The 3D genome and predictive gene regulatory models

Christina Leslie
Computational and Systems Biology Program
Memorial Sloan Kettering Cancer Center
http://cbio.mskcc.org/leslielab
Deciphering gene regulation

- Transcriptional regulation coordinated by transcription factors (TFs) binding to promoter and enhancer elements
- Distal enhancers may be >1Mb from promoters, physically interact via chromatin looping
- 1D epigenomic data (chromatin accessibility, histone marks) map presence of candidate enhancer elements but not their connectivity
Epigenomic data encodes regulatory information

- E.g. chromatin accessibility (ATAC-seq) maps local regulatory elements and encodes global differentiation state

![Graph showing chromatin accessibility and gene expression]

Functional CD8 T cells

Tumor-specific dysfunctional CD8 T cells

low $Pdcd1$, high $Ifng$ expression

gain of $Pdcd1$, loss of $Ifng$ expression

*Philip et al., Nature 2017*
Ascribing function to non-coding genetic variants

- Most GWAS signals reside in non-coding regions, causal variant assumed to be regulatory, i.e. alter regulation of target gene (possibly quite distal)
- Predictive models of gene regulation could infer the role of genomic elements, individual genetic variants on target gene expression

Example from Kumasaka et al., Nat Genet 2019
Predictive gene regulatory models

- Previous GRMs predict gene expression (or fold change) from DNA sequence and accessibility/activity of regulatory elements in order to decipher gene regulation

- Missing information: connectivity of promoter and enhancers
- Idea: use 3D interaction data in graph neural network GRMs
Mapping the 3D genome

• Hi-C, chromosome conformation capture
  – Capture 3D interactions: crosslink DNA (now in situ), restriction enzyme digest, proximity ligation, pull down, paired-end sequencing

  – Read pair = “contact”; build contact matrix for input cell population

Adapted from Arima product sheet
Hierarchical folding of chromatin

- TADs and CTCF/cohesin loops believed to play an “insulator” role in gene regulation
- 3D promoter-enhancer interactions can be more subtle than structural loops

Adapted from Wright et al., 2019
Methods matter: HiC-DC+

- “Hi-C direct caller”: use read counts from raw contact matrix directly, without normalization
  - Estimate background model (expected read count) directly from data using negative binomial regression
  - Covariates: genomic distance (spline fit), mappability, effective bin size (related to restricting enzyme density), GC content
  - Assign $P$ value (or $Z$-score) to interactions

HiC-DC+: Efficient code, extends to HiChIP, differential interactions between cell types

Carty et al., Nat Commun 2017; Sahin et al., in revision, bioRxiv 2020
Methods matter: HiC-DC+

- Gain of promoter-enhancer for diabetes gene *PDX1* in guided pancreatic differentiation
  - With Danwei Huangfu and Eftychia Apostolou (as 4D Nucleome project)

![Diagram of hESC to SC-β cells with HiC-West images](attachment:image.png)
GraphReg: graph neural networks for gene regulatory models

- Idea: use Hi-C/HiChIP to encode long-range chromatin interactions as a graph, propagate information via graph neural networks (GNNs)
- Nodes of graph = genomic bins, edges = 3D genomic interactions
- Input features: epigenomic data or DNA sequence
- Output: gene expression (at node)
Epigenome-based gene regulatory model, Epi-GraphReg

Kharbalayghareh et al., in preparation

3D input: regulatory chromatin interactions (H3K27ac HiChIP)

1D input: chromatin accessibility and histone modifications data

Output: CAGE-seq (gene expression at TSS)

- Predict gene expression from *activity* and *connectivity* of regulatory elements
- “Cell type agnostic”: can generalize to a new cell type given cell-type specific 1D and 3D inputs
Epi-GraphReg architecture

Inputs: DNase-seq, H3K4me3 (promoter mark), H3K27ac (enhancer mark)

GAT learns to weight edges

Output: CAGE-seq

- Train on 6Mb input regions
- Poisson loss on middle 2Mb bins
Prediction of gene expression

