Regulatory Activity Prediction with Attention-based Models

Geoffrey Angus*
Department of Computer Science
Stanford University
Stanford, CA 94305
gangus@stanford.edu

Sabri Eyuboglu*
Department of Computer Science
Stanford University
Stanford, CA 94305
eyuboglu@stanford.edu

Abstract

Phenotypic outcome prediction from genotypes is an important task that allows us to better understand the function of genomic regions. Inspired by the work of Kelley *et al.* on the task of identifying promoters and distal regulatory elements using machine learning, we propose a new neural network model that leverages attention mechanisms in order to predict transcriptional profiles from a given DNA sequence. Attention-based models have proven themselves to be effective in the task of determining long range dependencies in sequence data. We compare several attention-based model architectures with a set of other architectures in order to determine the performance of attention in finding long-range dependencies in DNA sequences with the ultimate goal of successfully predicting genome expression.

1 Introduction

Variations in non-coding regions of the human genome often underpin complex human phenotypes [1]. With genome-wide association studies (GWAS) we can identify the non-coding variations behind phenotypes of interest. However, even with the variations identified, it can be difficult the molecular mechanisms through which these variations affect the phenotype [2]. One common way non-coding variation can affect a phenotype is by altering gene expression. Thus, any insight into a variation's influence on gene expression could begin to explain the molecular mechanisms behind phenotypes linked to that variation. *In-silico* gene expression prediction is one way to gain this insight. For example, given a genome with a variation of interest, researchers can use predictive models to understand how the variation affects gene expression in different cell types.

Computational models that predict gene expression from genome sequence rely on variations in non-coding regions (e.g. promoters, enhancers and distal regulatory elements) to inform their predictions [2]. Until recently, computational models capable of predicting cell-type-specific gene expression have evaded researchers. Recent works, including the FANTOM and ENCODE projects, have produced large datasets of epigenetic and transcriptional experiments on variety of cell-types. These datasets have fueled the development of deep learning models for genome expression prediction.

1.1 Related Work

Distal regulatory elements are non-coding regions that affect the expression of genes thousands of base pairs away. Recent studies have confirmed the influence of distal regulatory elements on the expression of many genes [3], yet in most existing architectures, distal regulatory elements can't influence local predictions [4]. Two recent studies introduce convolutional networks for gene expression designed with distal regulatory elements in mind. Kelley *et al.*'s Basenji architecture

*Equal contribution. Listing order is random.

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

uses dilated convolutions to increase the receptive field of the model. With dilated convolutions, the model's expression prediction for a gene can be informed by base pairs tens of thousands of base pairs up or downstream [2]. Zhou *et al.*'s ExPecto method includes a convolutional neural network (CNN) trained to predict epigenomic profiles from input genome sequences [5]. The model then crafts a feature vector for each gene by computing a weighted sum of epigenomic profiles, where weights are a function of distance from the gene's transcription start site [5].

1.2 Present Work

Here, we present an attention-based neural network architecture designed to capture the influence of distal regulatory elements on gene expression. Like Kelley *et al.* and Zhou *et al.*, our model begins with a series of standard convolutions. Then, one or more multi-head attention layers expand the model's effective receptive field to include the entire input sequence. The model is trained and evaluated on a large dataset of CAGE, DNASE-seq and CHIP-seq assays. Given a genome sequence of input, it predicts experimental counts for a large number of .

Our attention-based model achieves achieves higher correlation with experimental counts ($\rho = 0.587$) than do models with standard convolutions alone ($\rho = 0.506$). It also outperforms ExPecto-like architectures that use constant exponential functions ($\rho = 0.538$). These results suggest that attention could be an effective way to incorporate long-range sequence context into local predictions. Averaged across all experiments, models that include dilated convolutions achieved the highest correlations ($\rho = 0.628$). However, our attention-based model outperformed dilated convolutions on CAGE experiments, suggesting that attention is particularly effective for gene-expression prediction.

