Epiphany: Predicting the Hi-C Contact Map from 1D Epigenomic Data

Arnav Das [1*] Rui Yang [2*] Vianne Gao [2] Alireza Karbalaghareh [2] William Noble [1] Jeff A. Bilmes [1] Christina Leslie [2]

[1]University of Washington, Seattle, WA, US [2]Computational & Systems Biology Program, Memorial Sloan Kettering Cancer Center, New York City, NY, US [*]Equal Contribution

Overview

Motivation Develop a model to predict the cell-type specific Hi-C contact map from 1D epigenomic data, and study the relationship between epigenomic signals and 3D genome structure.

Method Epiphany – a neural network to predict the Hi-C contact map from a small number of commonly generated epigenomic tracks.

Application We show that Epiphany generalizes to held-out chromosomes within and across cell types, and that Epiphany reveals potential contributions of epigenomic signals to 3D genome structure.

Materials and Strategy

Materials

- **Hi-C contact map** GM12878 cell line binned into 10Kb resolution, HiC-DC+ normalized observed/expected ratio.
- **Epigenomic tracks** 5 tracks of epigenomic signals: CTCF, H3K27ac, H3K27me3, and H3K4me3 ChIP-seq.

Prediction scheme

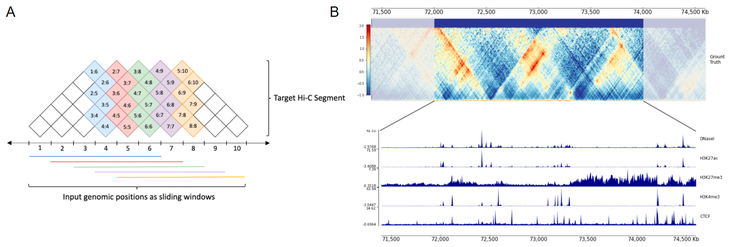

Figure 1. Prediction scheme. **(A)** Prediction scheme of Hi-C contact map and corresponding sliding windows from epigenomic tracks. **(B)** 1Mb × 2Mb target on Hi-C contact map, and its 3.2Mb receptive field from input epigenomic tracks.

Key ideas:

- **Vertical stripe prediction:** predict 1Mb orthogonal-to-diagonal vectors on the contact map (Fig 1A).
- **Receptive field:** for each 1Mb predicted vector, a window size of 1.2Mb centered at the corresponding region is used on epigenomic input tracks.
- **Predict multiple output vectors:** Epiphany captures dependencies between adjacent vectors and predicts a consecutive of output vectors simultaneously.

Epiphany Highlights

Epiphany: a neural network to predict chromatin 3D structure from epigenomic tracks.

- A framework to study the relationship between one or more epigenomic signals and 3D genome structure.
- A model to generate cell-type specific Hi-C contact maps for rare or new cell types.

Key structure:

- Bi-LSTM layer captures the contribution of distal elements.
- Adversarial loss pushes Epiphany to generate realistic Hi-C contact maps.

Model Architecture

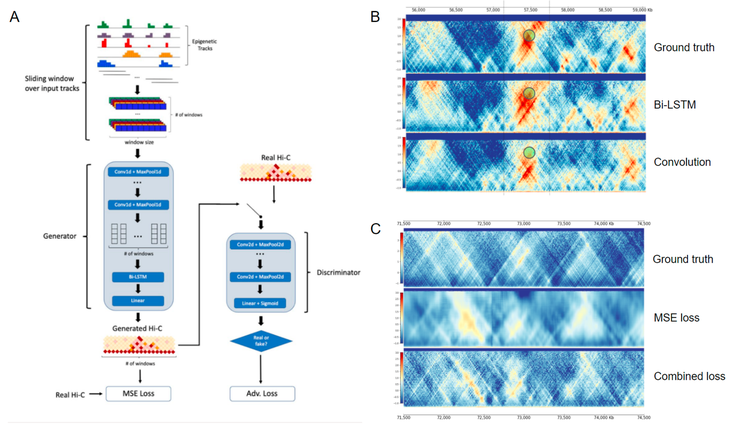

Figure 2. Model architecture and prediction scheme. **(A)** Model Architecture. **(B)** Prediction comparison between MSE only loss and combined loss **(C)** Prediction comparison between introducing bi-LSTM layer and convolution layer.

Key structure:

- **Bi-LSTM layer:** the introduction of a Bi-LSTM layer instead of a convolutional layer captures dependencies between adjacent predictions and increases the receptive field to capture distal elements.
- **Discriminator component:** introducing discriminative loss pushes Epiphany to predict realistic Hi-C contact maps.

