Predicting protein-DNA binding affinity from structure and sequence

Alex Tseng
amtseng
https://github.com/atseng95/tf_dna_prediction

0 Abstract

A significant problem in computational biology is predicting the binding affinity of proteins to sequences of DNA. Here, I attempt to employ a convolutional neural net to predict binding affinity, starting with a protein structure and a DNA sequence. This architecture requires one example of the protein in question in a crystal structure bound to any sequence of DNA. Although the immediate results of the model do not adequately utilize structural information in making predictions, more atomistic features of the protein and DNA are implied to be sufficient in improving the model.

1 Introduction

Transcription factor binding to DNA effects many cellular phenotypes and diseases, including simple Mendelian diseases and more complex diseases such as cancer. Transcription factors yield these cellular changes by binding directly to DNA and encouraging or repressing a gene's transcription. Understanding how transcription factors bind to DNA can yield a better understanding of disease, and targeted drug design to affect transcription

I propose a neural net to learn how to predict whether a protein will bind to a DNA sequence or not. The input consists of a protein structure and a DNA sequence. The protein structure is a large grid of features denoting its pattern of amino acid residues when bound do DNA. The DNA input consists of a sequence—one-hot encoded—and certain features denoting its sequence-dependent shape. The output is a binary decision of whether or not the protein will bind to the sequence.

An overarching motivation of such a neural net architecture is to be able to easily change a single residue or a single base, and understand how binding affinity will change. Additionally, we would be able to scan the genome for binding motifs based on a structure. For example, we could understand how mutations in transcription factors change binding sites across the genome.

2 Related work

Since the problem of protein-DNA binding affinity is a very important one, much past work has been done to explore how proteins bind to DNA. This research, however, has focused largely on only small parts of the problem, such as identifying what proteins are DNA binding to begin with (Stawiski *et. al.*, 2003), which residues are DNA-binding (Gao and Skolnick, 2008), and which transcription factors bind preferentially or promiscuously (Corona and Guo, 2016). Other work has focused on deriving statistical or integrated potentials of protein-DNA interaction using a small set of solved structures (AlQuraishi and McAdams, 2011; Liu, *et. al.*, 2005; Farrel, *et. al.*, 2016).

Other research on the DNA side has solely examined single proteins at a time, including some deep learning methods to learn which sequences are DNA-binding, but treating the protein as a category rather than a physical entity with features (Quang and Xie, 2016).

Instead of focusing on small pieces or on deriving a potential, we attack the general problem: given a protein structure and a DNA sequence, decide whether or not it will bind. We require a known structure of the protein bound to any sequence, and we hope to learn whether it will bind to any new sequence. While a potential may answer this question as well, potentials are limited in the assumptions that they make. The potentials derived above make assumptions on its constituents and form. For example, many of the above innovations included terms for pi-pi stacking and hydrogen bonds, and most assume microscopic interactions are pairwise additive.

A neural net is more expressive of other types of interactions; it can express more ways of composing microscopic interactions, and may even be capable of finding new uncharacterized interactions.

3 Dataset and Features

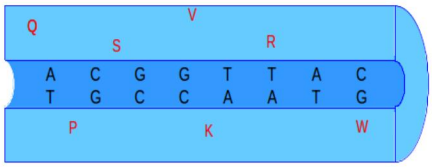
PDB (Berman, 2000) contains a large repository of protein structures, including some that are bound to some DNA sequence. There are roughly 4000 such protein-DNA structures, many which are redundant. A key insight on applying machine learning to this problem is that these data points can be greatly augmented by incorporating ChIP-seq data. Many of these proteins in the database are transcription factors of human value, and so there is knowledge of all locations in the genome which these transcription factors bind.

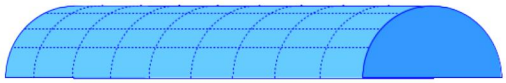
Out of PDB, there are 145 unique (nonredundant) proteins with available ChIP-seq data. Using the most recent ChIP-seq peaks from ChIP Atlas (Oki & Ohta, 2018), only the peaks of highest confidence ($q < 10^{-20}$) were selected. These peaks were then subjected to motif enrichment using HOMER (Heinz, *et. al.*, 2010), which found the most confident motifs ($p < 10^{-5}$) of length 12. Any ambiguous bases were filled in to expand to all possible sequences.

