Transcription factor binding prediction using constrained convolutional neural networks

Yakir Reshef **Fulton Wang**

Abstract

One of the promises of genomic medicine is the ability to interpret genetic variants observed in patients in order to inform diagnosis and treatment. A key part of this task is prediction of the effect of a variant on biological processes in the cell such as the binding to DNA of important regulatory proteins called *transcription factors*. Considerable progress has been made in recent years on this task, especially using convolutional neural networks; however, it is not clear whether existing methods, which are trained on sequences that do not contain genetic variation, are robust enough to be able to predict the effect on binding of variants that have never been seen before. A hallmark of interpretable models is that they are often constrained in some way that gives meaning to their parameters. Here we investigate the effect of constraining such predictors in both biologically informed and biologically agnostic ways as a first step toward understanding how to improve their interpretability.

1 Introduction

Advances in genomics and medicine have raised the possibility of leveraging a predictive understanding of the consequences of genetic variation as input for the diagnosis and treatment of patients. Specifically, one goal of precision medicine is to predict the effect of a genetic variant found in a patient on the phenotype[1] of that patient.

This remains very challenging in general, but there is one set of important phenotypes on which this problem is tractable: the binding to DNA of regulatory proteins called *transcription factors* (TFs). A TF is a protein that binds to stretches of DNA in a sequence-dependent manner and then turns nearby genes on/off. A given TF typically binds to DNA in many places in a coordinated fashion to achieve some goal such as cell division, response to an immune stimulus etc., and so TFs play an important role in many diseases [1]. Because TF activity is both disease-relevant and clearly tied to DNA sequence, there is considerable interest in understanding the genetic determinants of TF behavior.

Deep learning has led to advances in our ability to predict TF binding. Specifically, many convolutional neural network architectures have recently been proposed for predicting, given a short (say, 1000 base-pair) DNA sequence, whether a given TF will bind to that sequence or not [2–4]. Such predictors suggest the following approach to assessing the impact of a genetic variant in a patient: suppose the patient has, say, an A instead of a T some specific location and we wish to understand whether that variant affects the binding of a disease-relevant TF. We can query a predictor about the binding of the TF of interest by feeding in two different sequence inputs, one with an A at the location of the variant and one with a T at the location of the variant, and checking which gets a higher prediction and by how much.

However, the black-box nature of existing predictors means that the above method comes with many caveats. Most importantly, the "alternate" version of the sequence might not be represented in the

[1] Recall that a phenotype is the set of observable characteristics of an individual resulting from the interaction of their genetics with the environment.

distribution on which the predictor was trained, such that we don't have guarantees about the generalization capability of the predictor. This may lead to, for example, alternate sequences receiving systematically lower scores than the corresponding "reference" sequences, a behavior that has been observed empirically [5].

In this work, we propose leveraging a more mechanistic understanding of TF binding to "constrain" the predictor, with the ultimate goal of improving the interpretability of its predictions on sequences that are not well represented in the training set. Specifically, we use the fact that there are databases that contain short probabilistic DNA sequences (i.e., each position is a distribution over $\{A, C, G, T\}$) called *motifs* that have been observed experimentally to be bound by different transcription factors. Rather than training a generic convolutional neural network to predict binding, we train a neural network whose first layer of convolutions is fixed to match these motifs. This way, when we look at the difference in the predictions for two possible values of a genetic variant, we know that the variant must act by changing the strength of one of these biologically relevant motifs.

Building a convolutional neural network with meaningful convolutions has several advantages besides this parsimony and the generalization ability that it potentially implies. The first is philosophical: one interpretation of a convolutional neural network is that the first layer of convolutional filters learns "primitives" that are then combined in later layers to produce the output. As the existence of motifs suggests, many primitives governing TF binding are understood but how these primitives are combined to determine whether a TF binds or not remains mysterious. It therefore makes biological sense to spend more model capacity learning one of these than the other. Another advantage is interpretability of the model weights themselves rather than its predictions. In particular, since constrained convolutional filters map onto specific biological entities, there is potential to learn biology by seeing which first-layer convolutional filters are important for which TF.

