scGraphReg: modeling gene regulations in single cells using multiomics and chromatin interactions

Alireza Karbalayghareh[1], Divya Koyyalagunta[1], Christina S. Leslie[1]
[1]Computational and Systems Biology Program, Memorial Sloan Kettering Cancer Center, New York, NY 10065

Abstract

Transcriptional gene regulation is a complex process that involves the binding of transcription factors (TF) to a gene's enhancers and promoter and the formation of DNA loops between these elements. We have recently developed a method called GraphReg that models these gene regulatory mechanisms and identifies functional enhancers of genes in a bulk population of cells. In this work we introduce scGraphReg which uses single cell multiomics (scATAC and scRNA) as well as any form of chromatin interaction data to predict the gene expression of single cells or pseudo-bulk clusters of cells. We have shown that scGraphReg yields better gene expression prediction performance than its convolutional neural network (CNN) counterpart, which does not use any chromatin interaction data. This confirms the advantage of using chromatin interaction data and distal enhancer information in modeling gene regulation at the single cell level. Our main goal is to use scGraphReg to understand the role of distal enhancers and TFs in gene expression in single cells, which is an ongoing work.

Introduction

Deep learning methods have been widely employed in regulatory genomics to predict different readouts such as gene expression (CAGE-seq), DNA accessibility (DNase-seq and ATAC-seq), histone modifications and TF ChIP-seq data from genomic DNA sequence [1, 2, 3]. All of these works use convolutional neural network (CNN) architectures to predict different readouts from DNA sequences of length less than 150Kb. Epigenomic data can often be predicted quite well with these local models, since local DNA information (receptive field up to 30Kb for Basenji [1]) appears to be sufficient to predict many epigenomic signals based on learning TF binding motifs. However, these methods cannot accurately model gene expression because they do not consider information at distal enhancers, which can be 1Mb or farther from the gene promoter. Predictive models of gene regulation, therefore, require a more global sequence context than can be exploited with current CNN models. A recent method called Enformer [4] has attempted to increase the receptive field up to 100Kb by leveraging transformer architectures, but it still cannot capture more distal enhancers. We have recently developed a method called GraphReg [5] which uses chromatin interaction graphs extracted from HiC/HiChIP data and predicts gene expression values using graph attention networks (GAT). GraphReg uses distal enhancers up to 2Mb away from gene promoters and consequently can be used to find the functional enhancers of the genes.

The advent of single cell multiomics data provides the opportunity to model the relationship between chromatin accessibility and gene expression in single cells. A recent work, scBasset [6], has tried to predict chromatin accessibility of single cells from DNA sequence using CNNs. Here, we introduce scGraphReg which uses single cell multiomics and chromatin interaction data to model the gene expression in single cells or clusters of cells.

Method

We introduce two different versions of scGraphReg: Epi-scGraphReg (Fig. 1a) and Seq-scGraphReg (Fig. 1b), which respectively use chromatin accessibility or DNA sequence in the input layer. We can make predictions at either the single cell or cluster level. Since single-cell ATAC data are very sparse, here we report prediction results at the cluster level to increase the signal-to-noise ratio. At the single cell level, the input for Epi-scGraphReg is the accessibility signal (scATAC) and any 3D chromatin interaction graph for the whole cell population, and the output is its corresponding gene expression (scRNA) signal. The goal of Epi-scGraphReg is to learn enhancers of the genes in individual cells. By contrast, Seq-scGraphReg uses low-dimensional representations of DNA sequence and any 3D chromatin interaction graph to predict gene expression in single cells. The low-dimensional representations of DNA sequence can be learned by using a local CNN model to predict scATAC data from the DNA sequence, an approach similar to scBasset [6]. We show end-to-end models of Seq-scGraphReg in Fig. 1b, but in practice the CNN and GAT branchs can be trained separately. The main goal of Seq-scGraphReg is to find the association of distal TF binding motifs to the expression of genes in single cells. At the cluster level, we cluster the scATAC data and then sum the scATAC signals for all the cells in each cluster to come up with n pseudo-balk ATAC signals. We then sum the scRNA signals inside each cluster to create each cluster's corresponding pseudo-bulk RNA signal.