To identify non-binding sites, sequences of length 12 just adjacent to ChIP-seq peaks were found. These sequences are likely to be in the exact same cellular conditions as the peak itself (e.g. open chromatin), but the protein did not bind to this section of DNA. It is assumed that this is due to a lack of binding affinity rather than sterics.

In total, this yielded 60,000 sequences spanning 145 proteins, with balanced classes.

We treat a protein bound to DNA as a tunnel of residues surrounding a line in space. We break up this tunnel into 80 longitudinal sections of 1 angstrom, and 12 circumferential sections of 30 degree increments. Each tunnel encompasses a 20 base pair sequence of DNA. These cutoffs were selected based on histograms of how many bases are actually covered by residues.

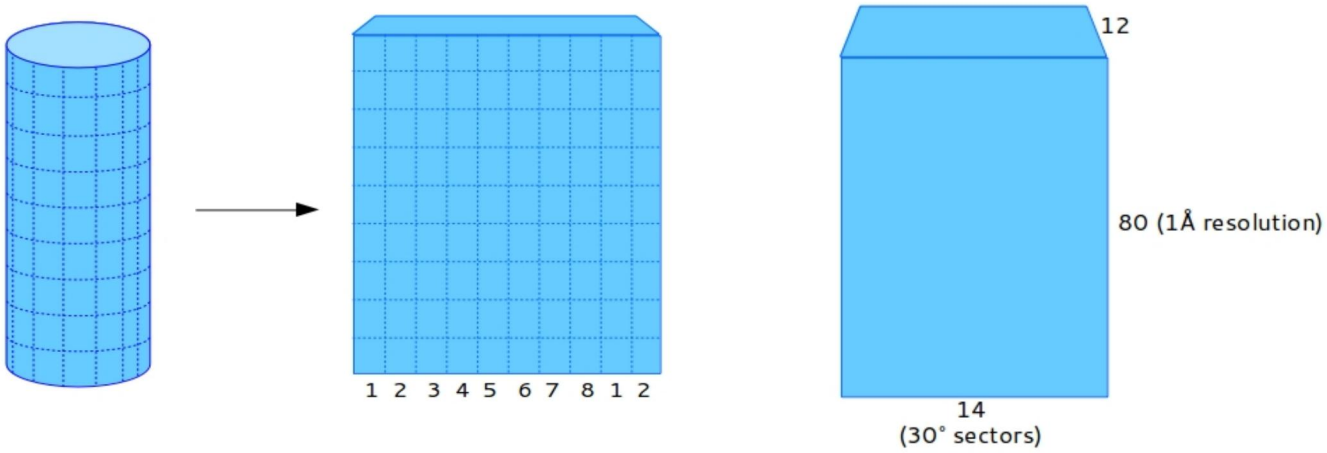




Each entry in the tunnel consists of a vector describing the closest residue to the central DNA axis, if any. This vector consists of locational properties (i.e. distance, and two angles to partially describe orientation in space), physical properties (e.g. charge, hydrophobicity, hydrogen bond donors, etc.), and identity properties (i.e. is it a glycine, is it a proline, is it a cysteine).

Note that although the example image above is simple, this process of extracting protein features is much more complex, because the DNA strand underneath is not straight, and so the cylinder curves around with the central core.