2 Neural Network Architectures for Sequential Regulatory Activity Prediction

Our models accept raw genomic sequences 131k bp in length as input. We can represent each sequence as a binary matrix $X \in \{0, 1\}^{4\times131,000}$. They are trained to predict experimental counts in $4,229$ CAGE, DNASE-seq and ChIP-seq assays. Predictions are made for $1,024$ non-overlapping bins of 128 bp each. Let $Y \in \mathbb{Z}_{\geq 0}^{4,229\times1024}$ represent the target experimental counts. Since the targets are counts, we use a log-Poisson loss function

$$L(f|X, Y) = -\sum_{i=1}^{1,024}\sum_{j=1}^{4,229} Y_{i,j} f(X)_{i,j} - e^{f(X)_{i,j}} \tag{1}$$

where $f : \{0, 1\}^{4\times131,000} \rightarrow \mathbb{R}^{4,229\times1024}$ is the model.

All of the architectures that we evaluated begin by applying standard convolutional layers interspersed with max-pooling layers of size 2, 4, 4, and 4. The max-pooling layers reduce the sequence length to 1,024 and create "bins" of 128 bp. The final max-pool layer outputs an activation $H \in \mathbb{R}^{d\times1024}$, where d is the number of filters in the last convolutional layer. This is effectively a d-dimensional embedding for each 128 bp bin.

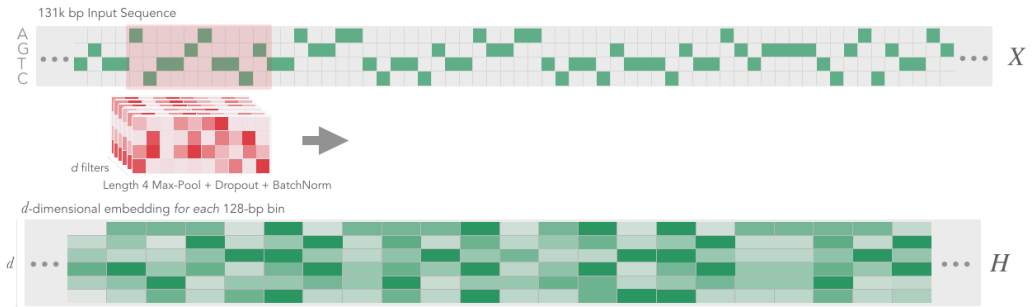

Figure 1: A dilated convolutional filter with $k = 3$ and $l = 2$. The filter is centered at bin j and weights are matched with every l^{th} bin around j.

The size of the receptive field around each bin is linear in the number of standard convolutional layers applied. Thus, each d-dimensional embedding h_i depends only on loci at most a few hundred base-pairs away. We explore three extensions of this basic architecture that increase the receptive field around each bin.

2.1 Dilated Convolutions

After the standard convolutional layers, the Basenji architecture uses dilated convolutional layers to expand the receptive field around the bins [2]. Dilated convolutions are favorable because the size of

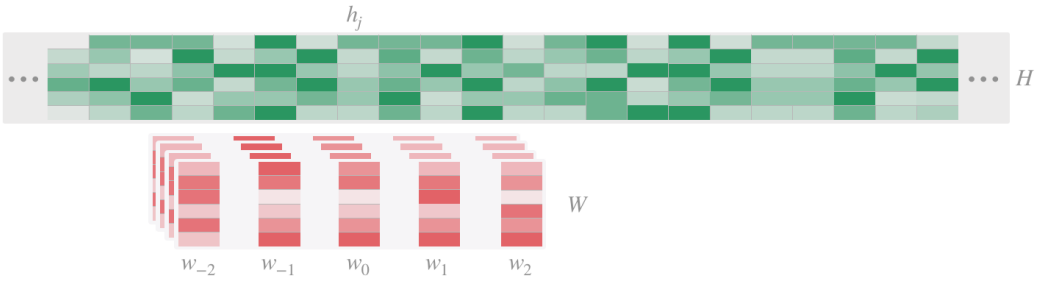

Figure 2: A standard convolutional filter applied to input genomic sequence $X \in \{0,1\}^{4\times 131{,}000}$

