Neural networks can extract thermodynamic DNA sequence affinities from *in vivo* binding profiles of transcription factor binding

Amr M. Alexandari
Computer Science Dept.
Stanford University

Connor Horton
Genetics Dept.
Stanford University

Avanti Shrikumar
Computer Science Dept.
Stanford University

Polly Fordyce
Genetics and Bioengineering Depts.
Stanford University

Anshul Kundaje
Genetics and Computer Science Depts.
Stanford University

Abstract

Transcription factors (TFs) bind genomic DNA in a sequence-specific manner to regulate gene expression. *In vitro* TF-DNA binding assays can measure the intrinsic DNA binding affinity of individual TFs. Complementary *in vivo* TF binding experiments can profile genome-wide TF occupancy that is influenced by several factors besides intrinsic sequence specificity. Deep learning models can accurately map genomic DNA sequence to *in vivo* TF binding profiles and predict effects of sequence mutations on binding occupancy. However, it has been difficult to provide biophysical interpretations of these predictions. Here, we show that neural networks trained to model base-resolution genomic occupancy profiles of TFs in yeast and humans can predict effects of sequence variation in and around core binding sites that are remarkably correlated with corresponding binding energy measurements from *in vitro* experiments. We show that the models can learn exquisitely detailed motif-flanking sequence preferences of paralogous TFs as well as effects of repetitive motif-flanking sequences on occupancy and affinity. We also find that binding affinity has a much stronger contribution to genomic occupancy signal of TFs in yeast as compared to occupancy profiles of TFs in humans. Our results indicate that with appropriate correction of experimental biases, deep learning models can learn to extract thermodynamic affinities *de-novo* from genomic occupancy profiles. This unique biophysical interpretation of predictions of deep learning oracles of genomic TF occupancy opens a new avenue to perform massive *in-silico* perturbation experiments to comprehensively decipher the influence of sequence context and variation on intrinsic affinity and *in vivo* occupancy.

1 Introduction

Gene expression is modulated by sequence-specific binding of transcription factors (TF) to regulatory DNA elements in the genome. Protein binding microarrays, HT-SELEX and BET-seq assays allow estimation of relative binding energies ($\Delta\Delta G$) for libraries of synthetic DNA sequences [Le et al., 2018]. PB-seq and PB-exo experiments measure binding affinity of TFs to *in vitro* purified genomic DNA [Rossi et al., 2018]. These *in vitro* data have been used to learn thermodynamic models of DNA binding affinity of individual TFs [Le et al., 2018, Rastogi et al., 2018]. However, *in vivo* binding of TFs to chromatin is not only a function of equilibrium binding affinity, but also influenced by DNA and nucleosome mediated cooperative and competitive interactions between TFs, local chromatin state, three dimensional chromatin architecture and local TF concentration. Assays such as Chromatin immunoprecipitation followed by sequencing (ChIP-seq, ChIP-exo) profile genome-wide *in vivo* TF occupancy. Deep learning models such as convolutional neural networks (CNNs) can accurately map genomic DNA sequences to *in vivo* binding profiles by learning predictive cis-regulatory motif syntax and predict the effect of sequence variation on binding occupancy [Alipanahi et al., 2015, Zhou and Troyanskaya, 2015, Avsec et al., 2019]. However, the quantitative relationship between these blackbox predictive models of *in vivo* TF occupancy and the classic thermodynamic models of binding affinity has been elusive [He et al., 2010]. Here, we show that CNNs trained to map genomic DNA to quantitative base-resolution in vivo TF occupancy profiles from ChIP-exo and ChIP-seq experiments in yeast and humans are able to extract binding affinities *de novo* from *in vivo* occupancy and can predict the effects of subtle sequence variation on relative binding energy as measured by in vitro BET-seq experiments with remarkable accuracy.