In general we might expect constraining a predictor to yield better interpretability but worse accuracy. Surprisingly, however, we show on real data that in the case of TF binding our constrained predictor achieves comparable accuracy to its unconstrained counterpart, such that the interpretability we gain with our simpler model comes seemingly "for free". This perhaps suggests that biologically informed constraints are a good idea. However, our results also show that a predictor with fixed convolutions that are chosen randomly but required to be "simple" (in a way that we formalize later) has higher accuracy than both of the above models. This fact is indicative of an intriguing structure to this problem that merits further investigation, and suggests that our understanding of the determinants of TF binding may be enhanced by more systematically comparing the performance of different types of constrained neural networks.

2 The task

To create a "test-bed" for studying the effect of constrained convolutions, we chose data from 24 real TF binding experiments, each corresponding to a particular TF profiled in a particular cell type (i.e., blood, brain, etc.). Our experiments each contain approximately 20,000 different locations where the TF in question is bound to the genome in the cell type in question. We represent each of these experiments $E^{(i)}$ as a set of sequences $\{z_j^{(i)}\}$ where the $z_j^{(i)} \in \{A, C, G, T\}^{1000}$ are DNA sequences centered around the binding locations in experiment i. The k-th task is then to predict, given a sequence from $\cup_i E^{(i)}$, whether or not it belongs to $E^{(k)}$. The 24 experiments we chose together represent 14 unique TFs with 14 known associated motifs, profiled in 2 different cell types. We had a total of $534, 275$ labeled examples, of which we held out $54, 031$ (chromosomes 10 and 11) as a validation set and $45, 458$ (chromosomes 8 and 9) as a test set.

3 The models

The architectures of state-of-the-art TF binding prediction models can be quite complex. Because our aim here is to elucidate the effect of constraining a convolutional neural network in isolation, we fix a relatively simple architecture loosely inspired by [2] as a baseline and then change it by constraining its convolutions in two different ways.

The unconstrained model with which we begin, which we call M_u is shown in Figure 1. It consists of 3 convolutional layers and two layers of fully connected hidden nodes. The convolutional layers

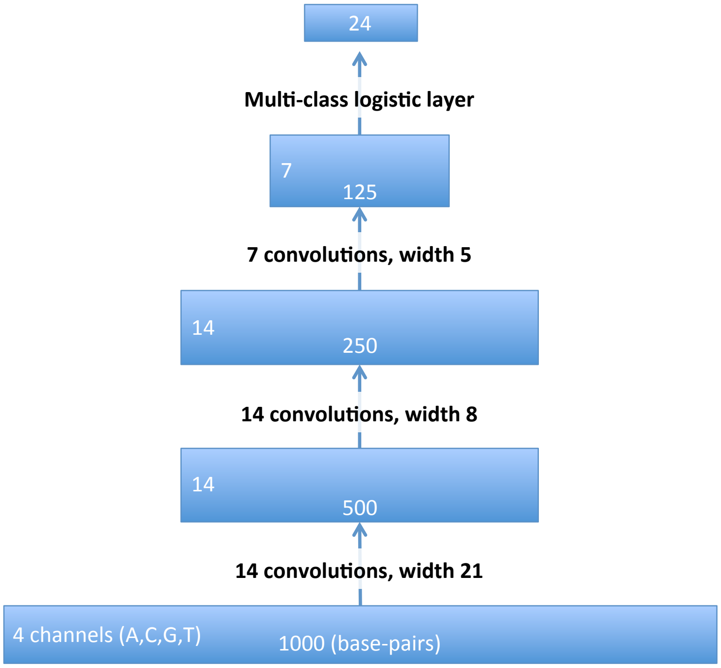

Figure 1: The architecture of the neural network we study. All convolutional layers are followed by rectified linear units and max-pool layers with width 2 and stride 2.

contain 14, 14, and 7 different convolutions of width 21, 8, and 5 respectively. Each convolutional layer is followed by a ReLU and a max-pool with width 2 and stride 2. This model has $24,269$ free parameters.

The first constrained model, which we call M_c, is identical to the unconstrained model except in that the 14 convolutions in the first layer are fixed to have filters corresponding to 14 known motifs for these transcription factors downloaded from an external database called the *Homo Sapiens* Comprehensive Motif Collection (HOCOMOCO) [6]. This model has $23,079$ free parameters.