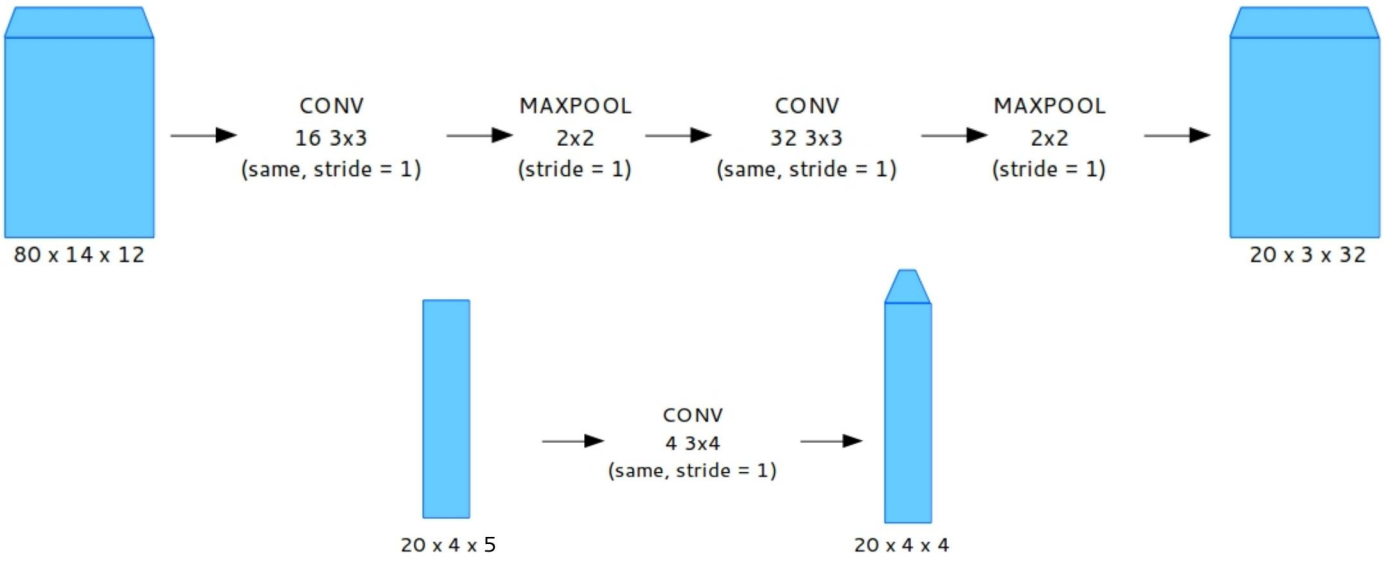
This protein tunnel is unrolled, and the last two columns (longitudinal slices) are replicated so that a convolutional filter may receive every portion of the cylinder equally.



The DNA sequence (of length 12) is one-hot encoded and zero-padded to length 20. DNAshapeR (Zhou, *et. al.*, 2013) was used to infer the minor groove width, propeller twist, helical twist, and roll of the DNA strand from the sequence.

4 Methods

The protein and DNA input are fed into convolutional layers separately as follows:



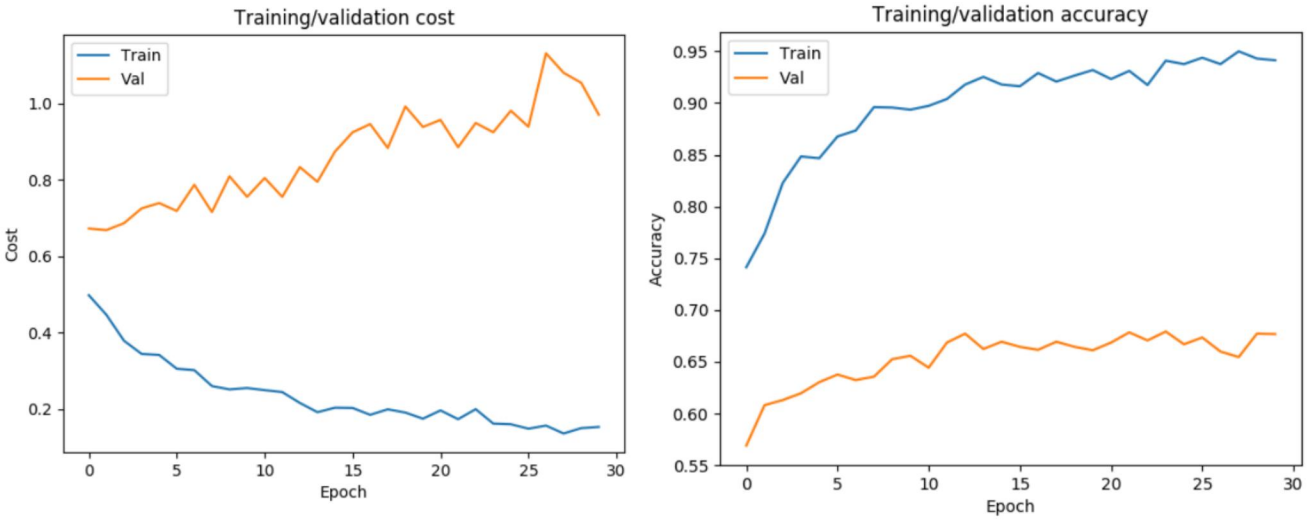
The results of each convolution are then flattened and concatenated, then fed into 4 fully-connected layers of 20, 20, 10, and 1 hidden unit(s).

