Final Report: Resolution Enhancement Finetuning on HiC Foundation with Sequence Information

Irene Lau

ABSTRACT

Hi-C data provides insights into the 3D organization of the genome, but generating high-resolution Hi-C maps is costly and data-intensive. The HiCFoundation model is a pretrained vision transformer designed for general-purpose chromatin modeling, including resolution enhancement. In this project, HiCFoundation is extended by incorporating DNA sequence information through a lightweight 1D CNN encoder, creating a dual-encoder model for predicting high-resolution contact maps from low-resolution inputs. Multiple versions of the model were trained—with and without sequence input—using rebinned micro-C datasets and GRCh38 reference sequences. Performance was assessed using Mean Squared Error (MSE) and Pearson Correlation Coefficient (PCC). While all finetuned models outperformed the no-finetune baseline, the addition of sequence information had only marginal improvements. These results suggest that the low-resolution Hi-C matrix is enough for this task given the current model constraints, but more expressive sequence encoders or wider resolution gaps could better leverage sequence data in future work.

Introduction

This project is based on the HiCFoundation model paper[1]. The goal of the project is to improve upon the resolution enhancement task by adding sequence information. Sequence information provides additional context the model can use to predict the interactions in a higher resolution contact matrix from a lower resolution matrix. Accurate higher resolution HiC-matrices are important for understanding and studying genome structure. Producing matrices at a high resolution is expensive, so providing a model which can accurately enhance low resolution input matrices makes research in chromosome structure more accessible.

Methods

Architecture

Overview of changes

The original HiCFoundation model architecture was preserved as much as possible. Sequence information was added by incorporating a sequence encoder, creating a dual encoder model. The embeddings produced by the encoders were concatenated and passed to the decoder. For the versions of the model trained without sequence information, the original model architecture was used.

sequence encoder

The sequence encoder used is a simple and very lightweight 1D CNN with three convolutional layers (with ReLU activation functions), average pooling, and one fully connected layer. This embedded one hot encoded sequence information.

Training

The original finetuning program was used to train the model with a few modifications. With the dual encoder model, the original vision transformer parameters were frozen and the sequence encoder was trained. The decoder was also trained following the original set up. The loss used was a combination of mean squared error (MSE) and the structural similarity index measure (SSIM). The MSE loss measures per pixel accuracy and SSIM measures similarity in terms of structures seen between the prediction and target. This deviates from the original paper which uses only MSE. Data used for training is discussed below. For hyperparameters, the following values were used: 32 epochs, batch size of 4, 2 workers, and a gradient accumulation of 4 iterations. These parameters, particularly the batch size, number of workers, and iterations of gradient accumulation were decided based off how much the GPU memory allows. The model itself is quite large which in combination with the large contact matrices makes it difficult to try other combinations of hyperparameters.

To understand the difference in model performance with and without sequence data, multiple versions of the model were trained (table 1 describes them). To see the resulting training and validation loss, see the plots in Figure 1 below. For the most part the training was very similar between all the different versions of the model trained.

Data

The data used was taken from the Orca paper[2]. They had rebinned micro-C matrices in mcool files and the associated GRCh38/hg38 reference genome which they used when developing their model. Using the processed files, the resolutions which most closely matched the original HiCFoundation resolution enhancement finetuning method. The original resolutions were 10kb and 4kb for the low resolution input matrices and high resolution target matrices respectively. The closest match was 8kb and 4kb. 224x224 windows of the matrices were taken and the respective DNA sequence was extracted from the reference genome. The windows of low and high resolution matrices and sequence data was saved to then use for training.

Evaluation

To evaluate model performance, predicted high-resolution Hi-C contact matrices were compared against ground truth using two metrics: Mean Squared Error (MSE) and Pearson Correlation Coefficient (PCC). MSE measures pixel-wise differences, describing overall reconstruction accuracy, while PCC captures the similarity between predicted and actual contact intensities, incorporating relative signal trends into the evaluation. These metrics were computed for each matrix in the test set, and the final results are reported as the mean and standard deviation across all samples (n = 1673). Refer to table 1 for the evaluation results.

Results

All finetuned models performed better than the baseline. However, the models which incorporated sequence data did not perform significantly better than the models without sequence data. The results are shown in the table below.