The second constrained model, which we call M_d, likewise has the 14 convolutions in the first layer fixed. However, rather than being biologically informed, the convolutions in this case are “biologically agnostic”, by which we mean that the values to which they are fixed are chosen at random, subject to the constraint that a) each filter’s values are binary, and b) each filter, at each position, must have exactly one non-zero entry out of the four entries corresponding to $\{A, C, G, T\}$. Like M_c, this model also has $23,079$ free parameters.

Finally, we also implemented a naive multi-task logistic regression model, which we call M_ℓ, with $96,000$ free parameters. The purpose of this model was to assess the performance of a method that is as non-biological as possible, in the sense that even a convolutional architecture encodes biological intuition about what causes TFs to bind to DNA.

We trained all four models for two epochs with a batch size of 50. We optimized the loss for each model using simple gradient descent with a learning rate of 0.1.

4 Results

The results of the four models on our test set are displayed in Figure 2, in which we quantify performance on each of our 24 tasks using the area under the precision-recall curve (AUPRC). (Our choice of AUPRC over AUROC is motivated by the fact that in cases of strong class imbalance AUROC can be high even for predictors with poor positive predictive value while AUPRC cannot.) We discuss three main takeaways from our results below and then describe two additional secondary analyses that we performed to better understand these results.

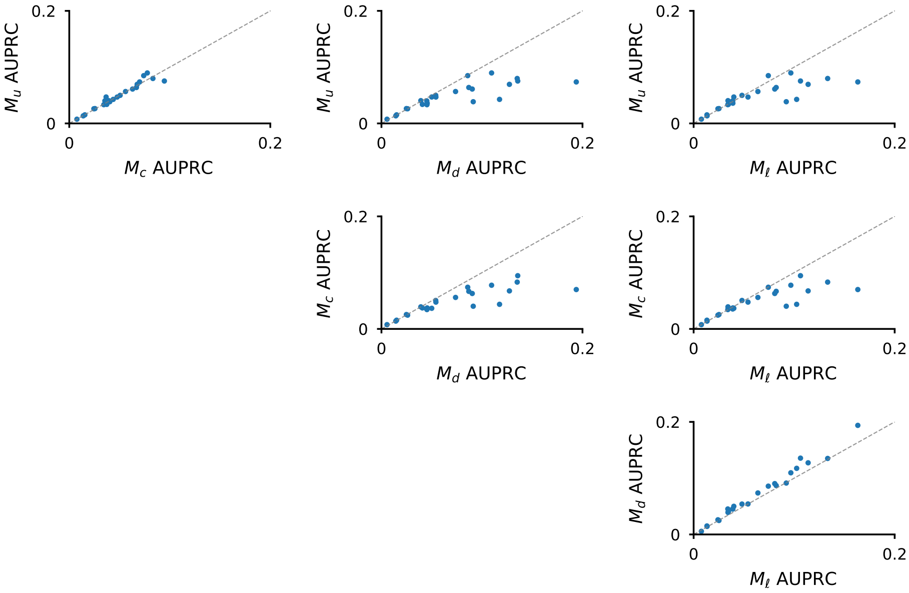

Figure 2: Results of all four models as quantified by area under the precision-recall curve (AUPRC) across all 24 tasks. Each plot corresponds to a pair of models. In each plot, we report for each task the AUPRC of the first model on that task against the AUPRC of the second model on that task. Recall that M_u is the unconstrained convolutional model, M_c is the biologically informed constrained convolutional model, M_d is the biologically agnostic constrained convolutional model, and M_ℓ is the logistic regression model.

4.1 M_c performs similarly to M_u

As the figure shows, the constrained model M_c does essentially as well as the unconstrained model M_u across all tasks. This leads to two major conclusions: first, it lends support to the prevailing theory that a major driver of TF binding is affinity to the characteristic, known motifs such as the ones that we downloaded. Second, it suggests that larger-scale predictive models of TF binding could benefit from incorporating mechanistic knowledge by using known motifs for their first-layer convolutions. This would have the advantage of interpretability, as previously described; for large-scale systems, it could also potentially reduce training times, though we did not observe this in our small-scale setting.