Epiphany architecture

- Sliding window: featurize epigenomic signals using sliding window approach (1.2Mb window over 3.2Mb region).
- Generator: 1D convolutional layers to learn local representations from the input tracks as well as bidirectional Long Short Term Memory (Bi-LSTM) layers to capture long term dependencies along the epigenome.
- Discriminator: 2D convolutional layers to distinguish true and predicted Hi-C contact map.

Loss function Convex combination of pixel-wise MSE and adversarial loss to enhance its ability to produce realistic Hi-C contact maps for downstream analysis. Given a dataset D and a trade-off parameter λ, Epiphany solves the following optimization problem during training:

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

$$\mathcal{L}_{adv}(\theta^{\mathcal{G}}, \theta^{\mathcal{D}}) = \mathbb{E}_{(X,Y)\sim D}[\log(\mathcal{D}(Y)) + \log(1-\mathcal{D}(\mathcal{G}(X)))]$$

$$\mathcal{L}_{MSE}(\theta^{\mathcal{G}}) = \mathbb{E}_{(X,Y)\sim D}\left[\sum_{i\in[n]}\sum_{j\in[d']}\left(Y_{ij} - [\mathcal{G}(X)]_{ij}\right)^2\right]$$

In our framework, $\mathcal{G}$ is the CNN-LSTM architecture described in the previous sections while $\mathcal{D}$ is a simple four layer 2D CNN. We used $\lambda = 0.05$ during training.

Model Performance

Benchmark loss functions

Loss Function	Pearson (all)	Pearson (train)	Pearson (test)	Spearman (all)	Spearman (train)	Spearman (test)
MSE only	0.7833	0.8045	0.6494	0.7381	0.7605	0.5963
MSE+GAN	0.7408	0.7687	0.5636	0.6899	0.7191	0.5048

Table 1. Mean Pearson and Spearman correlation for four models

Evaluate model predictions

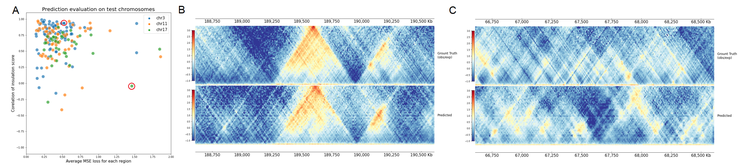

Figure 3. Evaluation of the predicted regions. **(A)** Scatter plot for each predicted region (window with size 1Mb from the diagonal × 2Mb along the diagonal) colored by test chromosomes. X-axis is the average MSE loss between ground truth matrix vs. predicted matrix, and y-axis is the Pearson correlation of insulation scored calculated from ground truth and predicted. **(B)** Comparison between the ground truth (top) and predicted (bottom) for one of the best predicted region (red circle on the top on (A)). **(C)** Comparison for one of the worst predicted region (red circle at the lower right in (A)).

Model Applications

Epiphany predicts cell-type specific contact map

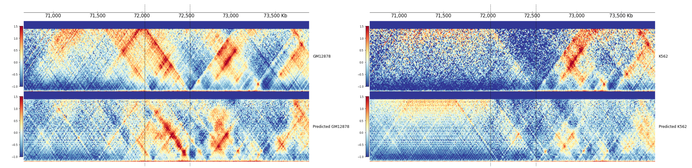

Figure 4. Cell-type specific prediction comparing GM12878 (left) vs. K562 (right) on example region (chr17:70670000-73880000). Ground truth (top) vs. Epiphany prediction (bottom).

Epiphany reveals the importance of epigenomic tracks for predicting 3D genome structure

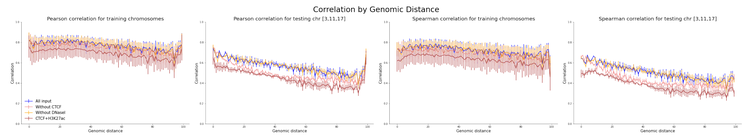

Figure 5. Distance-adjusted Pearson and Spearman correlation of predicted maps using different epigenomic tracks. Model with different epigenomic tracks are trained on all chromosomes except for chr3, 11, 17, and tested on chr3, 11, 17. Model trained with all five input tracks (blue), with CTCF masked (orange), with DNaseI masked (yellow), and using only CTCF+H3K27ac (red).

Summary

Epiphany

Epiphany: a neural network model uses epigenomic trakcs as input, and predict cell-type specific Hi-C contact maps.

Corresponding author: Christina Leslie cleslie@cbio.mskcc.org

International Conference on Machine Learning (ICML) Workshop on Computational Biology