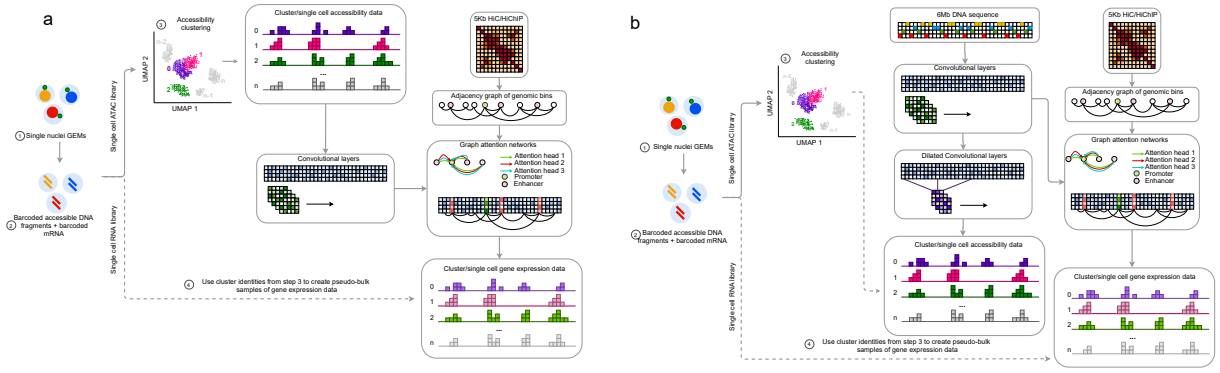

Figure 1: **A schematic overview of scGraphReg models. a.** The Epi-scGraphReg model uses pseudo-bulk cluster or single cell ATAC-seq to learn local features of genomic bins via convolutional neural networks and then propagates these features over adjacency graphs extracted from HiC/HiChIP contact matrices using graph attention networks, in order to predict gene expression (pseudo-bulk cluster or single cell RNA-seq) across genomic bins. **b.** The Seq-scGraphReg model uses DNA sequence as input, and after some convolutional and dilated convolutional layers predicts pseudo-bulk cluster or single cell ATAC-seq. This helps learn useful latent representations of DNA sequences that are then passed to the graph attention networks to be integrated over the adjacency graphs derived from Hi-C/HiChIP contact matrices and to predict gene expression values (pseudo-bulk cluster or single cell RNA-seq).

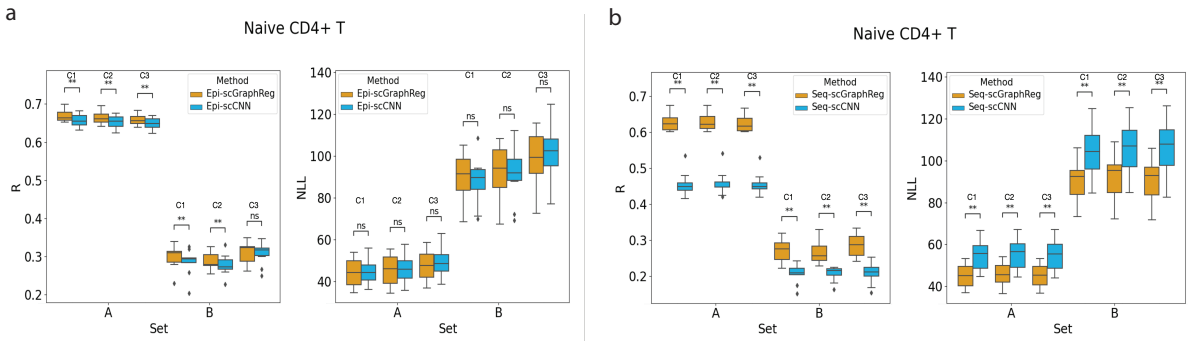

Figure 2: **Gene expression prediction results. a.** Pearson correlation (R) and negative log-likelihood (NLL) of predicted gene expression via Epi-scGraphReg and Epi-scCNN versus true gene expression values in three clusters of naive CD4+ T cells in two sets: A denotes the entire genes and B denotes the expressed genes having more than 5 counts. Improvements are consistent across all three clusters C1, C2, and C3. R for Epi-scGraphReg is significantly higher than Epi-scCNN over 10 runs, but NLL improvement is not significant. **b.** R and NLL of predicted gene expression via Seq-scGraphReg and Seq-scCNN versus true gene expression values in three clusters of naive CD4+ T cells in the two sets A and B. For the sequence models, scGraphReg achieves significantly higher R and lower NLL than scCNN in both sets A and B and all three clusters C1, C2, and C3. The amount of scGraphReg improvement in sequence-based models is higher than that in epigenome-based models.

