CSCI2952G: Deep Learning in Genomics, 2023, 1–5
doi: 2
Advance Access Publication Date: 17 December 2023
Paper

Structural Bioinformatics

DeepBio: Predicting HiC Contact Maps through Multimodal Input

Bumjin Joo, Evan Li, Gabrielle Shieh, Pavani Nerella and Serdar Sungun

[1]Department of Computer Science, Brown University, Providence, 02912, Rhode Island, United States of America

Abstract

Hi-C matrices provide information-rich data regarding the frequency of interactions of DNA sequences to determine the 3-D chromosomal structure of the genome. However, most available Hi-C data have low resolution as high-resolution Hi-C data is expensive and time-consuming to generate. The use of both epigenomic and HiC data improves upon previous approaches by utilizing pre-existing contact sequencing while additionally using meaningful epigenomic data to enhance these maps. The addition of the Generative Adversarial Network Discriminator Loss Term results in more learned latent structures that are essential to Hi-C matrices and enable more accurate predictions regarding long range interactions. Due to strong limitations in computational resources, we were unable to perform many experiments under ideal conditions. However, our model demonstrates improved performance as compared to certain baseline models and illustrates that utilizing pre-existing low-resolution Hi-C maps alongside readily available genomic data could yield higher quality upscaled contact maps.

Introduction

The three-dimensional structure of genomes is central to regulating gene expression. This genomic structure allows for sequence regions that are distant from each other in a flattened, sequential view to be located in close physical proximity to each other, forming chromatin loops. These loops give rise to promoter-enhancer interactions and increase gene transcription. Spatial genome interaction can be represented and understood using the high-throughput chromosome conformation capture (Hi-C) technique, which analyzes three-dimensional chromatin structure and organization to map chromatin interactions and contact (Lieberman-Aiden et al., 2009). With reads from long range genomic interactions, reads can be collected and compiled into a Hi-C matrix in which the frequency of interaction between two regions of a genome are mapped.

One of the many scientific breakthroughs made with Hi-C matrices was the discovery of topologically associating domains (TADs), which are genomic regions with high frequencies of interaction and play a fundamental role in regulating gene expression. Hi-C matrices are widely used by biologists in the realm of transcription regulation, the emergence of diseases, and evolution (Yardımcı et al., 2019). However, gathering the necessary data to produce high-resolution Hi-C data is time-consuming and expensive. Researchers have leveraged techniques such as CNNs, GANs, and RNNs to enhance high-resolution Hi-C data. CNNs, typically used in computer vision, have been adapted for feature extraction in Hi-C data. GANs, composed of a generator and a discriminator component, are notable for generating data imitative to real data and learning latent features, resulting in improved image resolution.

Several previous approaches have attempted to integrate Deep Learning (DL) techniques in an attempt to solve the biological and computational problem. Initial works approached the task through a computer vision approach, attempting to upsample low-resolution contact maps to match high-resolution ground truth Hi-C maps. Later approaches demonstrated the shortcomings of wholly treating Hi-C maps as images by achieving improved high-resolution contact maps by predicting directly on genomic and epigenomic features. In this paper, we introduce DeepBio, a deep learning model that predicts high-resolution Hi-C data through both low-resolution Hi-C data and epigenomic tracks.

Related Works

In previous works centered around the computational complexity of HiC maps, several pioneering methods have helped the path toward an enhanced understanding of intricate biological processes.

Initial approaches into Hi-C data originally focused on Hi-C as an image, and consequently attempted to take in more

available low-resolution Hi-C maps and upsample them into high-resolution reconstructions, utilizing several techniques from similar computer vision super resolution tasks. For instance, Liu et al. introduced hicGAN to attempt a super resolution task on HiC data. Generative Adversarial Networks (GANs) are deep neural networks that consist of two neural networks, the generator network and the discriminator network. The generator attempts to reconstruct realistic data, while the discriminator is trained to distinguish real data from generated data. Working in tandem, the generator starts to get better at producing realistic data and the discriminator becomes better at distinguishing fake data from real data. By utilizing a GAN-style objective function, hicGAN achieves a flexibility to learn complex properties defining realistic high-resolution Hi-C maps without the need to explicitly define them. hicGAN is able to effectively generate matrices closely resembling high-resolution Hi-C counterparts, unraveling essential insights into the formation of chromatin contacts.