Model	Mean MSE ↓	Mean PCC ↑	Notes
No Finetune	0.3478	0.0046	Baseline: no training
NoSeq_4DNFI643OYP9	0.000461	0.9083	Single dataset, Hi-C only
NoSeq_4DNFI9GMP2J8	0.000450	0.9030	Single dataset, Hi-C only
NoSeq_BothFiles	0.000390	0.9038	Both datasets, Hi-C only
Seq_4DNFI643OYP9	0.000419	0.9066	Single dataset, Hi-C + sequence
Seq_4DNFI9GMP2J8	0.000493	0.9023	Single dataset, Hi-C + sequence
Seq_BothFiles	**0.000304**	**0.9037**	**Best MSE, tied PCC with others**

Table 1. Evaluation metrics (Mean MSE and Mean PCC) for different model configurations. Lower MSE and higher PCC are better.

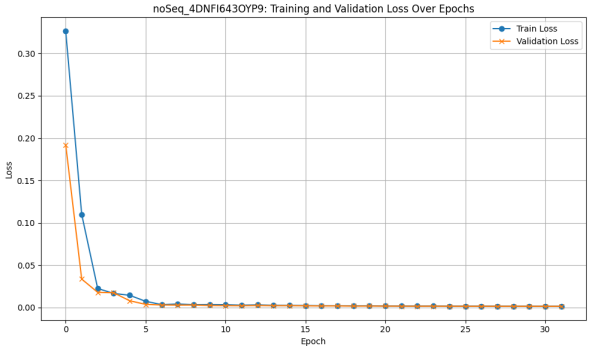
Discussion

The modified HiCFoundation model had marginally better performance than the version without sequence input. This could be attributed to the simplicity of the CNN used to encode the sequence input. If it is too simple, it may not be able to capture the necessary information in order for the model to make fine grained predictions. In the future, it would be good to explore more complex sequence encoders to see the performance difference. Originally, this project used Enformer, but training was not feasible with the memory constraints and the large size of the original HiCFoundation model, Enformer, and the data. Another interesting direction, would be to see how sequence data influences prediction accuracy when the difference in resolution between the target and the input resolution is greater.

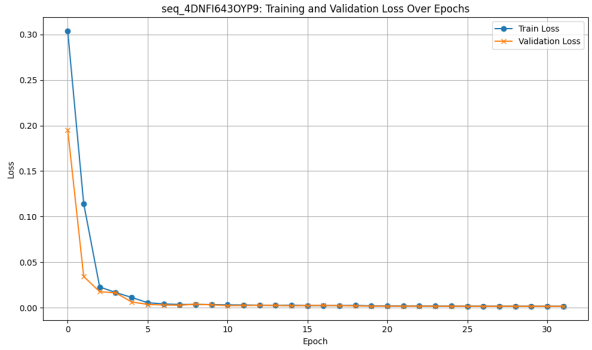
References

1. Wang, X. *et al.* A generalizable hi-c foundation model for chromatin architecture, single-cell and multi-omics analysis across species. *bioRxiv* DOI: 10.1101/2024.12.16.628821 (2024). https://www.biorxiv.org/content/early/2024/12/20/2024.12.16.628821.full.pdf.

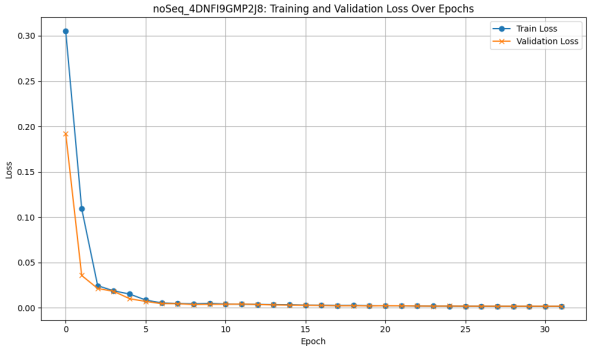
2. Zhou, J. Sequence-based modeling of three-dimensional genome architecture from kilobase to chromosome scale. *Nat. Genet.* **54**, 725–734, DOI: 10.1038/s41588-022-01065-4 (2022).



