

Student Names: Logan Martel

Student IDs: 260558097, 260481930

TFBS ML Classification Research (Submitted for McGill COMP 680-Mining DNA)

Quantifying family-wise specificity of intramolecular flanking region flexibility and structural motif interactions as features for transcription factor binding site classification

Logan Martel and Frederick Parsons

April 24, 2017

Blanchette M, COMP680

Introduction

Background Motivation: In developing a comprehensive understanding of functional biological systems, the genomic & biomedical implications of insights to gene regulation are profound (Kitano, 2002), (Barabasi and Oltvai, 2004). Of particular interest, the role of transcription factors in gene regulatory networks is well-evidenced as a controlling mechanism for gene expression (Ptashne and Gann, 1997), (Jacob and Monod, 1961). Transcription factor interactions are characterized highly by their DNA binding-specificity (Godoy, et al., 2011), (Dror, et al., 2016). Consequently, effective prediction & classification of transcription factor binding sites (TFBS) has been a key goal in bioinformatics for decades (Dror, et al., 2016), (Wasserman and Sandelin, 2004).

Classically, TFBS prediction algorithms have relied heavily on exhaustive search of position-weight matrices modeling sequence binding motifs from ChIP-seq data (Hannenhalli, 2008). Recently, the high availability of ChIP-seq datasets has motivated investigation into supervised learning approaches for TFBS classification (Libbrecht and Noble, 2015). In this domain, newly developed software to infer DNA shape from sequence (Zhou, et al., 2013) has enabled contemporary research to demonstrate considerable success of DNA structural properties as features to train an ensemble machine learning classifier for prediction of *in vivo* transcription factor binding.

Very recently, further research into DNA shape readouts across TF's has demonstrated family-specific preferences in structural DNA motifs (Yang, et al., 2017). These results fit with previous observations that protein-DNA binding specificity has been shown to depend on two main factors: (i) contact forces between amino acids and base pairs either directly or through water molecules (*intermolecular interactions*) and (ii) the dynamic conformational changes in the DNA, such as DNA shape and flexibility (*intramolecular interactions*) (Gromiha, 2005). Prior research attempting to characterize the relative contribution of these interactions (Gromiha, 2005) had also identified family-specific binding affinities where one type of interaction is more important to a particular TF family in contrast to greater importance of the other type of interaction in another family.

Building on this notion of family-wise specificity of intramolecular interactions, there are other structural features to consider. From a mechanical perspective, DNA bending rigidity (or *persistence length*¹) is known to be a crucial physical aspect to understanding DNA–protein binding (Geggier and Vologodskii, 2010). Though DNA is typically regarded as a highly stiff polymer – with a *persistence length* around 150bp (Manning, 2006) – there is evidence of extreme bendability in sub-100bp strands (Vafabakhsh and Ha, 2012).

As early as 1989, there has been evidence as to the effect of sequence composition on DNA’s conformational flexibility (Sarai, et al., 1989). Several sequence-based scales derived from epigenetic modification data (DNaseI cleavage and nucleosome positioning preference) were proposed to measure DNA stiffness in terms of bendability towards the major/minor groove. Here, we choose to focus instead on a more-general measure expressed in terms of *Young’s modulus*² for trinucleotide parameters. By definition, *persistence length* is directly proportional to *Young’s modulus* since:

$$\text{Persistence length} = \frac{B_s}{k_B T},$$

where $k_B T$ is an energy scaling factor (product of the Boltzmann constant and temperature)

and B_s is the bending stiffness $B_s = EI = E \frac{\pi a^4}{4}$, where I is the area moment of inertia

and E is the *Young’s modulus* $E = \frac{\text{tensile stress}}{\text{extensional strain}} = \frac{FL_0}{\Delta L}$ typically measured in N/m^2

Proceeding with *Young’s modulus* as a measure of DNA *stiffness* (the inverse of *flexibility*) therefore gives a natural physical interpretation to intramolecular interactions in our feature set. Additionally, the relative contribution of DNA *stiffness* to protein–DNA binding specificity has been well studied in a few TF families for which TFBS data is available (Gromiha, 2005).

¹ Persistence length is defined as the average projection of the end-to-end vector on the tangent to the chain contour at a chain end in the limit of infinite chain length (<https://goldbook.iupac.org/html/P/P04515.html>)

² Young’s Modulus is a measure of *stiffness* (the inverse of *flexibility*) defined as the ratio of the stress (force per unit area) along an axis to the strain (ratio of deformation over initial length) along that axis in the range of stress in which Hooke’s law holds (<http://goldbook.iupac.org/html/M/M03966.html>).