.

the receptive field increases exponentially in the number of dilated layers. In a l-dilated convolutional layer, the kernel weights are matched to every l^{th} input [6]. The update rule for h_j can be written as

$$h_j = \sum_{t=-\lfloor k/2 \rfloor}^{\lfloor k/2 \rfloor} h_{j+t*l} \cdot w_t \tag{2}$$

In Figure (1), we give an example of a dilated convolutional filter applied to the activations of the final standard convolutional layers, H.

We use the same architecture as Kelley *et al.* with seven standard convolutional layers followed by seven dilated convolutional layers. Dilation increases by a factor of two with each layer (i.e. the dilation of the i^{th} dilated convolutional layer is $l = 2^{(i+1)}$). The receptive field after the seventh layer is 32k bp, allowing us to capture some but not all distal regulatory interactions [3]. Below we discuss alternative architectures with receptive fields that can cover the entire input sequence of 131k bp.

2.2 Exponential Decay Functions

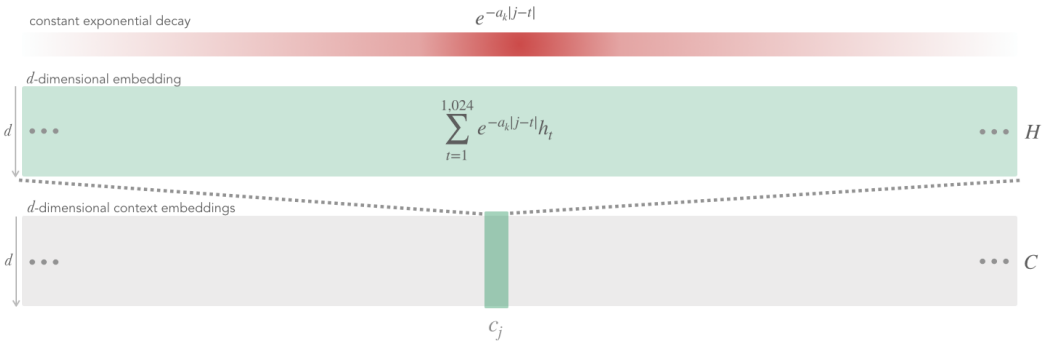

Figure 3: The context vector c_j for bin j is the exponentially weighted sum of all other h_t in the activation H. The decay factor α is a constant. The smaller α is, the more weight we give to distant bins.

Drawing inspiration from ExPecto, we develop a simple approach that uses exponential functions to expand the receptive field around a bin. Given the bin embeddings $H \in \mathbb{R}^{d\times 1024}$, we can develop new

context embeddings $C \in \mathbb{R}^{d \times 1024}$ by summing the embeddings around each bin with an exponentially weighted sum:

$$c_j = \sum_{t=1}^{1,024} e^{-\alpha_k |j-t|} h_t \tag{3}$$

where α_k is the decay constant for the exponential function, and h_j is the embedding for the j^{th} bin. We do this for five different decay factors $\alpha = [0.01, 0.02, 0.05, 0.1, 0.2]$ and concatenate the resulting context vectors together along with the original embedding h_j. These are the same decay factors as those used in ExPecto [5]. In Figure 2, we illustrate how the context vector c_j for bin j is simply the exponentially weighted sum of all other vectors h_j.

2.3 Attention

We introduce a novel architecture where the standard convolutional layers are followed by one or more multi-head attention layers. Like in the exponential decay architecture from section 2.2, our prediction for bin j is informed by all other bins t, not just those within the receptive field of the convolutional layers. However, unlike exponential decay model, ours learns which bins to attend to when making its prediction.

One layer of our attention architecture consists of k parallel attention heads that all receive the same input. The outputs of the heads are concatenated along with the original input and passed to the next layer. The full architecture is outlined in figure 4 below.

