Identifying composition rules for transcription factor circuits that control macrophage signal response with deep learning

Jenhan Tao
Department of Cellular and Molecular Medicine
University of California
San Diego, CA, USA
jenhantao@gmail.com

Gregory Fonseca
Department of Cellular and Molecular Medicine
University of California
San Diego, CA, USA
gfonseca@ucsd.edu

Christopher K Glass
Department of Cellular and Molecular Medicine
University of California
San Diego, CA, USA
ckg@ucsd.edu

KEYWORDS

genetic circuits; transcription factors; regulation of transcription; cell signaling; machine learning; deep learning

1 INTRODUCTION

Regulation of gene expression in eukaryotic cells is mediated in part by hundreds of sequence specific transcription factors (TFs) that bind to their individual binding motifs at genomic sequences proximal to a gene (promoters) as well as at distal elements (enhancers). Promoters and enhancers can interact with one another by looping together in three dimensional space. The binding of TFs at promoters and enhancers mediates the recruitment of cellular machinery necessary for transcription such as RNA Pol II [8]. Prior studies have suggested two classes of transcription factors: 1) lineage determining transcription factors (LDTFs) and and 2) signal dependent transcription factors (SDTFs). LDTFs play important roles in establishing cell type specific patterns of open chromatin (accessible regions of the genome) [6] whereas SDTFs bind in response to a cellular stimuli, resulting in cell-specific responses to signals [7] (Figure 1). These studies, as well as others, suggest that the context specific gene expression in a cell type is genetically encoded by combinations of TF binding motifs at millions of enhancers scattered throughout the genome [3].

Given the evidence that TFs act collaboratively, it naturally follows that individual TF binding motifs have been observed to be poor predictors of signal dependent activation of an enhancer. Previous studies have also demonstrated that the biological activity of an enhancer depends on the specific composition of the TF motif - arrangement and spacing between TF motifs, as well as the sequence degeneracy of each motif [5]. Given that transcription factors bind in combinations, and evidence that the arrangement of motifs help to determine transcriptional activity, we endeavored to teach an artificial neural network (ANN) to predict signal dependent activation of enhancers by reading arrangements of motifs present at open chromatin regions. We hypothesize that different arrangements of motifs can be used to predict the response to different cellular stimuli.

2 EXPERIMENTAL DESIGN

Using ATAC-seq [4], and ChIP-seq targeting H3K27Ac, an enhancer mark associated with active chromatin, we defined active enhancers in mouse macrophage cells stimulated with an array of cytokines (IFN-g, IL-1b, IL-4, IL-5, IL-6, IL-13, IL-23, LPS, TNF-a, TGF-b). This experimental model provides several key advantages:

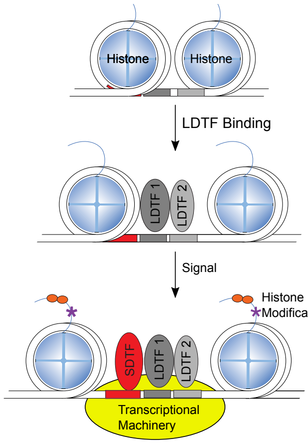

Figure 1: A collaborative hierarchical model for transcription factor binding. Lineage determining transcription factors (LDTFs) bind collaboratively to make cell type regions of chromatin accessible. In response to a signal, a signal dependent transcription factors (SDTFs) bind at sites bound by LDTFs.

(1) The macrophage is a well characterized immune cell with robust responses to signals such as cytokines.
(2) By comparing one signal to another, we can distinguish between SDTFs and general TFs that play a role in many contexts (Figure 2)
(3) Enhancers that respond to multiple signals offer an opportunity to study how elements that encode the response to individual signals can be composed together.

3 MODEL DESIGN

The sequence of each enhancer as well as the enhancersâĂŹ response to each signal, is used as input to train an artificial neural network (ANN) with an attention mechanism to predict signal dependent activation of an enhancer [2]. In contrast to traditional, black-box ANNs, which have been previously applied to predict enhancer activity [1, 9], ANNs with an attention mechanism highlight which regions of the inputs (subsequences of enhancers that presumably are part of TF binding motifs) the ANN is paying attention to as it makes each prediction, thereby divulging the "reasoning" of

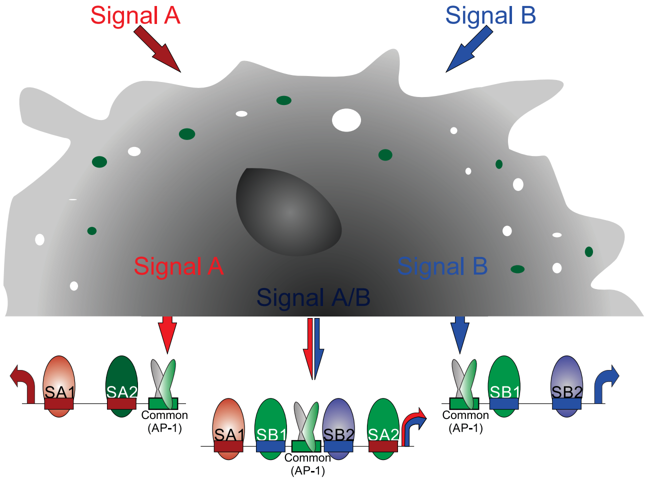

Figure 2: Signal response is encoded by combinations of transcription factor binding sites. Activation of enhancers that respond to signals A and B are mediated by distinct sets of SDTFs ([SA1, SA2] and [SB1,SB2] respectively) as well as general TFs. Enhancers that respond to both signals should contain TF motifs needed to mediate the response to signal A and signal B.