4.2 M_d dominates all models

A striking aspect of our results is that biologicall agnostic constrained convolutional model, M_d, is the best-performing model of all the ones tested. This is interesting because it is consistent with a more sophisticated understanding of TF binding wherein TFs do not just look for characteristic motifs but also care about more global properties of DNA sequence. To be more concrete, suppose that one length-10 motif were necessary and sufficient for the binding of a TF. This would be difficult for M_d to discern since it views the data through a basis consisting of just 14 randomly chosen non-probabilistic sequences of length 21. On the other hand, if more global aspects of DNA sequence, (such as the total proportion of G's and C's vs A's and T's in the sequence, which is known to influence how coiled versus uncoiled the sequence is[7]), played a dominant role, then these properties would likely still be detectable by M_d through its random basis. Our result therefore raises the intriguing dual possibility that a) there may be aspects of TF binding whose importance is currently under-appreciated, and b) that these may lead to more accurate predictive models in the future.

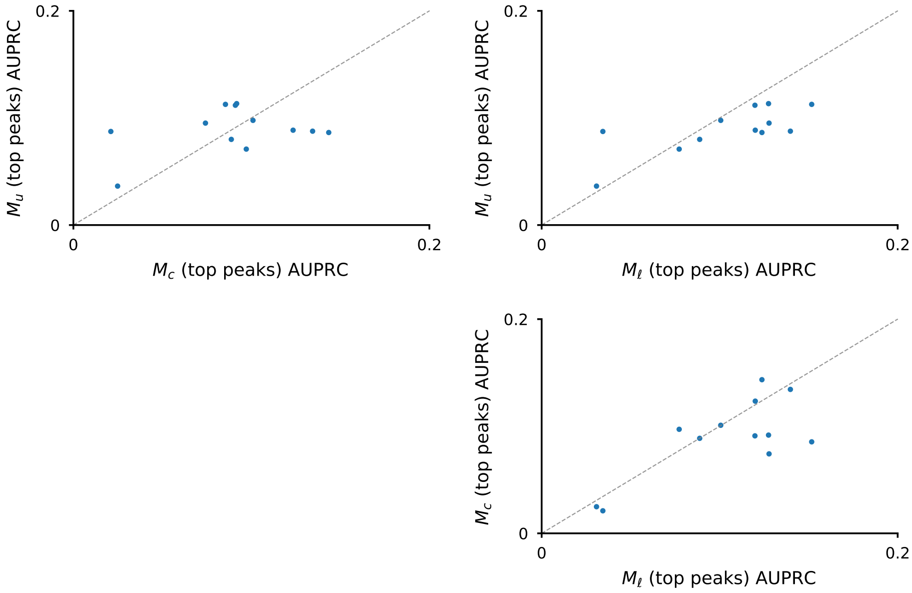

Figure 3: Results of models M_u, M_c, and M_ℓ as quantified by area under the precision-recall curve (AUPRC) across 12 tasks for which we had information about "top" TF binding sites. Each plot corresponds to a pair of models. In each plot, we plot for each task the AUPRC of the first model on that task against the AUPRC of the second model on that task.

4.3 M_ℓ outperforms M_u and M_c

We also found that the naive logistic regression model out-performed both the unconstrained convolutional network M_u and the biologically constrained convolutional network M_c. This too is consistent with global determinants of binding playing an under-appreciated role. To our knowledge, the performance of logistic regression for predicting TF binding with training data on the scale analyzed by modern convolutional network methods has not been characterized in previous work.

4.4 Additional analyses

We performed two additional secondary analyses to better understand our results.

4.4.1 Analysis of top peaks

TF binding sites have strengths associated with them, and some are stronger than others. In our work, we focused on a binary classification problem with our positive examples being any binding sites with strength above a relatively low threshold. It is possible that TF binding dynamics are different for the strongest binding sites of a given TF. To investigate this, we also analyzed a smaller set of "top binding sites" for the TFs in question. The story in these data was largely similar, though we perhaps did start to see differences between M_u and M_c on several of the tasks. See Figure 3

4.5 Single-task training rather than multi-task training

One unanticipated aspect of our training procedure that we did not appreciate initially was the subtlety of multi-task training. Recall that our classification task is "competitive" in that each example is treated as a positive for one task and a negative for all the other tasks. That is, each task is not about classifying TF binding sites away from non-binding sites but rather classifying binding sites for one TF away from binding sites other TFs. This has the significant advantage that it makes the classification task harder and ensures that our model does not simply learn, e.g., subtle ways that genomic sequences differ from random sequences.

