EpiAttend: A transformer model of gene regulation combining single cell epigenomes with DNA sequence

Anonymous Author(s)
Affiliation
Address
email

Abstract

Understanding cell type specific gene expression regulation requires models that integrate information across long genomic distances, such as enhancer-gene interactions spanning many tens of kilobases. Neural network models using deep convolutions and self-attention have achieved highly accurate prediction of cell type specific gene expression and other functional genomics measurements based on DNA sequence in local windows[2, 3]. By contrast, leading models for linking enhancers with target genes take advantage of cell type specific epigenomes[4]. Here, we propose a framework for combining DNA sequence with epigenetic data from single cell sequencing within a neural network to predict cell type specific functional readouts such as mRNA expression. This approach has the potential to identify long-range gene-regulatory interactions, linking enhancers with genes based on both the epigenome and DNA sequence binding motifs.

Single cell epigenome and transcriptome sequencing can finely resolve cell types within complex tissues such as the mammalian brain[1]. Using techniques for multimodal data integration, these data can estimate gene expression and epigenetic features such as open chromatin (using single nucleus ATAC-seq) and DNA methylation (single nucleus methylC-seq, snmC-seq) in the same cell types [7]. A recent study by the BRAIN Initiative Cell Census Network combining data from over 500,000 cells from the mouse primary motor cortex identified over 50 fine-grained neuronal and glial cell types, with matching functional data tracks across modalities. These data have the potential to reveal the regulatory networks that control cell type specific gene expression and define the mature identity of brain cell types. For example, correlations between peaks of chromatin accessibility can link enhancers with their target promoters [5, 4]. However, methods for linking *cis*-regulatory elements such as enhancers with their target genes based on epigenetic signals alone are limited, and may lead to substantial false-positives[6].

Here, we propose to combine epigenomic data from single-cell sequencing with rich representations of DNA sequence-based regulatory grammar learned by neural networks [2, 3]. The Enformer architecture leverages multi-headed attention layers to learn long-range interactions between sequence features in order to predict functional output tracks. The model was originally trained using a broad range of sequencing assays applied to bulk tissues and cell cultures[2]. We extend this approach to single cell data by applying it to pseudo-bulk tracks from mouse primary motor cortical neurons and glia[7]. The single-cell data include specialized cell types with distinct mRNA expression and epigenetic identities, which potentially use different DNA sequence motifs and exhibit different modes of regulation compared with cultured cells or bulk tissues. We reasoned that training an Enformer network using these data tracks could help to identify sequence motifs and enhancer locations relevant to brain cell types.

The original Enformer network use self-attention to predict regions of particular importance for predicting functional outputs based on the sequence itself. The single nucleus ATAC-seq and snmC-seq data provide additional, cell type-specific information about the activity of enhancers that could

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

enhance the attention mechanism's ability to focus on relevant regulatory regions. We therefore extended the Enformer framework by adding cell type-specific epigenetic data as additional inputs to the network. The model summarizes the epigenetic data in 128bp bins, which are processed by several layers of convolution before concatenation with the outputs of the Enformer sequence-processing network. The combined representation is then passed through additional layers of convolution and concatenated with sequenced-derived representations. The combined sequence and epigenetic representation is then input to several additional layers of convolution and multi-headed attention.

Our initial studies show that the Enformer architecture is capable of accurately predicting brain cell-type specific genomic signals, including mRNA expression and multiple epigenetic features. By using network interpretation methods, such as gradients or attention-based weights [2], our model will enable a new approach to predicting enhancer-gene links. This framework may help to advance beyond simple motif-based analyses, using long-range and non-linear dependencies between sequence and chromatin features across genomic regions to improve models of gene regulation.

References

[1] E. J. Armand, J. Li, F. Xie, C. Luo, and E. A. Mukamel. Single-Cell sequencing of brain cell transcriptomes and epigenomes. *Neuron*, 109(1):11–26, Jan. 2021.