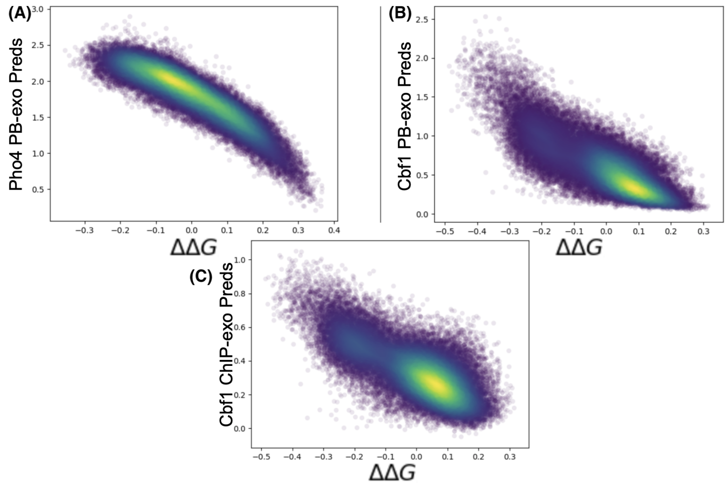

Figure 1: Model predictions against $\Delta\Delta G$ experimentally determined by BET-seq [Le et al., 2018]. Each point represents an affinity measurement for a sequence containing an E-box motif with unique 5 base pair flanks on either side with each plot containing $50,000$ predictions/measurements. Subfigure A contains predictions of PB-exo Pho4 model (Spearman correlation of -0.90). Subfigure B contains predictions of PB-exo Cbf1 model (Spearman correlation of -0.77). Subfigure C contains predictions of ChIP-exo Cbf1 model (Spearman correlation of -0.67).

2 CNN models of ChIP-exo and PB-exo binding profiles of two yeast TF paralogs can predict relative binding energies of sequence variation flanking core E-box motifs

We trained BPNet CNN models [Avsec et al., 2019] to map 546 base-pair DNA sequences from the *S. cerevisiae* genome to base-resolution PB-exo and ChIP-exo profiles of two yeast TFs Pho4 and Cbf1 that both bind to E-box motifs [Rossi et al., 2018]. The correlation of predicted and observed read coverage (TF occupancy) across peaks in held-out test chromosomes was on par with concordance between replicate experiments (pearson correlation = 0.93), despite the small size of training sets (< 4K peaks). We used the trained models to predict Pho4 and Cbf1 occupancy for millions of synthetic sequences which contained a high affinity core E-box motif (CACGTG) with systematic sequence variations in the ±5 bp flanking the motif. We found that these predictions from the PB-exo/ChIP-exo models had remarkable agreement with corresponding relative binding affinities ($\Delta\Delta G$) for the same library of sequences estimated using *in vitro* BET-seq experiments [Le et al., 2018] (Figure 1). We calibrated the predictions of the BPNet model to the binding energies using linear regression and isotonic regression [Chakravarti, 1989].

Next, we used the DeepSHAP [Lundberg and Lee, 2017]) feature attribution method to infer base-resolution importance scores for all bound sequences for each TF. These importance score profiles were used to learn non-redundant motif representations for Pho4 and Cbf1 using the TF-MoDISco algorithm [Shrikumar et al., 2018]). The derived motifs for the two TFs showed distinct flanking sequence preferences that matched previously derived binding affinity motif models from BET-seq data (Figure 2).

These results indicate the BPNet models are able to implicitly extract binding energies from the in vivo occupancy profiles and can distinguish subtle differences in flanking sequence preferences of two yeast TFs that bind very similar E-box motifs.

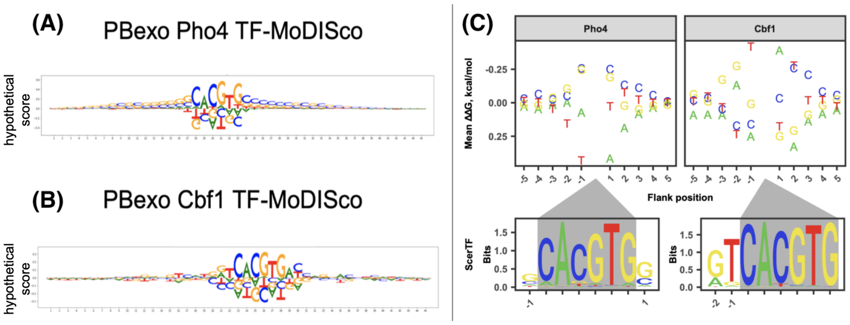

Figure 2: The top TF-MoDISco motifs for the PB-exo Pho4 model (subfigure A) and PB-exo Cbf1 model (subfigure B) corresponded with the most favorable sites by relative binding affinity $\Delta\Delta G$. From [Le et al., 2018]: subfigure C shows Pho4 and Cbf1 mean $\Delta\Delta G$ values as a function of flanking sequence position (Top), compared with the ScerTF database sequence logos (Bottom). Gray boxes show position of core consensus CACGTG.

