Prediction of transcription factor binding based on DNA physical properties

Claire Yang
ID: 260898597
youyou.yang@mail.mcgill.ca

ABSTRACT
This project studies how sequence and DNA shape features contribute to transcription factor binding prediction. Two factors with different binding behaviors, CTCF and BHLHE40, were selected to test whether structural information provides a useful signal beyond sequence. We built datasets from chromosome 1 using annotated binding sites, motif-matched negatives in accessible chromatin, and five genome-wide shape tracks. Logistic regression and SVM models were trained under three settings: sequence-only, shape-only, and a combination of the two. CTCF was easy to classify. The sequence alone reached perfect performance. When testing with BHLHE40, shape features gave slight gains for logistic regression, and combining sequence and shape produced the best AUPRC. These results show that the value of DNA shape depends on the transcription factor.
The code is open-sourced in github.com/wppqywq/COMP561_TF_binding.

1 INTRODUCTION

Transcription factors (TFs) are proteins that control gene expression by binding to specific DNA sequences. Their job is to find the specific sequence fraction on our DNA for gene regulation. Scientists can computationally scan the entire human genome to find sequences binding site that look like the right lock for a particular TF key, namely the 'motif'. Rohs et al. (2009) find that the TF binds to only a fraction of these potential sites. Many perfectly matching DNA sequences are simply ignored. This raises the finding that the three-dimensional structure of DNA also plays a crucial role. The DNA double helix is not perfectly regular. The reason is that the specific nucleotide sequence creates subtle variations in structural properties. These properties include: (i) Minor Groove Width (MGW): The width of the minor groove; (ii) Roll: The angle between adjacent base pairs; (iii) Propeller Twist (ProT): The twist angle within a base pair; (iv) Helical Twist (HelT): The twist angle between adjacent base pairs. These structural properties can influence protein-DNA interactions.(Tp et al., 2016) Thus, we can explain why some sequences that "look like" they should be bound based on motif matching, but are not actually bound in vivo. (Zhou et al., 2015)

In this project, we study this question: Can we build a model that learns to distinguish between DNA sites that a TF truly binds or not, by looking at both the DNA sequence and its predicted 3D shape?

2 METHODOLOGY

To investigate this question, the project builds machine learning classifiers that distinguish between genomic regions bound by transcription factors and regions that remain unbound within the same regulatory context. In contrast to using a single factor, we include two human transcription factors, **CTCF** and **BHLHE40**, to evaluate how sequence-dominated and shape-dependent binding behaviors differ under the same modeling framework.(Hu, 2020)

[h!]

Table 1. Dataset size on chr1 for the two transcription factors. Parentheses indicate the fraction of positives in the dataset.

TF	Positive Sites	Negative Sites	Total
CTCF	14,648 (57.6%)	10,780	25,428
BHLHE40	1,609 (11.8%)	11,995	13,604

2.1 Dataset Assembling

CTCF and BHLHE40 represent two contrasting families of transcription factors. CTCF is a well-studied architectural protein with a highly conserved 15–20 bp motif and strong sequence specificity, while BHLHE40 is a basic helix-loop-helix transcription factor whose binding is known to be more sensitive to local DNA structural variation. (X et al., 2022; Jd et al., 2008) Including both factors allows us to test whether the contribution of DNA shape features remains consistent across TFs with different binding mechanisms.

The whole dataset is built from available biological data:

- Human Genome Sequence (hg19): Sourced from the UCSC Genome Browser, these FASTA files were used to extract the raw DNA sequences for all genomic windows under investigation.

- Active DNA Regions (GM12878 cell line): This dataset from the ENCODE project defines regions of open chromatin in the GM12878 lymphoblastoid cell line. It was used to define the universe from which negative (unbound) samples were drawn.

- Known TF Binding Sites (Factorbook): A pre-existing catalog of genomic locations where the CTCF/BHLHE40 protein is experimentally known to bind.(J et al., 2012)

- DNA Shape Features: Shape features from precomputed genome-wide wig tracks: For experiments limited to chromosome 1, five precomputed shape features (MGW, ProT, Roll, Buckle, Opening) were extracted from genome-wide wig files.

The DNA shape features were standardized using z-score normalization. From these sources, two distinct groups of DNA sequences were created:

Positive Sites: All annotated binding sites for the TF were collected from Factorbook, and a 101 bp window was extracted around each center.

Negative Sites: These are DNA sequences located in active regions that are known to be unbound by the TF, even though they contain a motif. We sampled 101 bp windows from the active regulatory regions in GM12878 cells. A key design choice was to ensure these windows did not overlap with any known binding sites of the corresponding TF. This strategy contrasts with using a random genomic background, as it forces the models to distinguish bound sites from other accessible but unbound regions within the same regulatory context. A large set of negative controls was then sampled to create a dataset with a comparable number of instances in each class.