Following works attempted to ground predictions of 3D genome structure in genomic or epigenomic features. Fudenberg et al. propose Akita, a Convolutional Neural Network (CNN) architecture designed for predicting 3D genome folding. By employing dilated residual 2D convolutions and positional encoding of genomic distances, Akita learns the DNA motifs, grammar for genome folding, and relationships between features while considering dependencies among neighboring genomic bins in an attempt to predict contact maps. DeepC (Schwessinger et al., 2020) is another model that predicts 3D genome folding from megabase-scale DNA sequences. DeepC utilizes a transfer learning-based deep neural network to accurately predict high resolution Hi-C, including domain boundaries at high resolution, and identifies the sequence determinants of genome folding. DeepC's model training involves two phases; in the first pre-training phase, the model learns to predict a compendium of chromatin features across cell types, which are then used in the second fine-tuning training phase to refine and predict chromatin interactions. DeepC's performance is comparable to Akita.

Epiphany (Yang et al., 2021) marked a significant stride in predicting cell-type-specific Hi-C contact maps. Epiphany integrates bidirectional long short-term memory (Bi-LSTM) and an optional GAN to generate precise Hi-C contact maps from processed input data. The Bi-LSTM layers and convolution modules capture local dependencies of epigenomic tracks and to enhance prediction quality the GAN is trained using a combination of pixel-wise mean squared error (MSE) and adversarial loss. Epiphany not only accurately predicts cell-type-specific 3D genome architecture but also exhibits robust performance across various Hi-C normalization procedures and resolutions. Moreover, it facilitates the interpretation of specific epigenomic signal contributions to local 3D structures through feature ablation and attribution experiments.

Methods

We introduce DeepBio, a multimodal, adversarially trained deep learning model which utilizes both epigenomic and low-resolution Hi-C data to predict high resolution Hi-C maps. Our data samples consist of paired epigenomic tracks and a low-resolution HiC input and a high-resolution HiC matrix label. We source high-resolution (10kb) HiC matrices and their paired epigenomic tacks from Rao et al.. We follow the preprocessing pipeline from Bigness et al. (2021) as introduced in GC-MERGE to retrieve normalized HiC matrices as well as epigenomic features aligned with the bins of the HiC matrices. fff

Data

For our HiC reconstruction prediction task, our data samples consist of paired epigenomic marks and HiC matrix. We source 10kb HiC matrices as our high resolution reference, as well as paired histone modification (HM) marks from Rao et al. (2015) and the ENCODE data repository. Each HiC matrix is normalized with the Knight-Ruiz (or KR) normalization technique. We follow the preprocessing pipeline from Bigness et al. as introduced in GC-MERGE to retrieve normalized HiC matrices as well as epigenomic features aligned with the bins of the HiC matrices.

We then down-sample to $\frac{1}{16}$ of the reads of the high resolution HiC map to form a low resolution HiC matrix. This way, we can efficiently generate *in silico*, low resolution HiC samples without creating the need to experimentally gather more HiC samples. Our model is thus given a tuple of a low resolution HiC matrix, a high resolution HiC matrix, and epigenomic HM marks.

We train on all GM12878 chromosomes except 11 and 17, which are held out for validation, in order to align data usage with other reference works such as Epiphany and Akita.

Architecture

DeepBio's architecture consists of an upsampler network, or autoencoder, and a discriminator network.

The autoencoder takes as input the low resolution HiC map and the paired epigenomic marks. We predict that utilizing low resolution structural information already present in the HiC map alongside epigenomic marks that could dictate the genomic structure will allow us to predict high quality HiC contact maps, relative to baseline studies. Given significant computational bottlenecks, we decided to predict on slices of the HiC map alongside the corresponding slices of the epigenomic marks. We only found it necessary to predict on half of the square HiC map, given the inherent symmetry across the diagonal. First, the HiC input is encoded into a latent vector through a 2D Convolutional network (CNN). Next, The epigenomic marks are encoded with a 1D CNN. The two latent encodings are flattened and concatenated into a single latent vector, which is finally processed through a 2D Transpose Convolutional network to predict the high resolution HiC map.