the ANN. The architecture of our neural network is shown in Figure 3. We extract the highlighted binding motifs and then represent the spatial arrangement of TF binding motifs present at each enhancer as a network as shown in Figure 4 - each instance of a motif is represented as a node, and adjacent nonoverlapping motifs are connected with an edge. We can then calculate arrangements of motifs that are enriched at enhancers that respond to a specific cytokine or combination of cytokines. Thus, we can determine a compositions of TF motifs that encodes the transcriptional response to each cytokine, yielding insights into compositional rules for transcription factor circuits.

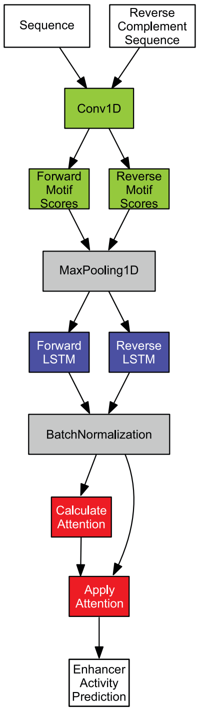

Figure 3: Architecture of neural network

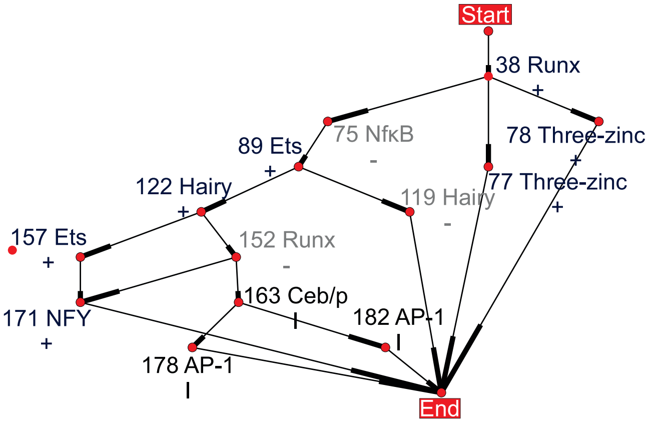

Figure 4: An enhancer represented as a network of TF motifs. Individual motifs are represented as nodes. Adjacent, nonoverlapping motifs are connected with edges. The name of each motif, the position, as well as the orientation is annotated at each node.

REFERENCES

[1] Babak Alipanahi, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. 2015. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol* 33, 8 (2015), 831–838. https://doi.org/10.1038/nbt.3300

[2] Jianpeng Cheng, Li Dong, and Mirella Lapata. 2016. Long Short-Term Memory-Networks for Machine Reading. (jan 2016). arXiv:1601.06733 http://arxiv.org/abs/1601.06733

[3] The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 7414 (sep 2012), 57–74. https://doi.org/10.1038/nature11247

[4] Leighton J Core, Joshua J Waterfall, and John T Lis. 2008. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science (New York, N.Y.)* 322, 5909 (2008), 1845–8. https://doi.org/10.1126/science.1162228 arXiv:NIHMS150003

[5] Emma K. Farley, Katrina M. Olson, Wei Zhang, Daniel S. Rokhsar, and Michael S. Levine. 2016. Syntax compensates for poor binding sites to encode tissue specificity of developmental enhancers. *Proceedings of the National Academy of Sciences* 113, 23 (jun 2016), 6508–6513. https://doi.org/10.1073/pnas.1605085113

[6] Sven Heinz, Christopher Benner, Nathanael Spann, Eric Bertolino, Yin C. Lin, Peter Laslo, Jason X. Cheng, Cornelis Murre, Harinder Singh, and Christopher K. Glass. 2010. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell* 38, 4 (may 2010), 576–589. https://doi.org/10.1016/j.molcel.2010.05.004

[7] S Heinz, C E Romanoski, C Benner, K A Allison, M U Kaikkonen, L D Orozco, and C K Glass. 2013. Effect of natural genetic variation on enhancer selection and function. *Nature* 503, 7477 (nov 2013), 487–92. https://doi.org/10.1038/nature12615 arXiv:NIHMS150003

[8] Sven Heinz, Casey E Romanoski, Christopher Benner, and Christopher K Glass. 2015. The selection and function of cell type-specific enhancers. *Nature reviews. Molecular cell biology* 16, 3 (2015), 144–54. https://doi.org/10.1038/nrm3949

[9] Daniel Quang and Xiaohui Xie. 2016. DanQ: A hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Research* 44, 11 (2016), 1–6. https://doi.org/10.1093/nar/gkw226