3 CNN model of ChIP-seq binding profiles of the NR3C1 TF in the human genomes accurately predicts *in vitro* binding energies of mutations in motif

We trained similar BPNet CNN models to map 1346 bp sequences from the human genome to base-resolution *in vivo* ChIP-seq profiles of the glucocorticoid receptor (NR3C1) in the A549 human cell-line [Vockley et al., 2016]. Unlike the yeast TFs, the NR3C1 CNN model showed moderate correlation between predicted and observed occupancy for peaks in held-out test chromosomes compared to replicate concordance (Fig 3A,B), suggesting that local sequence cannot fully predict *in vivo* occupancy of this TF in humans. Next, we used the trained CNN to predict occupancy for hundreds of synthetic sequences containing the NR3C1 consensus motif sequence with random sequence variation at one or two random positions in the in the motif. Remarkably, we found that these predictions from ChIP-seq CNN model had very strong agreement with corresponding relative binding affinities inferred from *in vitro* microfluidic experiments (Figure 3C). These results indicate that even for human TFs, the CNN models are able to extract binding energies from the in vivo data despite the fact that these in vivo measurements contain extraneous signal not encoded in the local DNA sequence.

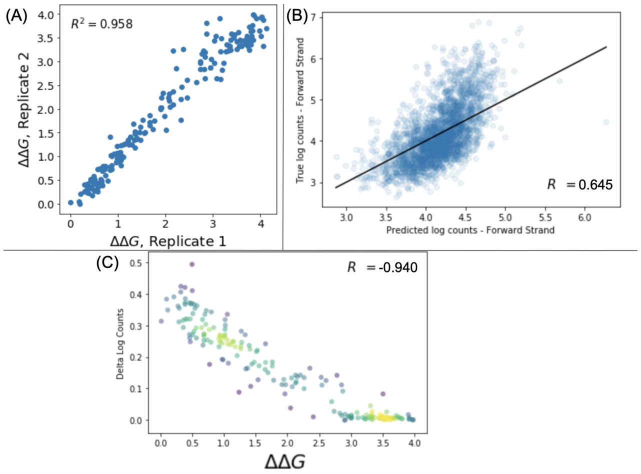

Figure 3: Subfigure A shows the replicate vs replicate correlation for the relative binding energy ($\Delta\Delta G$) obtained from microfluidic experiments. Subfigure B shows the correlation of observed and model predicted read coverage across peaks in held-out test chromosomes. Subfigure C shows the correlation between the observed $\Delta\Delta G$ and the predictions from the same model.

4 Estimating the Binding Influence of Repetitive Sequences Flanking E-box Motifs via *In-Silico* Experiments

Prior work [Afek et al., 2014] has shown that repetitive sequences may increase binding *in-vitro*. Motivated by this, we systematically interrogated the models with synthetic DNA sequences containing different types and lengths of repetitive sequences flanking core E-box motifs to understand their influence on binding affinity. (Figure 4) highlights two such experiments to understand the effect of 'CT' dinucleotide repeats and 'GT' repeats of varying lengths on Cbf1 binding. The trends observed from the model predictions strongly agreed with smaller scale *in-vitro* measures of binding affinity, which demonstrate that repetitive sequences can modulate binding affinity *in-vitro*.