Because the orientation and relative position of the DNA is critical for these spatial relationships, it is important for the network to know the correct directionality and the correct sliding window to position the DNA. To solve this problem, we simply input all 18 possible locations and reversals for a 12-base DNA strand in 20 spaces, and only output (and thus perform backpropagation) on the maximum score.

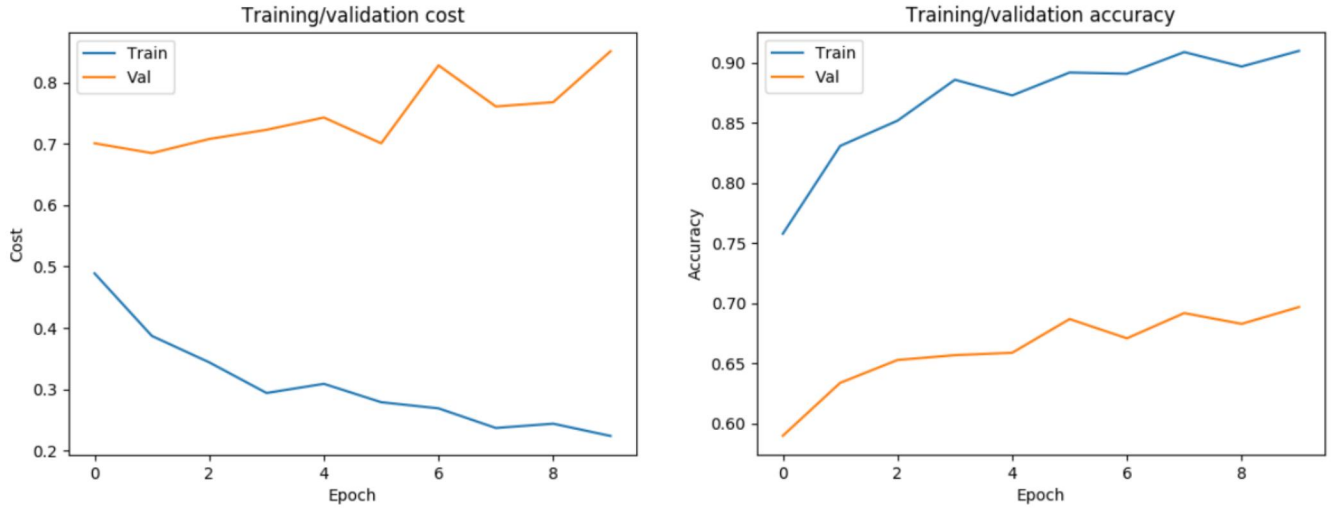
The model is trained using TensorFlow (Abadi, *et. al.*, 2016) with binary cross-entropy loss, using an Adam optimizer with an initial learning rate of 0.005 and batch sizes of 64. Hyperparameter tuning suggests this optimal learning rate, and a relatively inconsequential selection of optimizer and batch size.

5 Results and Discussion

After training for 30 epochs on the training set, the following cost and accuracy patterns are as shown:



Since the positive and negative classes are balanced, a validation accuracy of almost 70% without regularization or dropout can seem promising. However, the same results are obtained when the protein features are all set to 0:



Additionally, the same pattern occurs when the experiment is repeated with only the canonical DNA sequences found in the crystal structures. That is, without the ChIP-seq data at all.