Results

We obtained human PBMC multiomics data from the 10X website and then clustered immune cells based on scRNA-seq. We annotated the clusters based on enrichment of marker genes. As we have access to the HiChIP data for several classes of CD4 T cells [7], we extracted the CD4 cells from the dataset. As the scATAC-seq is very sparse, we clustered the CD4 cells based on scATAC-seq using ArchR [8] and ended up with three clusters. We generated pseudo-bulk signals by summing up the scATAC-seq and scRNA-seq of all the cells inside each cluster. We assigned the pseudo-bulk gene expression values of each gene to their corresponding promoter bin. The pseudo-bulk ATAC signals are binned at 100bp while pseudo-bulk gene expression signals are binned at 5Kb which is the resolution of HiChIP data. We trained both scGraphReg models 10 times with different sets of train/validation/test chromosomes (except X and Y). In each experiment, we held out two chromosomes for validation, two for test, and trained on the remaining 18 chromosomes. For benchmark models that do not use chromatin interaction graphs, we used CNNs and substituted the GAT layers of scGraphReg with dilated CNN layers, which we denote as Epi-scCNN and Seq-scCNN. Figure 2 shows the performance of both scGraphReg and scCNN models in terms of Pearson correlation (R) and negative log-likelihood (NLL). We see in Fig. 2a that Epi-scGraphReg models have significantly higher R ($p < 0.05$, Wilcoxon ranked paired rank test) than Epi-scCNN models in all three clusters C1, C2, and C3. Sets A and B denote the entire genes and expressed genes having more than 5 counts, respectively. Figure 2b shows the distribution of R and NLL over 10 runs for the Seq-scGraphReg and Seq-scCNN models across the three clusters and the sets A and B. We see in Fig. 2b that Seq-scGraphReg models have significantly higher R and lower NLL ($p < 0.05$, Wilcoxon ranked paired rank test) than Seq-scCNN models in all three clusters C1, C2, and C3 and both sets A and B. We notice that the improvement gap between Seq-scGraphReg and Seq-scCNN is higher than that between Epi-scGraphReg and Epi-scCNN, and this is because of the 1D input to these models; scATAC has only DNA accessibility information, while low-dimensional representation of DNA sequence may contain rich TF binding motif information whose association with gene expression can be learned more effectively using scGraphReg.

References

[1] David R Kelley, Yakir A Reshef, Maxwell Bileschi, David Belanger, Cory Y McLean, and Jasper Snoek. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Research*, 28(5):739–750, 2018.

[2] Žiga Avsec, Melanie Weilert, Avanti Shrikumar, Sabrina Krueger, Amr Alexandari, Khyati Dalal, Robin Fropf, Charles McAnany, Julien Gagneur, Anshul Kundaje, and Julia Zeitlinger. Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nature Genetics*, 2021.

[3] Jian Zhou, Chandra L Theesfeld, Kevin Yao, Kathleen M Chen, Aaron K Wong, and Olga G Troyanskaya. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nature Genetics*, 50(8):1171–1179, 2018.

[4] Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R. Ledsam, Agnieszka Grabska-Barwinska, Kyle R. Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R. Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *bioRxiv*, 2021.

[5] Alireza Karbalayghareh, Merve Sahin, and Christina S. Leslie. Chromatin interaction aware gene regulatory modeling with graph attention networks. *bioRxiv*, 2021.

[6] Han Yuan and David R Kelley. scbasset: Sequence-based modeling of single cell atac-seq using convolutional neural networks. *bioRxiv*, 2021.

[7] Maxwell R Mumbach, Ansuman T Satpathy, Evan A Boyle, Chao Dai, Benjamin G Gowen, Seung Woo Cho, Michelle L Nguyen, Adam J Rubin, Jeffrey M Granja, Katelynn R Kazane, et al. Enhancer connectome in primary human cells identifies target genes of disease-associated dna elements. *Nature genetics*, 49(11):1602–1612, 2017.

[8] Jeffrey M Granja, M Ryan Corces, Sarah E Pierce, S Tansu Bagdatli, Hani Choudhry, Howard Y Chang, and William J Greenleaf. Archr is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nature genetics*, 53(3):403–411, 2021.