Previous HiC approaches have classically optimized a pixel-wise Mean Square Error (MSE) metric. This follows the reasoning that HiC matrices are images whose similarity can be accurately measured through pixel differences. However, as the authors of Epiphany demonstrate, MSE loss can give perceptually incorrect regions of significant interactions. The adversarial loss can address this shortcoming by forcing the model to learn more realistic representations of underlying HiC structures. In particular, we utilize both MSE and GAN-style losses; where a convolutional discriminator network takes in either a high resolution HiC matrix or a predicted upscaled HiC matrix and predicts the likelihood of realism. We introduce a weight hyperparameter $\lambda = 0.65$, in accordance with Epiphany (Yang et al., 2021) to balance the linear combination of the two loss terms. The objective function is thus defined as

$$\min_{\theta^G} \max_{\theta^D} = (1-\lambda)\mathcal{L}_{\text{MSE}}(\theta^G) + \lambda\mathcal{L}_{\text{adversarial}}(\theta^G, \theta^D)$$

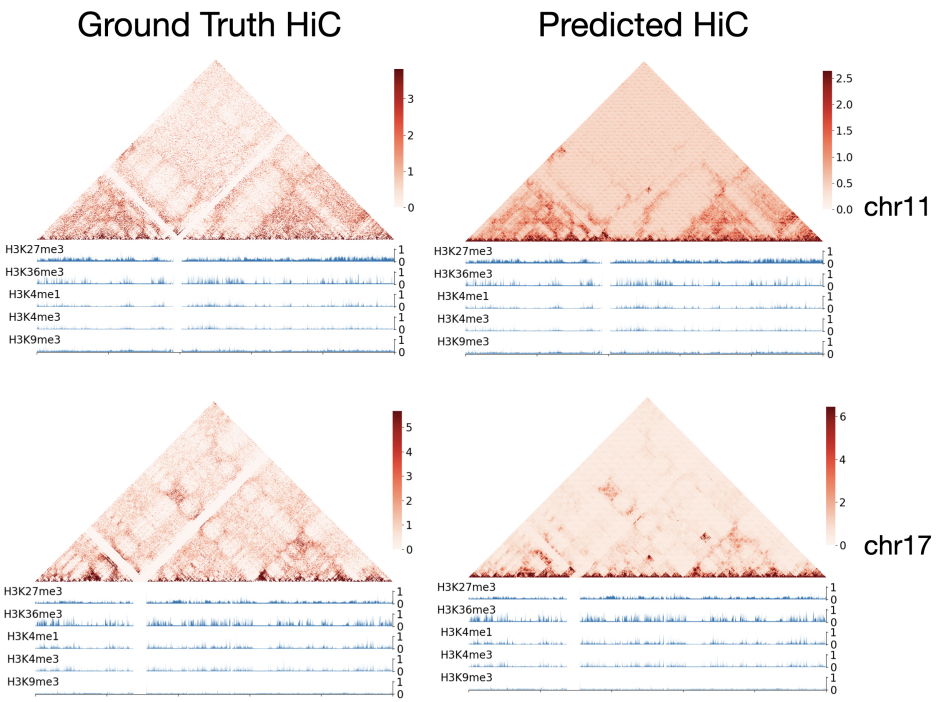

Fig. 1: DeepBio achieves high resolution contact maps within 2 training epochs. Comparison of the predicted, high resolution HiC map and the ground truth high resolution (10kb) HiC map for GM12878 chr11 and chr17. Though the predicted HiC maps have perceptually block and fuzzy regions, structural elements within the contact map are still discernible.

where $\mathcal{L}_{\text{MSE}}(\theta^G)$ represents the MSE loss between the predicted HiC map and ground truth, high resolution HiC map, and $\mathcal{L}_{\text{adversarial}}(\theta^G, \theta^D)$ represents the adversarial loss between the HiC autoencoder and the discriminator networks.

We posit that, akin to Epiphany, the additional GAN Discriminator Loss Term will force the model to learn latent structures inherent to HiC matrices, resulting in higher quality HiC predictions.

Naturally, this model has several hyperparameters to tune. To achieve the results noted in the following sections, we train this model on 2 epochs on a T4 GPU through Google Colab Pro, with batch size 32 and HiC slice size of 256 bins. For each epoch, we iterate through each batch per chromosome. Chromosomes in the training set greater than 5 (*i.e.* chromosomes 6 - on) are trained on 3 times. For each batch, the discriminator is trained for 15 steps, followed by the generator (or autoencoder) being trained for 35 steps: given the complexity of high resolution HiC data, we concluded - with preliminary experimental data as support - that significantly more generator steps (as compared to discriminator steps) were necessary to achieve high quality predictions of high resolution HiC maps. This process took around 7 hours

Results

Despite its innovative approach and the improvements it brought to Hi-C map prediction, DeepBio did not achieve state-of-the-art performance when benchmarked against the latest models in the field. However, the model still demonstrates notable performance after training for a short amount of epochs.

Evaluation of Predicted HiC Maps

We directly predicted and visualized HiC maps for GM12878 to qualitatively evaluate model performance. The visualized results can be observed in Fig. 1. Though we are not sufficiently qualified to properly perform quantitative analysis or downstream interpretation on the HiC maps, we still though a qualitative analysis would be insightful to evaluate the visual quality of DeepBio predictions.

