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

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

Christopher Benner
Department of Medicine
University of California
San Diego, CA, USA

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 by looping together in three dimensional space. The binding of TFs at promoters and enhancers mediates the recruitment of cellular machinery necessary for transcription. Prior studies have suggested two classes of TFs: 1) lineage determining TFs (LDTFs) and and 2) signal dependent TFs (SDTFs). LDTFs play important roles in establishing cell type specific patterns of open chromatin (accessible regions of the genome) [5] whereas SDTFs bind in response to a cellular stimuli, resulting in cell-specific responses to signals [6] (Figure 1). These studies, and 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 motifs have been observed to be poor predictors of activation of an enhancer. The biological activity of an enhancer may depend on the specific composition of TF motifs - arrangement and spacing between TF motifs, as well as the sequence degeneracy of each motif [4], 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, 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: 1) The macrophage is a well characterized immune cell with robust responses to signals such as cytokine. 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.

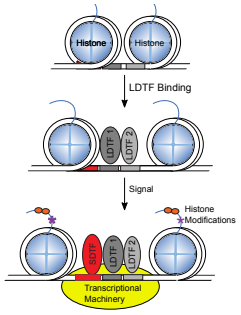

Figure 1: A collaborative hierarchical model for TF binding. Lineage determining TFs (LDTFs) bind collaboratively to make cell type regions of chromatin accessible. In response to a signal, a signal dependent TFs (SDTFs) bind at sites bound by LDTFs.

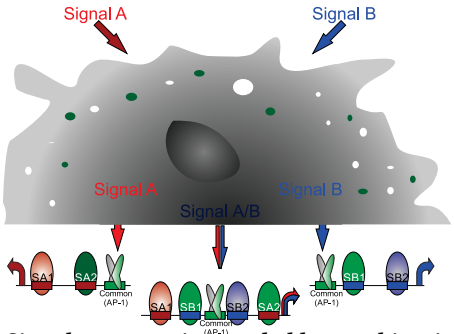

Figure 2: Signal response is encoded by combinations of TF binding sites. Activation of enhancers that respond to signals A and B are mediated by distinct sets of SDTFs ([SA1, SA2] and [SB1,SB2] respectively). Enhancers that respond to both signals should contain TF motifs that mediate both signals.

3 MODEL DESIGN

The sequence of each enhancer as well as the enhancers' response to each signal, is used as input to train an artificial neural network

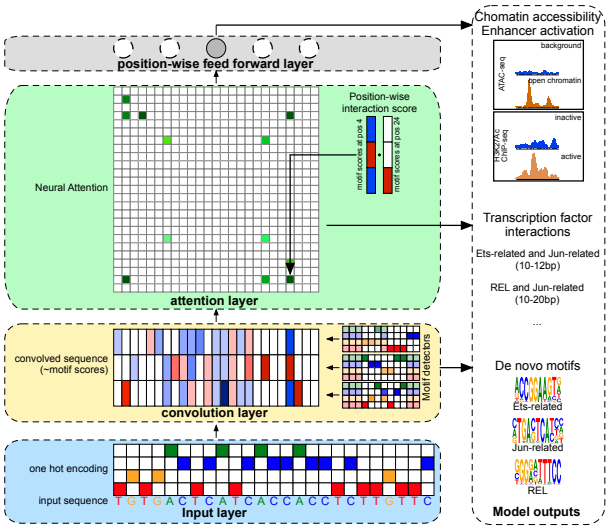

Figure 3: Overview of model

			Model		
			Att	Conv	Large-Conv
		# params	10753	2162	10850
Tx	Veh	Acc.	0.854	0.822	0.846
		Prec.	0.838	0.804	0.830
	KLA-1h	Acc.	0.859	0.807	0.839
		Prec.	0.857	0.791	0.826
	IL4-24h	Acc.	0.862	0.832	0.847
		Prec.	0.858	0.809	0.836

Table 1: Model performance. Mean performance metrics (n=3), accuracy (Acc.) and precision (Prec.), of 3 models: our attentive model (Att.), a convolutional network (Conv), and a large convolutional network (Large-Conv) are shown for macrophages under 3 treatment conditions

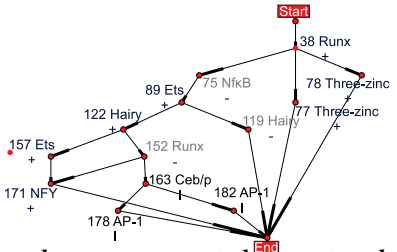

Figure 4: An enhancer represented as a network of TF motifs. Motifs are represented as nodes. Adjacent, motifs are connected with edges. The position are annotated at each node.

(ANN) with an attention mechanism to predict signal dependent activation of an enhancer. In contrast to traditional ANNs that combines the input data in a cryptic fashion (via a fully connected layer) to predict enhancer activity [1, 7], ANNs with an attention mechanism highlight which regions of the inputs (subsequences of enhancers that presumably are TF binding motifs) the ANN is paying attention to as it makes each prediction, thereby divulging the "reasoning" of the ANN. Here we implement a convolutional neural network that uses a dot product attention mechanism [2] to use genomic sequence alone to predict enhancer activity. The architecture of our neural network is shown in Figure 3.

4 PRELIMINARY RESULTS

To assess the performance of our model architecture, we compared the performance of our model against the current state of the art, a convolutional network. We trained our model and an implementation of DeepBind, a previously described convolutional network,[1], to distinguish accessible enhancers from random genomic sequences. The performance of our model exceeded that of the convolutional model, in terms of model accuracy and precision, at detecting enhancers present in macrophages in 3 separate treatment conditions (Table 1 Att versus Conv). Our model's increase in performance versus the convolutional network can be potentially attributed to the greater number of free parameters used (Table 1). And so, we also trained a large convolutional network (with 54 convolution kernels and 108 dense neurons versus 16 convolution kernels and 32 neurons in the original model). The improved performance of our model suggests that the attention mechanism is capable of extracting useful information.

5 FUTURE WORK

While we are encouraged by the performance of our model, we believe the insights we can extract from the network more important. We are currently extracting TF binding sites highlighted by our model and representing each enhancer as a network of TF motifs (Fig. 4). Next, we will calculate arrangements of motifs that are enriched at enhancers that respond to a specific cytokine. Thus, we can determine a compositions of TF motifs that encodes the transcriptional response to each cytokine, yielding insights into compositional rules for signal specific TF circuits.

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] 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

[5] 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

[6] 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

[7] 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