After filtering and processing from the whole chromosome 1 sequence (legnth=249,250,621), the final dataset used for the analysis is presented in Table 1. Comparing to BHLHE40, CTCF has a substantially larger number of experimentally annotated sites on chr1 and a more balanced dataset.

Based on that, we use a coordinate-based split, allocating the first 60% of the chromosome for training, the next 20% for validation, and the final 20% for testing.

2.2 Models and Evaluations

To isolate and quantify the effect of sequence versus shape, we restrict the modeling framework to two classical classifiers:

Linear Logistic Regression: A generalized linear model trained either on one-hot encoded sequence features or on the five DNA shape features. Logistic regression provides an interpretable baseline and helps identify whether class separation can be achieved in a linear feature space.

Support Vector Machine (SVM) with RBF Kernel: A nonlinear model trained with identical feature inputs. The RBF kernel

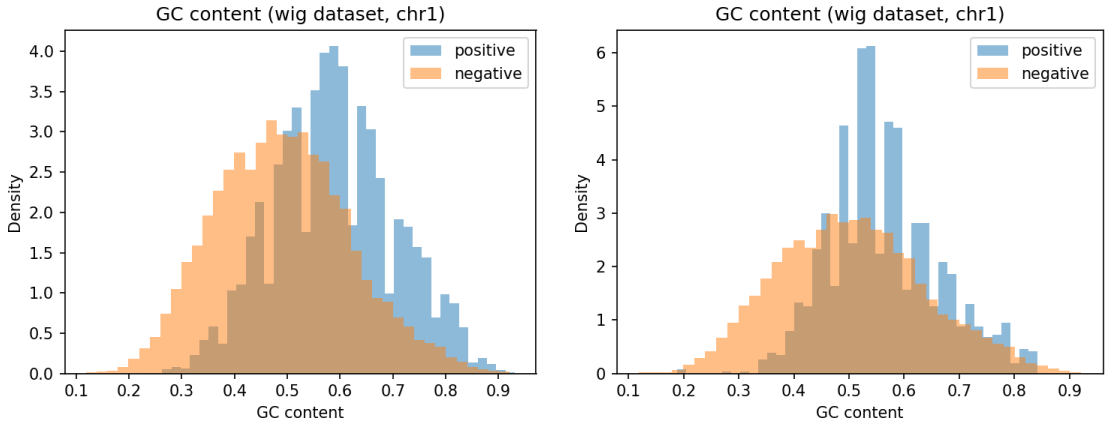

Figure 1. GC Content Distribution of Positive and Negative Sites. Left: CTCF; right: BHLHE40.
For CTCF, the positive sites show a strong shift toward higher GC levels (mean = 0.589), while negative sites cluster around a lower mean of 0.491. The two distributions have limited overlap. For BHLHE40, the contrast is weaker. Positive sites have a mean GC content of 0.563, while negatives average 0.505, and the two distributions largely overlap.

allows the classifier to capture higher-order interactions and nonlinear dependencies between shape features and sequence context.(Zhou et al., 2015)

By training each model under three configurations: sequence-only, shape-only, and sequence-shape-combination. We can directly compare how sequence-dominated (CTCF) and shape-sensitive (BHLHE40) factors differ in model performance.

Model performance was evaluated using the Area Under the Receiver Operating Characteristic curve (AUROC) and the Area Under the Precision–Recall Curve (AUPRC). AUROC measures discrimination ability across thresholds, while AUPRC is more informative in imbalanced datasets. For both metrics, 1.0 indicates perfect classification.

3 RESULTS

3.1 GC Content Distinguishes Bound and Unbound Sites

We compared the GC content of the 101 bp windows for both transcription factors in Figure 1. We aim to evaluate whether bound and unbound regions differ in basic sequence composition. For CTCF, the positive sites show a clear shift toward higher GC content, consistent with the GC-rich nature of the canonical CTCF motif.

For BHLHE40, the pattern is noticeably different. The GC content of positive and negative sites largely overlaps. This indicates that GC composition alone does not separate bound regions from unbound ones for BHLHE40. Negative sites are distributed at lower GC levels, with a wider range of accessible chromatin sequences. This contrast with CTCF highlights how different transcription factors rely on distinct sequence cues, and it provides a baseline expectation for the relative performance of sequence-based classifiers across the two TFs.

3.2 Performance

Across both transcription factors, we evaluated linear logistic regression and nonlinear SVM models under various conditions. Results are in Figure 2 and Table 2.

We first trained the linear models on CTCF, given its well-balanced dataset. Sequence-only models achieved near-perfect performance (AUROC = 1.00) with a very short runtime of $\sim$ 1s. Shape-only model performance slightly drops (AUROC $\sim$ 0.98), with higher runtime. The combining features did not meaningfully improve performance, indicating that the most discriminative signal for CTCF arises directly from the sequence. The SVM gives a similar but slightly lower performance than the linear model, while requiring a much longer training time ($\geq$200s with shape features included.)