Observing the HiC maps, we can observe a TAD boundary on the left side of both the ground truth and predicted HiC. Moreover, we can generally observe high interaction regions shared between both the ground truth and predicted HiC maps. We thus argue that DeepBio, on minimal training, achieves enhanced HiC maps with biologically interpretable features.

However, the regions of high interaction in the predicted HiC maps are also visibly blocky, where the regions seem to be larger but also more suddenly transition to lower interaction regions as compared to ground truth. We posit that this perceptually-low quality comes as a consequence of our model predicting small slices of high resolution samples.

Ultimately, we conclude that though DeepBio achieves improved resolution HiC contact maps at a visual level, further training is required to properly assess the scale of the architecture's abilities.

	MSE	Pearson	Train Epochs
deepC	–	0.3	14
Akita	0.14	0.61	12
Epiphany	0.8858	0.7028	"Until Convergence"
HiCPlus	0.85	0.976	160,000*
hicGAN	.0078	0.958	Until Convergence
DeepBio	1.201	0.635	2

Table 1. Summary of baseline HiC prediction models and DeepBio performance at validation time, quantified through both Mean Squared Error (MSE), and Pearson Correlation Coeffecient (Pearson), given the number of epochs for which each model was trained. Though our model does not outperform state of the art models, we present near-state of the art performance on just 2 epochs. *It should be noted that HiCPlus only trained on chromosomes 1-8

Comparison with Existing Models

DeepBio's innovative approach in Hi-C map prediction, despite its advancements, did not achieve state-of-the-art performance when benchmarked against the latest models in the field. However, it demonstrated superior performance compared to earlier models such as Akita and DeepC, indicating significant advancements in certain aspects of Hi-C map prediction.

Comparison to Baseline Models

DeepBio's utilization of both epigenomic and low-resolution Hi-C data, combined with adversarial training, allowed for the generation of Hi-C maps with greater detail and accuracy compared to those produced by Akita and DeepC. This advantage was particularly evident in the finer structural details of the maps generated by DeepBio. The inclusion of an adversarial loss term in DeepBio's training regime addressed some of the shortcomings associated with pixel-based metric optimization. This approach led to more realistic representations of underlying Hi-C structures, marking a significant improvement over the results from Akita and DeepC, which did not use such methods. In terms of biological relevance, DeepBio demonstrated a higher consistency with known biological interactions and principles of chromatin organization, suggesting that its predictions might be more biologically relevant and offering insights into chromatin interactions that were less apparent with the earlier models.

However, DeepBio was unable to surpass the performance of the computer vision (CV)-based models HiCPlus and hicGAN. These CV models also seem to surpass the predictive abilities of all other epigenomic and sequence-based models, as can be seen in Table 1. We posit that the vision-based architectures which take advantage of down sampled HiC matrices actually act as very strong predictors which enable better HiC prediction capabilities. Moreover, it should be noted that HiCPlus was trained for significantly many more epochs, while hicGAN was allowed to train until convergence. We believe that DeepBio, as a GAN model, likewise requires more training epochs to overcome the well-documented issue of GAN training instability (Kodali et al., 2018). We strongly believe that training until convergence for the generator of the GAN will enable DeepBio to outperform state of the art models in high resolution HiC prediction.

Despite poorer performance under certain metrics when compared to certain models, it is also critical to take note of computational bottlenecks: while most other models either trained for roughly 10 to 15 epochs (Fudenberg et al. (2020), Schwessinger et al. (2020)) and some even trained beyond 100000 epochs or until convergence (Yang et al. (2021), Zhang et al. (2018)), we were only able to train for 2 epochs. We re-emphasize that further training would be necessary to properly evaluate the quality of high resolution predictions made by DeepBio.

While DeepBio represents a significant step forward from primarily genomic models like Akita and DeepC, particularly in generating more detailed and biologically relevant Hi-C maps, there remains potential for further enhancements to achieve state-of-the-art performance in the field.

Conclusion

In this paper, we introduced DeepBio, a multimodal approach to predicting high-resolution Hi-C from low resolution Hi-C data and gene expression. Our model consists of convolutional layers and a GAN, allowing us to extract fundamental features from low-resolution Hi-C data and generate high-resolution Hi-C data similar to real-life data.

A natural expansion to our model formulation would be to observe the importance of different input features on a well trained discriminator. By analyzing the important input features for the discriminator, we could perform further validation to determine whether the model is being penalized or rewarded for biologically motivated contact map features.