• Train on cell line data, assess performance on held-out chromosomes

E-GraphReg, mESC

Epi-GRM, $R = 0.797$

245.0

640.5

mESC expression

E-GraphReg, GM12878/K562

$m<5$: $R = 0.629$, MSE = 3.785

$m>=5$: $R = 0.636$, MSE = 1.909

GM12878 vs. K562

log fold change
Sequence-based gene regulatory model, Seq-GraphReg

Kharbalayghareh et al., in preparation

- Predict expression and 1D epigenomic signals from genomic DNA sequence + 3D connectivity
- “Cell type specific”: captures TF binding signals that are specific to the training cell type
• Sequence-to-1D-epigenome component of the model is similar to Basenji (Kelley et al., 2018)
• Learn DNA sequence features that predict regulatory element activity, combined over HiChIP graph to predict expression
Prediction of gene expression

- Train on ENCODE GM12878 and K562 cell line data, assess performance on held-out chromosomes

**S-GraphReg, mESC**

- n=0: R = 0.784, NLL = 483.2
- n>0: R = 0.757, NLL = 917.5

**S-GraphReg, GM12878/K562**

- m<5: R = 0.361, MSE = 5.812
- m>=5: R = 0.362, MSE = 2.482

mESC expression

GM12878 vs. K562
log fold change
Prediction performance

- Graph NN models outperform baseline sequence models (1D dilated CNNs) in all cases
- Sequence-based prediction is more difficult
- Prediction of expression *per se* is not the point: want to interpret the model

A: all genes, B: all expressed genes, C: expressed genes at least 1 HiChIP edge, D: expressed genes with at least 5 HiChIP edges
Feature attribution to predict functional enhancers

- DeepSHAP identifies features/genomic bins that contribute most to specific gene predictions

DHPS

true signal
predicted signal
feature attribution (DeepSHAP) for the gene DHPS

Epi-GraphReg
Evaluation of enhancer prediction with FlowFISH

- CRISPRi-FlowFISH: CRISPR inactivation screen against candidate enhancers, reads out expression change of target gene
- Activity-by-contact (ABC): score for predicting functional enhancers based on activity (DNase, H3K27ac) and Hi-C contacts

Fulco et al., Nat Genet 2019
GraphReg improves functional enhancer prediction

- Use FlowFISH experiments sufficient data on distal elements (2906 candidate elements for 21 genes)
- GraphReg models with DeepSHAP or saliency outperform CNN models, ABC
GraphReg models access distal information unavailable to CNNs

Epigenome-based models

Sequence-based models

- Dilated CNN has wide receptive field, but feature attribution shows they rely on promoter-proximal inputs
Conclusions

• Graph neural network model can predict gene expression (TSS output) across large genomic regions from 3D and 1D data, or from DNA sequence using 1D epigenomic prediction as auxiliary task
• Epi-GraphReg and Seq-GraphReg outperform baseline dilated 1D CNN models for gene expression prediction
• More importantly, can use feature attribution to predict functional enhancers for genes
• GraphReg outperforms CNN models and ABC score for identifying enhancer elements, as validated by CRISPRi-FlowFISH

Rapid developments in machine learning, epigenomics/3D genomics, and genome editing enable advances in modeling and deciphering gene regulation
Acknowledgements

Leslie lab

Alireza Karbalayghareh
Merve Sahin
Wilfred Wong
Yuri Pritykin (now PI at Princeton)
Erik Ladewig
Rose DiLoreto
Alli Pine
Zakieh Tayyebi
Vianne Gao
Rui Yang
Karen Chu
Sneha Mitra (intern)
Brennan Lee (MS)
Saloni Vishwakarma (MS)

Collaborators (3D projects)

Effie Apostolou
Danwei Huangfu
Aaron Viny
Ari Melnick
Steve Josefowitz
Alexander Rudensky
Andrea Ventura
Ping Chi
Joseph Sun