Enhancing Hi-C data resolution with deep convolutional neural network HiCPlus

Kaustubh Deshpande

**Introduction**

DNA in living cells exists within a highly compacted and structured dynamic state. A cell’s transcription machinery compels access to this DNA in order to carry out its work. When chromatin de-condenses, DNA is vulnerable to entry of molecular mechanisms. This conveys to us that transcription is therefore contingent on the structural state and that the sequence alone does not have the final say. **[1]** This knowledge has driven study of chromosomal design and long-range chromatin interactions. To achieve this goal, chromosome conformation capture techniques have become pertinent over the years.Chromosome confirmation procedures are used for analysis of spatial organization of chromatin within a cell. These methods calculate the connections between genomic loci that are nearby in 3-D space. These interaction frequencies are both analyzed directly, and they are translated to distances and subsequently used to reconstruct 3-D structures. **[1]**

The chromosome conformation capture (3C) experiment computes interactions amid a single pair of genomic loci. 3C can be used to check a potential promoter-enhancer interaction. Chromosome conformation capture-on-chip (4C) apprehends interactions between one locus and all genomic loci. 4C technique does not need the previous knowledge of both interacting chromosomal regions. Chromosome conformation capture carbon copy (5C) perceives interactions between all constraint fragments within a region, with this region's size normally no greater than a mega base. This method is unsuitable for running genome-wide complex interactions as that will require millions of 5C primers. **[1]** Hi-C uses high-throughput sequencing to obtain the nucleotide sequence of fragments and makes use of paired end sequencing, which obtains a short sequence from ligated fragment terminals. **[1]** Thus, two sequences attained should represent by two different restriction fragments that were ligated together during the ligation step. The pair of sequences are independently united to the genome, thus determining the fragments involved in that ligation event. Thus, all probable pairwise interactions between fragments are verified. **[1]**

A screenshot of a cell phone

Description automatically generated

High-throughput chromosome conformation capture (Hi-C) technique is an efficient tool for studying spatial organization of chromosomes. Hi-C data is presented in the form of a n × n contact matrix. **[1]** The genome is distributed into n equally sized bins. The value in each cell of the matrix designates the quantity of pair-ended reads between a pair of bins. The bin size is referred to as 'resolution'. Resolution is vital as it affects down the line analysis. For example, predicting enhancer–promoter interactions. Although Hi-C technology is one of the most popular tools for studying 3D genome organization, due to sequencing cost, the resolution of most Hi-C datasets are coarse and cannot be used to link distal regulatory elements to their target genes. **[1]** Due to high sequencing cost, a majority of available Hi-C datasets possess relatively low resolution such as 25 or 40 kb, as the linear surge of resolution requires a quadratic rise in the total quantity of sequencing reads. Low-resolution Hi-C datasets are used to define large-scale genomic patterns like A/B compartment or TADs. However, they cannot be used to recognize more refined structures like sub-domains or enhancer–promoter interactions. **[1]** With all these factors in mind, the original researchers felt a need to develop a computational method to make better use of currently available Hi-C datasets and produce higher-resolution Hi-C interaction matrix.

The original researchers developed HiCPlus. Their project was an initial attempt to infer high-resolution Hi-C interaction matrices from low-resolution Hi-C samples. HiCPlus was applied to get high-resolution matrices for 20 tissue/cell lines from the low-resolution Hi-C datasets that were obtained using only 1/16th of the sequence reads. **[1]** Their work provided means to study chromatin interactions. Their research also established a method to predict high-resolution Hi-C matrix with a fraction of sequencing cost and identified latent features fundamental to the formation of 3D chromatin interactions such as loops.

**Method and Results**

The training process for the HiCPlus algorithm essentially divides the input Hi-C matrix into multiple square-like sub-regions with permanent size, and each sub-region is treated as one sample. **[2]** Once the sectioning is complete, the ConvNet ascertains the mapping relationship between the inputted high-resolution Hi-C matrix and the inputted low-resolution Hi-C matrix at feature levels. Once the model is trained, it can be applied it to enhance any Hi-C interaction matrix with low-sequencing depth. **[2]** Once each block of interactions has been predicted, all the blocks are amalgamated into chromosome-wide interaction matrix. As the samples have a adjacent padding region that is disconnected during the prediction by ConvNet, an overlap is indispensable when dividing the Hi-C interaction matrix to the samples. This pipeline can be visualized efficiently in the following image. **[1]**

A close up of a map

Description automatically generated

Similar to the original research paper, I first obtained data from online websites. I was unable to use the original dataset as it was not provided in the GitHub repository. **[4]** Consequently, I downloaded a high-resolution matrix and then down sampled the values by a factor of 1/16 sequence reads to obtain a low-resolution matrix. **[3]** I consider the matrix generated from down-sampled sequencing reads as low resolution as it would have been processed at a lower resolution than the sequencing depths in practice. Next, I then trained my own model using values at each position in the high-resolution matrix as the response variable and using its neighboring points from the down-sampled matrix as the predictors. My goal was to verify the original researchers claim that the ConvNet framework can precisely predict values in the high-resolution matrix using values from the low-resolution matrix. The major difference between my model and the researcher’s model was substituting ELU (Exponential Linear Unit) instead of ReLU (Rectified Linear Unit) as my activation function. ELU is similar to ReLU except for negative inputs. ELU smoothens slowly whereas ReLU sharply smoothens. **[2]**

A picture containing screenshot

Description automatically generated

I wanted to replicate the original results using another dataset and I was successfully able to do that. I trained a model and then tested it on the downloaded data set. Similar to the original study, the trained model was able to infer high-resolution Hi-C interaction matrices from low-resolution Hi-C samples. The image below depicts the results obtained by the original researchers in comparison to the results that I obtained testing on another dataset. The final output was the enhanced image that can be seen in the column in the middle.