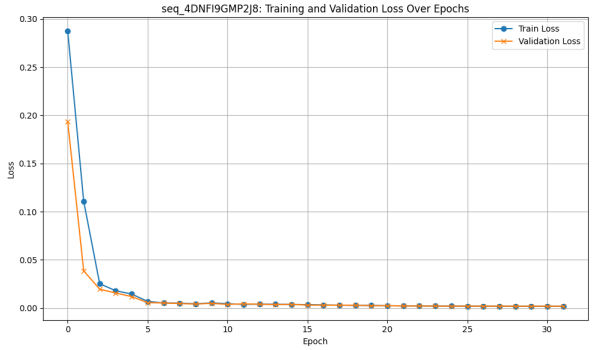
(a) noSeq_4DNFI643OYP9



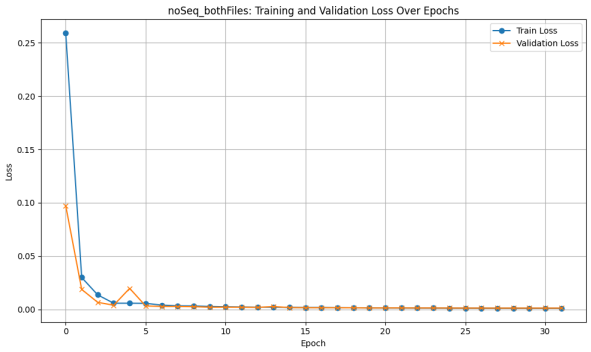
(b) seq_4DNFI643OYP9



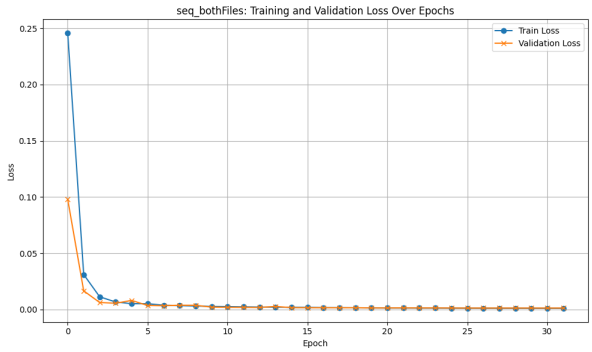
(c) noSeq_4DNFI9GMP2J8



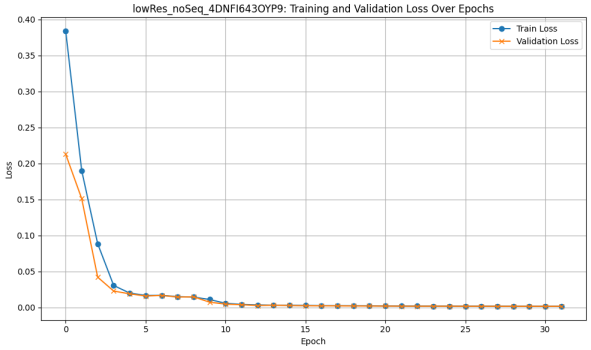
(d) seq_4DNFI9GMP2J8



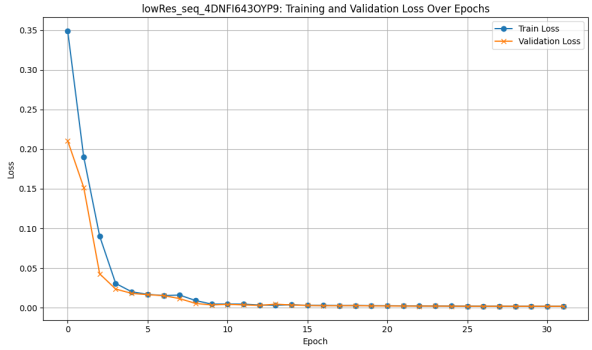
(e) noSeq_bothFiles



(f) seq_bothFiles



(g) lowRes_noSeq_4DNFI643OYP9



(h) lowRes_seq_4DNFI643OYP9

Figure 1. Training and validation loss plots for various model configurations. Towards the end of the project, I got a bit curious and wanted to test training a model to predict from 16kb to 4kb which is what the "lowRes" plots are from. This direction was not fully explored, so the report doesn't discuss this.