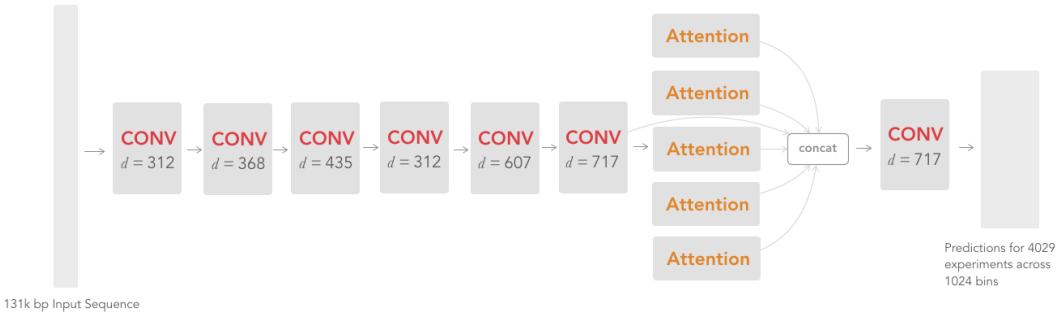

Figure 4: Our attention architecture begins with six standard convolutional layers followed by one or more multi-head attention layers. Each attention layer consists of k parallel attention heads. We follow the attention layers with one last convolutional layer and an affine layer, which makes predictions on all 4,229 experiments.

2.3.1 Attention Heads

For each bin j, we compute a context vector c_j, which is a weighted sum of the embeddings h_t of all other bins t. The weights of this sum represent the attention paid to the other bins. Formally, we define the context vector c_j as

$$c_j = \sum_{t=1}^{1,024} \text{softmax}_t(s(h_t, h_j)) h_t \tag{4}$$

where $s : \mathbb{R}^{1,024}, \mathbb{R}^{1,024} \to [0, 1]$ is an attention scoring function that outputs a scalar weight between 0 and 1 for each bin t. There are many variations of scoring functions that could be used. We focus on one that takes some inspiration from ExPecto's exponential decay function. In ours, we first pass h_j through an affine layer with equal input and output features and a $\tanh$ activation. This transforms h_j into a query vector $q_j \in \mathbb{R}^d$. To determine the relevancy of bin t, we dot q_j with h_t and pass it through a softmax function over all t's. Finally, we scale the score with a function of bin positions i, t. The full function s is given by

$$s(h_j, h_t) = f(j, t) \tanh(h_j^T W + b) \cdot h_t) \tag{5}$$

where $W \in \mathbb{R}^{d \times d}$ and $b \in \mathbb{R}$ are learned weights and $f(j,t)$ is a learned function of bin positions i, t. This scoring function assigns high attention scores to any bin t where h_t is similar to q_j. We explored two options for the function of positions f.

The first, which takes inspiration from ExPecto, uses an exponential function centered at j.

$$f(j,t) = e^{-\alpha|j-t|} \tag{6}$$

where $\alpha \in \mathbb{R}_{\geq 0}$ is the learned exponential decay. The smaller α is, the more weight is given to bins far from j. However, the function always assigns those near j the most weight, regardless of the value of α. In an effort to avoid this constraint, we propose a different choice of f.

We introduce another function f which allows each head to focus on a specific region up or downstream of j. We refer to this position function as the two-sided learned exponential.

$$f(j,t) = \min(e^{\alpha_r(t-j-512)}, e^{\alpha_l(j-t-512)}) \tag{7}$$

where α_r is the learned decay factor for the right exponential function and α_l is the learned decay factor for the left exponential function. In Fig. 8, we plot the two-sided learned exponential for each of our five attention heads, illustrating the regions the model learns to attend to.