[h!]

Table 2. Performance comparison of Logistic Regression and SVM (RBF) on chr1 for CTCF and BHLHE40.

TF	Model	Features	AUROC	AUPRC	Runtime (s)
CTCF	LogReg	Shape	0.9829	0.9839	1.99
	LogReg	Seq	**1.0000**	**1.0000**	1.07
	LogReg	Seq + Shape	1.0000	1.0000	1.18
CTCF	SVM (RBF)	Shape	0.9790	0.9804	200.53
	SVM (RBF)	Sequence	**1.0000**	**1.0000**	56.70
	SVM (RBF)	Seq + Shape	0.9999	0.9999	286.74
BHLHE40	LogReg	Shape	0.9954	0.9641	1.91
	LogReg	Seq	0.9924	0.9297	1.05
	LogReg	Seq + Shape	**0.9992**	**0.9939**	1.08
BHLHE40	SVM (RBF)	Shape	0.9906	0.9703	19.81
	SVM (RBF)	Sequence	**0.9986**	0.9881	16.48
	SVM (RBF)	Seq + Shape	0.9982	**0.9901**	40.44

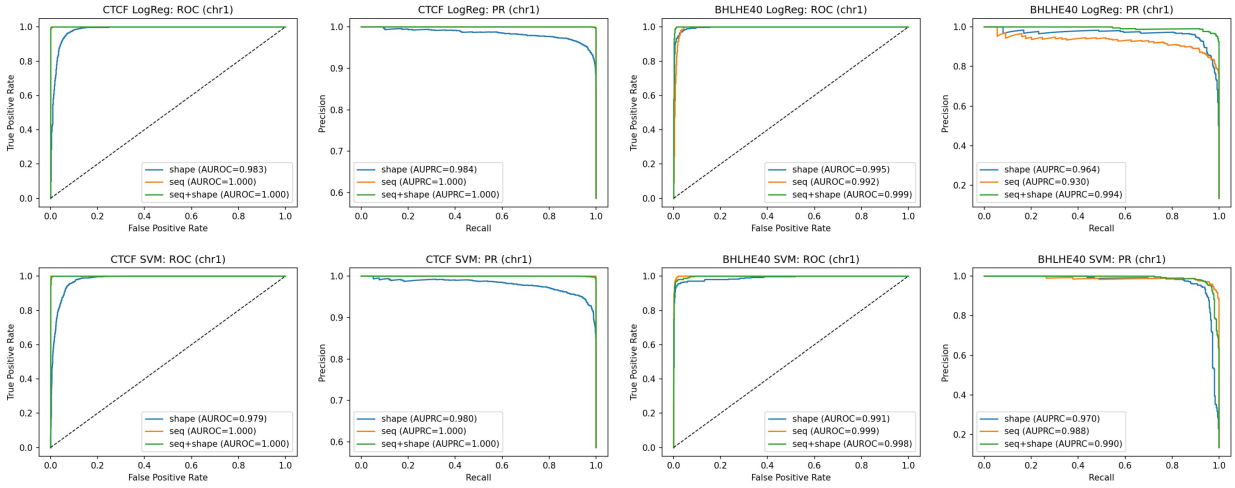

Figure 2. ROC and PR curves using logistic regression and SVM under three feature settings. Left: CTCF; right: BHLHE40.

Given the performance does note increase when shape features are added, under the CTCF data, we continue the further experiment with the BHLHE40 data. In contrast, BHLHE40 displayed a different pattern. For logistic regression, shape features give the highest stability, with AUROC near 0.995 and a clear advantage in the AUPRC curve over sequence-only. The sequence-only model performs slightly worse, especially in precision at higher recall, which matches the earlier observation that GC content does not separate positive and negative sites for this factor. Combining sequence and shape produces the strongest results for logistic regression, reaching AUROC = 0.999 and a large improvement in AUPRC. This answers our initial question about whether shape contributes beyond sequence.

For SVM, all three feature sets perform well, but the pattern is slightly different. Sequence-only SVM reaches AUROC = 0.9986, while shape-only falls behind. The fused model does not show a large gain in AUROC but gives more stable precision at high recall. Across both models, BHLHE40 behaves as a factor whose binding cannot be explained by sequence composition alone, and structural features give consistent improvements in recall. The runtime cost with BHLHE40 SVM is much lower than the CTCF given its limited data size ($\sim$20s with single feature sources.).