This suggests that structure is consistently being ignored. That is, the model is not learning how residues interact with DNA bases, but how different DNA sequences (and perhaps different DNA shapes) tend to be binding. This is an interesting result in itself, because it shows that there is some signal in the pattern of DNA sequences that tend to bind or not bind to proteins in general. Unfortunately, however, it is not the goal of this initiative.

Additionally, the same pattern of structure being ignored is present even when the dataset is subset only to the canonical sequences in the crystal structures. This removes any dependency on the quality of ChIP-seq peaks and motif enrichment. Since structure remains to be unhelpful to performance in this case, it intimates that the structural features are simply insufficient.

As previous work has demonstrated, structural features from an atomistic view are sufficient to derive statistical and integrated potentials that are reflective of binding affinity. Thus, such features must also exist that would allow structure to be helpful in this current classification task.

6 Conclusions and Future Work

The work here represents an attempt to solve a rather challenging, but significant problem. Predicting the binding affinity of proteins to DNA sequences is an impactful problem, and previous attempts using statistical and integrated potentials show it is possible. In order to fix this current attempt using deep learning, different structural features must be used.

Future work will focus on identifying these features needed to propagate binding signal through the network. A balance must be struck on the specificity of the features to the underlying physics, and the ability to model in new sequences and residues. Features that sit closer to an atomistic view also sit closer to models that are known to perform decently, but render the model difficult or impossible to include new sequences or proteins for which there are no crystal structures. Ideally, a middle-ground can be found, which brings the physical signal through the network, but remains general enough to allow for relative ease in modeling new sequences and proteins.

7 Contributions

All of the work presented here was completed by Alex Tseng, although a lot of thought and advice arises from Joe Paggi, Ron Dror, and David Eng.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., … Zheng, X. (2016). TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. Retrieved from http://arxiv.org/abs/1603.04467

AlQuraishi M, McAdams HH. Direct inference of protein–DNA interactions using compressed sensing methods. *Proceedings of the National Academy of Sciences of the United States of America*. 2011;108(36):14819-14824. doi:10.1073/pnas.1106460108.

Berman HM. The Protein Data Bank. Nucleic Acids Res. 2000;28(1):235-242. doi:10.1093/nar/28.1.235

Corona RI, Guo J. Statistical analysis of structural determinants for protein-DNA binding specificity. *Proteins*. 2016;84(8):1147-1161. doi:10.1002/prot.25061.

Farrel A, Murphy J, Guo J. Structure-based prediction of transcription factor binding specificity using an integrative energy function. *Bioinformatics*. 2016;32(12):i306-i313. doi:10.1093/bioinformatics/btw264.

Heinz S, Benner C, Spann N, Bertolino E et al. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Mol Cell 2010 May 28;38(4):576-589.

Gao M, Skolnick J. DBD-Hunter: a knowledge-based method for the prediction of DNA–protein interactions. *Nucleic Acids Research*. 2008;36(12):3978-3992. doi:10.1093/nar/gkn332.

Liu Z, Mao F, Guo J, Yan B, Wang P, Qu Y, Xu Y; Quantitative evaluation of protein–DNA interactions using an optimized knowledge-based potential, *Nucleic Acids Research*, Volume 33, Issue 2, 1 January 2005, Pages 546–558, https://doi.org/10.1093/nar/gki204

Oki S, Ohta T, et al. Integrative analysis of transcription factor occupancy at enhancers and disease risk loci in noncoding genomic regions. *bioRxiv* 262899; doi: https://doi.org/10.1101/262899

Stawiski EW, Gregoret LM, Mandel-Gutfreund Y; Annotating Nucleic Acid-Binding Function Based on Protein Structure, *Journal of Molecular Biology*, Volume 326, Issue 4, 2003, Pages 1065-1079, https://doi.org/10.1016/S0022-2836(03)00031-7.

Zhou T, Yang L, Lu Y, et al. DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. Nucleic Acids Res. 2013;41(W1):W56-W62. doi:10.1093/nar/gkt437