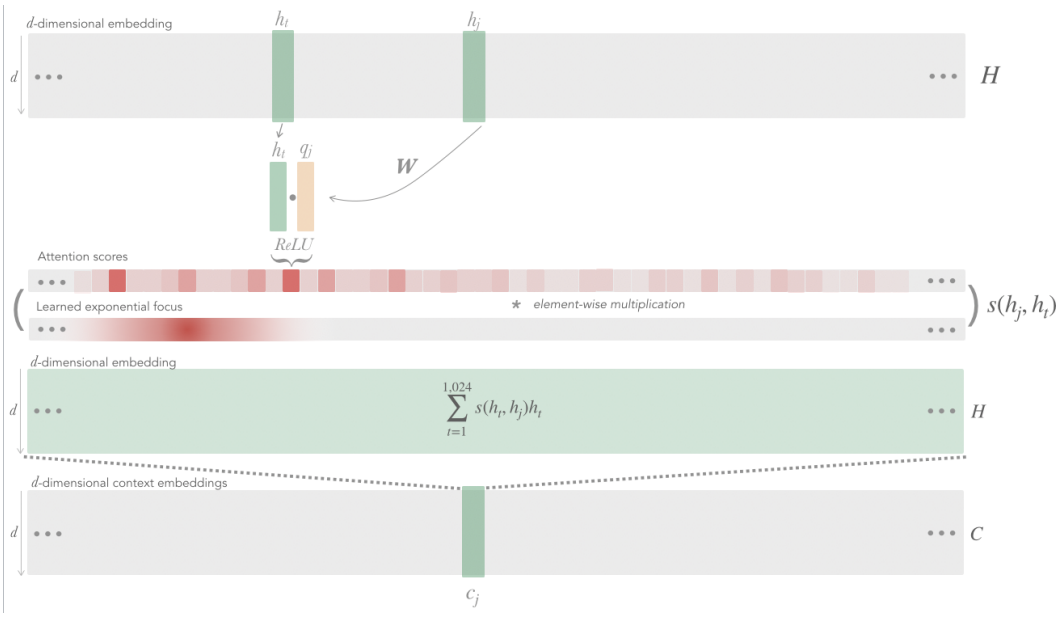

Figure 5: A single attention head with a two-sided learned exponential function. In this toy example, the model learns to attend to regions upstream of the current bin j.

3 Results

With attention-based architectures, we are able to more accurately predict transcriptional profiles than with standard convolutions alone. When averaging over all experiments, attention-based models achieve a Pearson correlation of $\rho = 0.589$, greater than $\rho = 0.506$ with only standard convolutions. Our attention-based models also outperformed ExPecto-like models ($\rho = 0.536$). Models with dilated convolutions were the best performers when averaging over all experiments ($\rho = 0.628$). In figure 7, we show the average performance across all experiments for a subset of our models.

The poor performance of standard convolutions alone could be explained by their inability to learn long distance dependencies outside of its small receptive field. Almost all architectures designed to expand the receptive field achieve a boost in performance. This supports the hypothesis that expanding the receptive field allows the model to capture the influence of distal regulatory elements when making local predictions.

The attention-based models with the highest average correlation employ the two-sided exponential function described in section 2.3.1. Perhaps these models perform the best because they are able to focus on specific regions up and downstream of the current bin. We can observe this effect in figure 8, which shows the two-sided exponential functions learned by our model.

Model	Dilated Convs.	Attn. Layers	Attn. Heads	Exp. Fn.	ρ	r^2
BASENJI	7	0	0	NONE	0.628	0.412
ATTENTION	0	1	5	2-SIDED	0.589	0.356
ATTENTION	0	2	$5, 5$	VARIABLE	0.571	0.325
ATTENTION	0	1	5	VARIABLE	0.566	0.32
ATTENTION	0	1	16	VARIABLE	0.554	0.297
EXPCONSTANT	0	0	0	CONSTANT	0.536	0.235
EXPVARIABLE	0	0	0	VARIABLE	0.528	0.265
STANDARD	0	0	0	NONE	0.506	0.277
ATTENTION	0	1	5	NONE	0.503	0.273

Figure 6: The performance of six models on the validation set. Attention-based models outperform the model reliant on constant exponential decay functions and the standard model.