A screenshot of a cell phone

Description automatically generated

**Discussion and Future Application**

For this project, I attempted to replicate the results of HiCPlus, while making minor modifications to the hyperparameters and configuration parameters. My final code structure along with the values used for training can be found in the GitHub repository link I have provided in the appendix. For reference, I have included some code samples in the appendix as well. Through my results I verified the original claim that the framework can construct an interaction matrix with similar quality despite only using 1/16th of the sequencing reads. I applied my modified HiCPlus model to generate high-resolution matrices for a new hic dataset, which only contained low-resolution Hi-C data. I successfully tested the claim that using the Hi-C plus algorithm you can enhance low-resolution Hi-C data matrix to higher-resolution Hi-C data, which is essentially an image-enhancing problem. My takeaway from this project is that HiCPlus is trained by learning certain patterns and information from the training datasets. These patterns are used for enhancing Hi-C matrix in the prediction process. A potential topic for future studies can be to further explore these patterns and hopefully obtain a more biologically meaningful interpretation of the results.

The original GitHub repository contained multiple syntactic issues within the code base. **[4]** I believe this is because the code may have been outdated. Additionally, due to the format of the pretrained model, I was unable to load and use it to form predictions. In order to overcome this obstacle, I was forced to adjust my project. I trained my own model and saved the output in standard PyTorch (.pth) format. My model was used to make the predictions which are displayed in the images above. After my final presentation for this project, I was able to obtain more hic datasets using the resources listed by other students. My intention was to cross validate my model using these resources. However, these hic datasets were much larger in comparison to the ones I had originally used. Since my laptop doesn’t have a GPU, I was unable to successfully run the algorithm on the newly obtained datasets. My laptop overheated causing the program to freeze. As a result, I could not preform validation on a novel dataset.

Given access to a strong GPU, I would have liked to conduct some additional evaluation tests. The original hypothesis is that the Hi-C matrix contains repeating local patterns, and the interaction intensity of each point is not independent to its local neighboring regions. As a result, it should be feasible to predict the interaction frequency of any cell in the Hi-C matrix with the interaction frequencies from its neighboring regions. To test this hypothesis, I would have used my trained model to systematically predicted interaction matrices in another set of chromosomes. As a final evaluation I would have computed both the Pearson and Spearman correlation coefficients between the predicted values and the real values at each genomic distance. **[1]**

Another major point of discussion was the use of a common model between varying cell types. The original researchers claimed and verified that a model trained on a particular cell type could be used to form accurate predictions on low resolution Hi-c matrices for a different cell type. More specifically, the researchers trained the ConvNet model on three different cell types (GM12878, K562, IMR90) with similar sequencing depths and tested their prediction performances in K562 cells. **[1]** While looking for data at the start of my project, I found hic data for various mouse cells. It would be interesting to test whether a ConvNet trained on human cells hi-c matrices can be used to form accurate predictions for mouse hi-c interaction matrices. This is something that demands strong computing power and thus once again a GPU is required to carry out this testing.

In summary, HiCPlus presents the first deep learning framework for enhancing the resolution of Hi-C interaction matrices. By leveraging interaction frequencies from neighboring regions and learning regional patterns from available high-resolution Hi-C data, HiCPlus can generate high-resolution Hi-C interaction matrices at a fraction of the original sequencing reads. **[1]** Based on my experience this quarter, I believe that HiCPlus is a powerful tool for the study of 3D genome regulation and gene organization.

**References**

[1] Zhang, Yan, et al. “Enhancing Hi-C Data Resolution with Deep Convolutional Neural Network HiCPlus.” *Nature News*, Nature Publishing Group, 21 Feb. 2018, [www.nature.com/articles/s41467-018-03113-2](http://www.nature.com/articles/s41467-018-03113-2).

[2] “Convolutional Neural Network Architecture: Forging Pathways to the Future.” *MissingLink.ai*,missinglink.ai/guides/convolutional-neural-networks/convolutional-neural-network-architecture-forging-pathways-future/.

[3] “Welcome to Hi-C Project at Ren Lab!” *Welcome to the Hi-C Project at Ren Lab!*, chromosome.sdsc.edu/mouse/hi-c/download.html.

[4] wangjuan001. “wangjuan001/Hicplus.” *GitHub*, github.com/wangjuan001/hicplus.

**Appendix**

**1**. My code source: - <https://github.com/kaus0399/Hi_C>

**2. My modified Neural Network class with ELU instead of ReLU.**

A screenshot of a cell phone

Description automatically generated

**3. Prediction images obtained using python interactive and inference scripts.**

A screenshot of a computer

Description automatically generated

**4. 2D-Gaussian-kernel smoothing predictions. ConvNet predictions were much better.**

A picture containing clock

Description automatically generated