4 DISCUSSION & FUTURE WORK

This project compares two transcription factors with different binding behaviors and evaluates whether DNA shape features improve prediction. The results show a clear contrast. For CTCF, sequence information alone was enough to reach perfect performance. Both logistic regression and SVM achieved AUROC and AUPRC values of 1.0 using only the one-hot sequence. Adding shape features did not increase performance, and in some cases increased runtime by a large margin. This fits the observations from the data. CTCF has a long and conserved motif, and the GC shift between positive and negative sites almost separates the two classes by itself. In this setting, the models do not need additional structural features, and the learning problem becomes simple.

BHLHE40 behaves differently. Sequence alone gave strong results, but shape features provided clear gains. Logistic regression on shape features outperformed the sequence-only model, and combining the two produced the best AUPRC. The overlap in GC content explains why sequence alone does not dominate. Shape captures local structural preferences that are not encoded by simple motif composition, and the linear model was able to use these differences. The SVM showed less separation between feature types, partly because the nonlinear kernel already captures interactions that the linear model cannot. Even so, combining sequence and shape improved stability at high recall, which matches the expectation that BHLHE40 does not rely on a strict motif.

Several limitations of the current design suggest directions for improvement. All experiments were restricted to chromosome 1. This reduces the genomic diversity of both positive and negative examples and may inflate performance, especially for the sequence-only CTCF models. Expanding to the whole genome would make the evaluation more realistic. The negative sampling strategy also affects the difficulty of the task. Negative windows were drawn from accessible chromatin and required to contain the motif, but this still does not guarantee a uniform distribution of local sequence patterns. Alternative sampling schemes or cross-cell-type negatives could increase robustness.

The shape features used here come from precomputed wig tracks and represent only five dimensions. These tracks do not capture all aspects of DNA geometry. Other shape descriptors, including higher-order roll interactions or models based on molecular simulations, may provide more detailed signals. The current shape window is also fixed to 101 bp. Different TFs may rely on shape information at different distances from the motif, so learning a TF-specific window or weighting scheme would be a natural extension.

Finally, both models in this study were classical classifiers. They are easy to interpret and fast to train, but their capacity is limited. A neural model that learns shape patterns jointly with sequence, or a kernel that incorporates shape similarity directly, could improve performance for shape-sensitive factors without relying on manually selected features.

Overall, the results support a simple view. Some transcription factors, such as CTCF, are sequence-dominated and can be predicted with basic models. Others, such as BHLHE40, slightly gain from structural signals that sequence alone does not capture. Future work should focus on datasets that cover more TFs and on models that can adjust to the different binding strategies used across the genome.

5 CONCLUSIONS

The experiments demonstrate that sequence and DNA shape do not contribute equally across transcription factors. CTCF is dominated by sequence composition, and both models reached perfect performance without using shape. This matches its strong consensus motif and the clear GC shift in the data. BHLHE40 relies on different signals. Shape features improved performance for both models, and the combined setting gave the most stable precision–recall behavior. This indicates that local DNA geometry captures aspects of binding that are not reflected by simple motif encoding.

The study also shows that classical models can detect these differences. Logistic regression was enough to separate the feature types for both TFs, while the SVM reduced the gap because of its nonlinear kernel. The runtime differences further suggest that linear models are effective when the feature space matches the factor's binding mechanism.

Future work should extend the analysis beyond chromosome 1, explore richer shape features, and evaluate more transcription factors with diverse binding modes. A broader set of TFs and more flexible models would help build a clearer picture of when structural information provides real predictive value.

REFERENCES

Hu, G., e. a., 2020, Systematic screening of CTCF binding partners identifies that BHLHE40 regulates CTCF genome-wide distribution and long-range chromatin interactions - PubMed.

J, W., Z. J, I. S, L. Xy, G. Mc, K. Bh, M. J, P. Bg, D. X, V. D, B. E, H. Jh, and W. Z, 2012, Factorbook.org: a Wiki-based database for transcription factor-binding data generated by the ENCODE consortium: PubMed.

Jd, S., R. Eh, and S. Mk, 2008, Phylogenetic and expression analysis of the basic helix-loop-helix transcription factor gene family: genomic approach to cellular differentiation: PubMed.

Rohs, R., S. M. West, A. Sosinsky, P. Liu, R. S. Mann, and B. Honig, 2009, The role of DNA shape in protein–DNA recognition: Nature, **461**, 1248–1253. (Publisher: Nature Publishing Group).

Tp, C., C. F, Z. T, Y. L, P. R, and R. R, 2016, DNAshapeR: an R/Bioconductor package for DNA shape prediction and feature encoding: PubMed.

X, S., Z. J, and C. C, 2022, CTCF and Its Partners: Shaper of 3D Genome during Development: PubMed.

Zhou, T., N. Shen, L. Yang, N. Abe, J. Horton, R. S. Mann, H. J. Bussemaker, R. Gordân, and R. Rohs, 2015, Quantitative modeling of transcription factor binding specificities using DNA shape: Proceedings of the National Academy of Sciences, **112**, 4654–4659. (Publisher: Proceedings of the National Academy of Sciences).