Multi-task training also has many drawbacks, such as the danger that the model may attend to some tasks more than others, or that correlated tasks may result in the model up-weighting some types of inference over others. However, there is an additional drawback to our scheme that we had not anticipated: if the binding sites of two different TFs occur in similar looking genomic sequences, the model is effectively forced to assign that class of sequences to only one of the TFs. We observed this in our data, in the form of tasks on which our models did worse than chance. When we retrained the model on one such task in isolation, the area under the ROC curve (which would be 50% for a random guess on a balanced task) went from 34% (multi-task) to 74%. Moving forward, this poses questions about when and whether existing methodology[2], which also uses this competitive multi-task setup, could likewise be improved.

5 Limitations

Our work has several limitations. First, our analyses covered only a fraction of the TF binding data available, and our models were very simple relative to the ones used in the field. This is largely because such models require weeks to train, even on GPUs. For this reason, we only draw comparisons here between the simple models described in this work. However, given more time, it would be interesting to see if the phenomena we observed are present at scale.

Additionally, we did not perform an extensive hyperparameter search to optimize, e.g., the number of layers and sizes of convolutions in our architectures. Likewise, we did not optimize our learning rate, batch size, training time, etc. Perhaps for this reason, together with the issues of scale described above, the AUPRCs of our models are substantially lower than those of state-of-the-art models. Though our focus was on elucidating general properties of TF binding prediction through internally consistent comparisons of simple models, it is possible that our conclusions are partly a function of the "sand-box" style of our analysis.

Despite these limitations, we see this work as a first step toward a more systematic, data-driven analysis of TF binding dynamics and their relationship to existing prediction paradigms. Given that convolutional TF binding models are quickly becoming workhorses of biomedicine, understanding the interpretability, robustness, and complex behavior of these models is an increasingly urgent goal.

References

[1] Samuel A Lambert et al. "The human transcription factors". In: *Cell* 172.4 (2018), pp. 650–665.

[2] David R. Kelley, Jasper Snoek, and John Rinn. "Basset: Learning the Regulatory Code of the Accessible Genome with Deep Convolutional Neural Networks". In: *Genome Research* (May 3, 2016), gr.200535.115. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.200535.115. pmid: 27197224. URL: http://genome.cshlp.org/content/early/2016/05/03/gr.200535.115 (visited on 06/17/2017).

[3] Jian Zhou and Olga G. Troyanskaya. "Predicting Effects of Noncoding Variants with Deep Learning-Based Sequence Model". In: *Nature Methods* 12.10 (Oct. 2015), pp. 931–934. ISSN: 1548-7091. DOI: 10.1038/nmeth.3547. URL: http://www.nature.com/nmeth/journal/v12/n10/full/nmeth.3547.html (visited on 10/02/2017).

[4] Babak Alipanahi et al. "Predicting the Sequence Specificities of DNA- and RNA-Binding Proteins by Deep Learning". In: *Nature Biotechnology* 33.8 (Aug. 2015), pp. 831–838. ISSN: 1087-0156. DOI: 10.1038/nbt.3300. URL: http://www.nature.com/nbt/journal/v33/n8/full/nbt.3300.html?foxtrotcallback=true (visited on 10/02/2017).

[5] Yakir A Reshef et al. "Detecting genome-wide directional effects of transcription factor binding on polygenic disease risk". In: *bioRxiv* (2017), p. 204685.

[6] The ENCODE Project Consortium. "An Integrated Encyclopedia of DNA Elements in the Human Genome". In: *Nature* 489.7414 (Sept. 6, 2012), pp. 57–74. URL: http://www.nature.com/nature/journal/v489/n7414/full/nature11247.html (visited on 06/17/2017).

[7] Christopher D. Scharer et al. “ATAC-Seq on Biobanked Specimens Defines a Unique Chromatin Accessibility Structure in Naive SLE B Cells”. In: *Scientific Reports* 6 (June 1, 2016), p. 27030. ISSN: 2045-2322. DOI: 10.1038/srep27030. URL: http://www.nature.com/srep/2016/160601/srep27030/full/srep27030.html (visited on 05/03/2017).