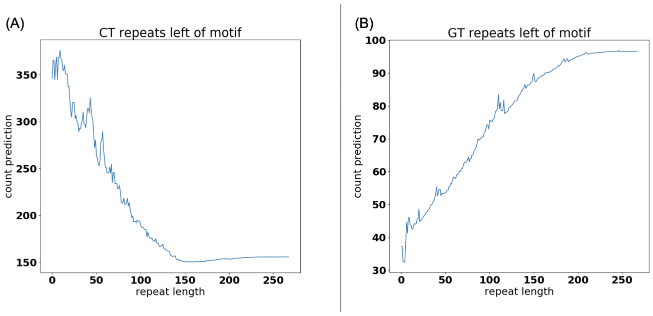

Figure 4: We chose a low count dinucleotide shuffled test region background and inserted the extended motif GTCACGTGAC. Then we added increasing lengths from 0 to 267 to the left of the motif and measured the counts prediction of the model. Subfigure A shows the unfavorable effect of increasing lengths of 'CT' dinucleotide repeats on Cbf1 binding as predicted by ChIP-exo derived model. Subfigure B shows the favorable effect of increasing lengths of 'GT' repeats on Cbf1 binding as predicted by PB-exo derived model. Interestingly, both figures show a saturating reaction to the increasing repeat lengths up to 150 bp.

5 Conclusion

Our results indicate that deep learning models can extract thermodynamic affinities *de-novo* from *in vivo* occupancy profiles of TFs in yeast and humans. This biophysical interpretation of predictions of deep learning oracles of genomic TF occupancy opens a new avenue to perform massive *in-silico* perturbation experiments to comprehensively decipher the influence of sequence context and variation on intrinsic affinity and *in vivo* occupancy.

References

Daniel D Le, Tyler C Shimko, Arjun K Aditham, Allison M Keys, Scott A Longwell, Yaron Orenstein, and Polly M Fordyce. Comprehensive, high-resolution binding energy landscapes reveal context dependencies of transcription factor binding. *Proceedings of the National Academy of Sciences*, 115(16):E3702–E3711, 2018.

Matthew J Rossi, William KM Lai, and B Franklin Pugh. Genome-wide determinants of sequence-specific dna binding of general regulatory factors. *Genome research*, 28(4):497–508, 2018.

Chaitanya Rastogi, H Tomas Rube, Judith F Kribelbauer, Justin Crocker, Ryan E Loker, Gabriella D Martini, Oleg Laptenko, William A Freed-Pastor, Carol Prives, David L Stern, et al. Accurate and sensitive quantification of protein-dna binding affinity. *Proceedings of the National Academy of Sciences*, 115(16):E3692–E3701, 2018.

Babak Alipanahi, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. *Nature biotechnology*, 33 (8):831–838, 2015.

Jian Zhou and Olga G Troyanskaya. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature methods*, 12(10):931–934, 2015.

Žiga Avsec, Melanie Weilert, Avanti Shrikumar, Amr Alexandari, Sabrina Krueger, Khyati Dalal, Robin Fropf, Charles McAnany, Julien Gagneur, Anshul Kundaje, et al. Deep learning at base-resolution reveals motif syntax of the cis-regulatory code. *BioRxiv*, page 737981, 2019.

Xin He, Md Abul Hassan Samee, Charles Blatti, and Saurabh Sinha. Thermodynamics-based models of transcriptional regulation by enhancers: the roles of synergistic activation, cooperative binding and short-range repression. *PLoS Comput Biol*, 6(9):e1000935, 2010.

Nilotpal Chakravarti. Isotonic median regression: a linear programming approach. *Mathematics of operations research*, 14(2):303–308, 1989.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, pages 4765–4774, 2017.

Avanti Shrikumar, Katherine Tian, Žiga Avsec, Anna Shcherbina, Abhimanyu Banerjee, Mahfuza Sharmin, Surag Nair, and Anshul Kundaje. Technical note on transcription factor motif discovery from importance scores (tf-modisco) version 0.5. 1.1. *arXiv preprint arXiv:1811.00416*, 2018.

Christopher M Vockley, Anthony M D'Ippolito, Ian C McDowell, William H Majoros, Alexias Safi, Lingyun Song, Gregory E Crawford, and Timothy E Reddy. Direct gr binding sites potentiate clusters of tf binding across the human genome. *Cell*, 166(5):1269–1281, 2016.

Ariel Afek, Joshua L Schipper, John Horton, Raluca Gordân, and David B Lukatsky. Protein- dna binding in the absence of specific base-pair recognition. *Proceedings of the National Academy of Sciences*, 111(48):17140–17145, 2014.