To understand how models perform on different assays, we split the experiments into three broad categories: CAGE, DNASE and ChIP-seq. For each category, the Pearson correlation was averaged across all experiments in that category. In figure 6, we plot the Pearson correlation for five representative models: dilated convolutional layers, multi-head attention with two-sided exponential functions, multi-head attention with standard exponential functions, standard convolutions with constant exponential functions and standard convolutions alone.

Our multi-head attention model outperforms Kelley *et al.*'s dilated convolutions on CAGE datasets. However, it performs poorly on ChIP-seq and DNASE experiments.

4 Methods

4.1 Dataset and Pre-processing

We use the same datasets and preprocessing techniques that Kelley *et al.* presented [**?**]. The dataset includes a total of 973 CAGE experiments, 949 DNase-seq experiments and 2,307 histone modification ChIP-seq experiments spanning a diverse set of cell-types (CAGE experiments measure gene expression levels in a given cell-type or tissue, ChIP-seq assay identify protein binding sites on DNA sequences and DNase-seq experiments locate regulatory regions in a sequence). These experiments were carried out by the FANTOM5, ENCODE and Roadmap Epigenomics projects [7, 8, 9]. The output of these experiments is a set of reads (short DNA sequences) mapped to experimental counts.

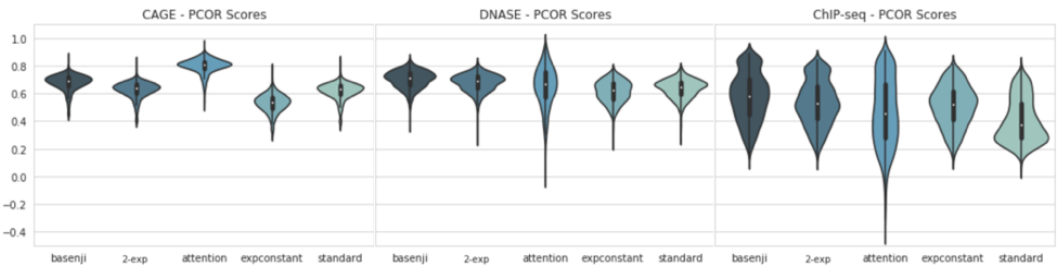

Figure 7: The performance of each model on three different experiment classes—CAGE, DNASE, and ChIP-seq. The manuscript model (*basenji*), attention model with two-sided exponential functions (*2-exp*), the two layer attention model (*attention*), the ExPecto-like model (*expconstant*), and the no-convolution model (*standard*) are shown.

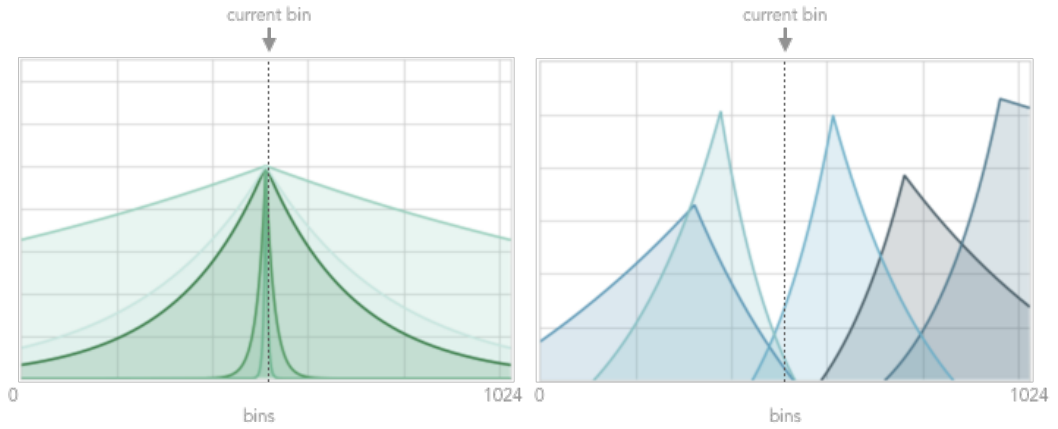

Figure 8: Plots from the parameters of two separate experiments exhibiting the shape of the standard exponential decay function (left) and the two-sided exponential function (right). Note that the two-sided exponential functions need not be centered. This allows the model to specify regions of interest up and down stream of the current bin.