A very significant limiting factor throughout our work was our very limited computational resources. With increased computational resources, our proposed architecture might be able to achieve performance closer to state of the art models by processing entire chromosomal regions at once rather than processing slices of both input modalities. Moreover, using more advanced hyperparameter selection techniques like grid search and learning rate scheduling, instead of approximate experimental data, might further enhance the capabilities of our model. Processing longer epigenomic sequences might be able to better recover long range interactions between genomic regions, while processing more of the HiC data at once might provide a better, rough contextual structure to the latent encoding of the inputs. Likewise, more complex model structures might better represent the HiC structure and lead to more representative predictions. In particular, we would be very interested in exploring Variational Autoencoder (VAE) structures in an attempt to better define the latent space of paired genomic bin interaction levels given epigenomic features.

In summary, DeepBio represents a proof-of-principle for merging multiple modalities to better predict chromosomal structure from a broader set of contextual information. In particular, we demonstrate that models can take advantage of both available low-resolution contact maps and epigenomic features to predict high-resolution contact maps.

Contributions

Bumjin

Problem formulation / alteration over time, developing model architecture, debugging incompatible code from Epiphany/GC Merge/DeepC/CAESAR, developing train and validation step design, hyperparameter tuning, running code, citations, figure generation

Gabrielle

Introduction, abstract, and revisions, data processing, gathering Hi-C data.

Pavani

Epiphany and related works, data processing, gathering Hi-C data.

Evan

Related works and results, gathering data from GC-merge, preprocessing code into per chromosome data, adding metrics + down sampling to model, modifying model architecture.

Serdar

Discriminator model formulation, related works, background information, data processing

Acknowledgments

The authors thank Professor Singh, Atishay Jain, and the anonymous reviewers for their valuable suggestions and guidance.

References

J. Bigness, X. Loinaz, S. Patel, E. Larschan, and R. Singh. Integrating long-range regulatory interactions to predict gene expression using graph convolutional networks. *bioRxiv*, 2021. doi: 10.1101/2020.11.23.394478. URL `https://www.biorxiv.org/content/early/2021/01/28/2020.11.23.394478`.

G. Fudenberg, D. R. Kelley, and K. S. Pollard. Predicting 3d genome folding from dna sequence with akita. *Nature Methods*, 17:1111–1117, 2020. doi: 10.1038/s41592-020-0958-x. URL `https://doi.org/10.1038/s41592-020-0958-x`.

N. Kodali, J. Hays, J. Abernethy, and Z. Kira. On convergence and stability of GANs. 2018. URL `https://openreview.net/forum?id=ryepFJbA-`.

E. Lieberman-Aiden, N. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B.R., Lajoie, P. S. abd M.O. Dorschner, R. Sandstrom, B. Bernstein, M. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. Mirny, E. Lander, and J. Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science (New York, N.Y.)*, 326(5950):289–93, 2009.

Q. Liu, H. Lv, and R. Jiang. hicGAN infers super resolution Hi-C data with generative adversarial networks. *Bioinformatics*, 35(14):i99–i107, 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz317. URL `https://doi.org/10.1093/bioinformatics/btz317`.

S. Rao, M. Huntley, N. Durand, E. Stamenova, I. Bochkov, J. Robinson, A. Sanborn, I. Machol, A. Omer, E. Lander, and E. Aiden. A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 162:687–688, 08 2015. doi: 10.1016/j.cell.2015.07.024.

R. Schwessinger, M. Gosden, D. Downes, R. C. Brown, A. M. Oudelaar, J. Telenius, Y. W. Teh, G. Lunter, and J. R. Hughes. Deepc: predicting 3d genome folding using megabase-scale transfer learning. *Nature Methods*, 17:1118 – 1124, 2020. doi: 10.1038/s41592-020-0960-3. URL `https://doi.org/10.1038/s41592-020-0960-3`.

R. Yang, A. Das, V. R. Gao, A. Karbalayghareh, W. S. Noble, J. A. Bilmes, and C. S. Leslie. Epiphany: predicting hi-c contact maps from 1d epigenomic signals. *bioRxiv*, 2021. doi: 10.1101/2021.12.02.470663. URL `https://www.biorxiv.org/content/early/2021/12/03/2021.12.02.470663`.

Y. Zhang, L. An, J. Xu, B. Zhang, W. J. Zheng, M. Hu, J. Tang, and F. Yue. Enhancing hi-c data resolution with deep convolutional neural network hicplus. *Nature Communications*, 9:750, 02 2018. doi: 10.1038/s41467-018-03113-2.