.

Chang, H., Yeung, D.-Y., and Xiong, Y. (2004). Super-resolution through neighbor embedding. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 1, pages I–I. IEEE.

Dong, C., Loy, C. C., He, K., and Tang, X. (2014). Learning a deep convolutional network for image super-resolution. In *European conference on computer vision*, pages 184–199. Springer.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Hong, H., Jiang, S., Li, H., Du, G., Sun, Y., Tao, H., Quan, C., Zhao, C., Li, R., Li, W., *et al.* (2020). Deephic: A generative adversarial network for enhancing hi-c data resolution. *PLoS computational biology*, **16**(2), e1007287.

Johnson, J., Alahi, A., and Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer.

Kim, J., Kwon Lee, J., and Mu Lee, K. (2016). Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1646–1654.

Li, Z. and Dai, Z. (2020). Srhic: A deep learning model to enhance the resolution of hi-c data. *Frontiers in Genetics*, **11**, 353.

Lim, B., Son, S., Kim, H., Nah, S., and Mu Lee, K. (2017). Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144.

Liu, Q., Lv, H., and Jiang, R. (2019). hicgan infers super resolution hi-c data with generative adversarial networks. *Bioinformatics*, **35**(14), i99–i107.

Liu, T. and Wang, Z. (2019a). Hicnn: a very deep convolutional neural network to better enhance the resolution of hi-c data. *Bioinformatics*, **35**(21), 4222–4228.

Liu, T. and Wang, Z. (2019b). Hicnn2: Enhancing the resolution of hi-c data using an ensemble of convolutional neural networks. *Genes*, **10**(11), 862.

Maier, A., Syben, C., Lasser, T., and Riess, C. (2019). A gentle introduction to deep learning in medical image processing. *Zeitschrift für Medizinische Physik*, **29**(2), 86–101.

Razzak, M. I., Naz, S., and Zaib, A. (2018). Deep learning for medical image processing: Overview, challenges and the future. In *Classification in BioApps*, pages 323–350. Springer.

Timofte, R., De Smet, V., and Van Gool, L. (2013). Anchored neighborhood regression for fast example-based super-resolution. In *Proceedings of the IEEE international conference on computer vision*, pages 1920–1927.

Wang, X., Yu, K., Dong, C., and Change Loy, C. (2018). Recovering realistic texture in image super-resolution by deep spatial feature transform. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 606–615.

Yamashita, R., Nishio, M., Do, R. K. G., and Togashi, K. (2018). Convolutional neural networks: an overview and application in radiology. *Insights into imaging*, **9**(4), 611–629.

Zhang, Y., An, L., Xu, J., Zhang, B., Zheng, W. J., Hu, M., Tang, J., and Yue, F. (2018a). Enhancing hi-c data resolution with deep convolutional neural network hicplus. *Nature communications*, **9**(1), 1–9.

Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., and Fu, Y. (2018b). Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 286–301.

Zhang, Y., Tian, Y., Kong, Y., Zhong, B., and Fu, Y. (2018c). Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2472–2481.