The first step in preprocessing is to align these short DNA sequences to a reference genome. The reads were aligned with Bowtie2 [10]. The counts for multi-mapped reads were distributed and smoothed using an EM algorithm [11]. Finally, GC-content bias in read counts is accounted for using a normalization technique. For more details on preprocessing, we refer the reader to Kelley *et al.*

To generate training, validation and testing examples we extract 131 kbp, non-overlapping sequences from the reference genome. These raw, nucleotide sequences serve as input to the model. Each 131 kbp sequence is split into 1,024, 128 bp bins. We sum the experimental counts of the nucleotides in each bin to produce count estimates for that bin. We extracted a total 14,533 sequences, and used the same train, validation and test splits as Kelley *et al.*.

$$\text{Input sequence:} \quad X \in \{0,1\}^{4x131,072}$$

$$\text{Output counts:} \quad Y \in \mathbb{Z}_{\geq 0}^{4,229x1024}$$

4.2 Training

For all of our models, we initialized the weights of the standard convolutional layers with those learned by Kelley *et al.*'s Basenji architecture. We then fine-tuned those weights and trained the attention layers via stochastic gradient descent with learning rates determined by the Adam optimizer. In order to reduce overfitting, we augmented the dataset, replicating the approach taken by Kelley *et al.*, which focus primarily on two strategies— reverse complementing the DNA sequences every other epoch and shifting the sequences between 0 and 3 nucleotides either to the left or the right. We took an empirical approach to hyperparameter tuning for the attention-based architecture, specifically experimenting with the number of attention layers, the decay factor, the strength of the L2 regularizer and the dropout probability rate.

4.3 Evaluation

We evaluate the performance of the three model architectures of interest with two primary metrics. The first is the coefficient of determination, often denoted as r^2, which can be expressed as the result of the formula

$$r^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2},$$

where y is the vector of true values, $\bar{y}$ is the mean of y, and $\hat{y}$ is the prediction of the model itself. In the context of this task, these values can be interpreted as the counts associated with each of the

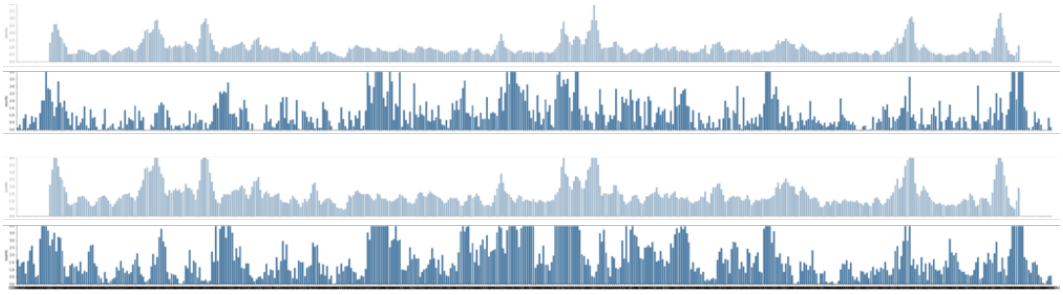

Figure 9: A comparison of the output of the proposed Attention model to the target values on a DNASE experiment run on different donors on a region of the CHR2 chromosome. We see that many of the peaks in the predicted sequence (dark blue) align with those of peaks in the experimental reads (light blue).

experiments, thus intuitively we expect a high r^2-score from a model that minimizes the difference between its prediction and the actual counts for each experiment.