[2] Ž. Avsec, V. Agarwal, D. Visentin, J. R. Ledsam, A. Grabska-Barwinska, K. R. Taylor, Y. Assael, J. Jumper, P. Kohli, and D. R. Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods*, 18(10):1196–1203, Oct. 2021.

[3] D. R. Kelley. Cross-species regulatory sequence activity prediction. *PLoS Comput. Biol.*, 16(7): e1008050, July 2020.

[4] J. Nasser, D. T. Bergman, C. P. Fulco, P. Guckelberger, B. R. Doughty, T. A. Patwardhan, T. R. Jones, T. H. Nguyen, J. C. Ulirsch, F. Lekschas, K. Mualim, H. M. Natri, E. M. Weeks, G. Munson, M. Kane, H. Y. Kang, A. Cui, J. P. Ray, T. M. Eisenhaure, R. L. Collins, K. Dey, H. Pfister, A. L. Price, C. B. Epstein, A. Kundaje, R. J. Xavier, M. J. Daly, H. Huang, H. K. Finucane, N. Hacohen, E. S. Lander, and J. M. Engreitz. Genome-wide enhancer maps link risk variants to disease genes. *Nature*, 593(7858):238–243, Apr. 2021.

[5] H. A. Pliner, J. S. Packer, J. L. McFaline-Figueroa, D. A. Cusanovich, R. M. Daza, D. Aghamirzaie, S. Srivatsan, X. Qiu, D. Jackson, A. Minkina, A. C. Adey, F. J. Steemers, J. Shendure, and C. Trapnell. Cicero predicts cis-regulatory DNA interactions from Single-Cell chromatin accessibility data. *Mol. Cell*, 71(5):858–871.e8, Sept. 2018.

[6] F. Xie, E. J. Armand, Z. Yao, H. Liu, A. Bartlett, M. Margarita Behrens, Y. E. Li, J. D. Lucero, C. Luo, J. R. Nery, A. Pinto-Duarte, O. Poirion, S. Preissl, A. C. Rivkin, B. Tasic, H. Zeng, B. Ren, J. R. Ecker, and E. A. Mukamel. Robust enhancer-gene regulation identified by single-cell transcriptomes and epigenomes. Oct. 2021.

[7] Z. Yao, H. Liu, F. Xie, S. Fischer, R. S. Adkins, A. I. Aldridge, S. A. Ament, A. Bartlett, M. M. Behrens, K. Van den Berge, D. Bertagnolli, H. R. de Bézieux, T. Biancalani, A. S. Booeshaghi, H. C. Bravo, T. Casper, C. Colantuoni, J. Crabtree, H. Creasy, K. Crichton, M. Crow, N. Dee, E. L. Dougherty, W. I. Doyle, S. Dudoit, R. Fang, V. Felix, O. Fong, M. Giglio, J. Goldy, M. Hawrylycz, B. R. Herb, R. Hertzano, X. Hou, Q. Hu, J. Kancherla, M. Kroll, K. Lathia, Y. E. Li, J. D. Lucero, C. Luo, A. Mahurkar, D. McMillen, N. M. Nadaf, J. R. Nery, T. N. Nguyen, S.-Y. Niu, V. Ntranos, J. Orvis, J. K. Osteen, T. Pham, A. Pinto-Duarte, O. Poirion, S. Preissl, E. Purdom, C. Rimorin, D. Risso, A. C. Rivkin, K. Smith, K. Street, J. Sulc, V. Svensson, M. Tieu, A. Torkelson, H. Tung, E. D. Vaishnav, C. R. Vanderburg, C. van Velthoven, X. Wang, O. R. White, Z. J. Huang, P. V. Kharchenko, L. Pachter, J. Ngai, A. Regev, B. Tasic, J. D. Welch, J. Gillis, E. Z. Macosko, B. Ren, J. R. Ecker, H. Zeng, and E. A. Mukamel. A transcriptomic and epigenomic cell atlas of the mouse primary motor cortex. *Nature*, 598(7879):103–110, Oct. 2021.