Second, we use Pearson correlation ρ in order to determine the linear correlation between the model's predictions and the true counts found within each experiment.

5 Conclusion and Future Work

We've shown that augmenting standard convolutional networks with attention-based mechanisms that expand the receptive field improves performance in transcriptional profile prediction. In doing so, we've provided more evidence that to predict accurate transcriptional profiles from sequence alone, models need to distant sequences into local predictions.

Furthermore, we've shown that attention-based mechanisms can achieve high correlation when predicting transcriptional profiles. We find that they are particularly powerful for predicting the output of CAGE experiments, even outperforming Kelley *et al.*'s model with dilated convolutions. The strong performance on CAGE datasets could mean that our models are indeed attending to distal regulatory elements, and in turn, making more accurate gene expression predictions. At the very least, this suggests that applying attention-mechanisms to gene expression prediction is a promising avenue for continued work.

6 Peer Review Revisions

We addressed all of the feedback given in our set of peer reviews. We adjusted our experiment methodology such that all models were pre-trained using weights from the manuscript model, in order to decrease training time. With this in place, we were able to ensure that each of the models could have the same number of filters in the standard convolutional layers. With this modification we were also able to train every one of our models to convergence.

We expanded our results section with analyses regarding the ExPecto and Standard model results, specifically comparing their performance to that of the other models. We ran multiple new experiments in order to isolate and compare the effects of the exponential functions and the attention mechanisms independently. We experimented with learned decay constants in order to move away from the implementation found in the ExPecto paper. We expanded our explanation of dilated convolutions in our model architecture section and created a more detailed architecture section and discussed the potential use case of attention mechanisms in our task in the abstract.

Since the poster session as well, we implemented several new aspects to our project, including a model using stacked attention and the model constructed with a two-sided exponential function, as well as more rigorous analysis of the model performance on unseen data across the three different datasets. Thank you to the groups that provided feedback to our team and helping us craft a stronger paper.

References

[1] Evan A. Boyle, Yang I. Li, and Jonathan K. Pritchard. An expanded view of complex traits: from polygenic to omnigenic. *Cell*, 169(7):1177–1186, June 2017.

[2] 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, March 2018.

[3] Borbala Mifsud, Filipe Tavares-Cadete, Alice N. Young, Robert Sugar, Stefan Schoenfelder, Lauren Ferreira, Steven W. Wingett, Simon Andrews, William Grey, Philip A. Ewels, Bram Herman, Scott Happe, Andy Higgs, Emily LeProust, George A. Follows, Peter Fraser, Nicholas M. Luscombe, and Cameron S. Osborne. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nature Genetics*, 47(6):598–606, June 2015.

[4] David R Kelley, Jasper Snoek, and John L Rinn. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome research*, 26(7):990–9, July 2016.

[5] 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, August 2018.

[6] Fisher Yu and Vladlen Koltun. Multi-Scale Context Aggregation by Dilated Convolutions. *arXiv:1511.07122 [cs]*, November 2015. arXiv: 1511.07122.

[7] The ENCODE Project ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science (New York, N.Y.)*, 306(5696):636–40, October 2004.

[8] The FANTOM Consortium and the Riken Pmi. A promoter-level mammalian expression atlas. *Nature*, 507, 2014.

[9] Bradley E. Bernstein, John A. Stamatoyannopoulos, Joseph F. Costello, Bing Ren, Aleksandar Milosavljevic, Alexander Meissner, Manolis Kellis, Marco A. Marra, Arthur L. Beaudet, Joseph R. Ecker, Peggy J. Farnham, Martin Hirst, Eric S. Lander, Tarjei S. Mikkelsen, and James A. Thomson. The NIH roadmap epigenomics mapping consortium. *Nature Biotechnology*, 2010.

[10] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with Bowtie 2. 2012.

[11] Qi Zhang and S Und Uz Keles¸1keles¸1. Genome analysis CNV-guided multi-read allocation for ChIP-seq. 30(20):2860–2867, 2014.