Reflecting deeper on known intramolecular contributions to protein-DNA binding, it is important to recognize the interaction between DNA stiffness and geometry. Many standard DNAShape parameters (Rise, Twist, Tilt, and Roll) are known to have both a static component and a dynamic one (Gabrielian, et al., 1998) and it has been observed that the energetics of DNA flexibility also contribute to the structural propensity for DNA to form a particular geometry (Travers, 2004). Taken in tandem with DNA shape, these structural and mechanical properties complement each other as predictors for DNA-protein binding (Ortiz and de Pablo, 2011).

In this regard, it is no surprise that there is evidence for the conservation of structural features such as curvature, bendability, and nucleosome positioning preference in DNA promoter regions (Bansal, et al., 2014). Whereas promoter regions are typically at least 100bp in length (Sharan), previous supervised-learning research into DNA shape features has looked primarily at top motif hits of around 15bp (Mathelier, et al.). This key observation shall form the basis for our research.

Problem Statement: Given the evidence that DNA flexibility varies with sequence composition and is a key physical factor in DNA-protein binding, can we improve the predictive power of a DNA-shape trained TFBS classifier by incorporating an estimate of promoter-region flexibility derived from an extended high-scoring pssm (position-specific scoring matrix) sequence?

Research Hypothesis: Since DNA bendability can be parameterized via several trinucleotide-step scales (Brukner, et al., 1995) (Gromiha, 2000), we should see improvement in a classifier trained on extended flanking region trinucleotide-steps in addition to pssm-hit DNA shape in protein families where intramolecular interactions are known contributors to protein-DNA binding specificity.

Objectives: Developing on the DNAShapedTFBS project (Mathelier, et al.), we seek to:

1. Improve TFBS classification by incorporating an estimate of flanking region flexibility.
2. Investigate the impact of not-bound training examples on classification performance.
3. Explore the family-wise interactions and trends between DNA shape and flexibility features

Methodology

Tools: The bulk of our project was developed by extending the open-source DNAsHapedTFBS project, which comprises numerous python libraries (including Numpy, Scipy, Biopython, and scikit-learn, among others). Additionally, we also made extensive use of Bedtools (Quinlan and Hall, 2010) in transforming our dataset and a modified version of the BiasAway tool (Worsley Hunt, et al., 2014) to generate our not-bound background sequences from the provided GM12878 regulatory region. For computing Friedman's H-statistic on the relative strength of interaction effects between input features of an ensemble classifier (Friedman and Popescu, 2008) we use Ralph Haygood's sklearn-gbmi python package (v1.0.0 - 2017)³. Lastly, where possible, the Readout⁴ tool (Ahmad, et al., 2006) is used to quantify energy Z-scores for indirect protein-DNA affinity of several TF's with known protein-DNA bound complexes available via PDBe⁵ (Krissinel and Henrick, 2007)

Procedures:

Objective 1:

Our primary goal was to leverage and extend the DNAsHapedTFBS project in order to achieve improved TFBS classification protein families whose DNA binding specificity has been shown to depend heavily on intramolecular interactions. To this end, we chose to focus most on the ETS and Zinc-Finger families whose binding free energies have been positively correlated with DNA stiffness (Gromiha, 2005) Below, we will discuss the methods for constructing our feature vectors X_i (where X_i is the i - th feature vector investigated).

Assume, at this stage, that our bound and not-bound .bed & FASTA files per protein (4 files) are formatted to fit the project's requirements. As an additional requirement, note that the subset of our data that we were able to use was restricted to proteins for which there was a human position-specific scoring matrix (pssm) available on the public JASPAR CORE database (Mathelier, et al., 2015).

³ <https://pypi.python.org/pypi/sklearn-gbmi/1.0.0>

⁴ <http://www.abren.net/readout/>

⁵ <https://www.ebi.ac.uk/pdbe/node/1>

The basics of the DNAShapedTFBS algorithm are to train & test a Gradient Boosting ensemble machine learning classifier on the pssm score & DNA shape features (retrieved from a feature-annotated human genome) of a set of bound & not-bound sequences. These bound & not-bound sequences are each chosen as the top pssm hit for each of our input foreground & background strings⁶. First and second order DNA shape metrics for minor groove width (MGW), roll (Roll), propeller twist (ProT), and helix twist (HelT) are retrieved at each position of the pssm hit via hg19 bigWig files⁷ precomputed by DNAShape.

From here, the primary task was to construct input features to evaluate the flexibility of the flanking region from a pssm hit sequence. Firstly, whereas each of our foreground sequences was constructed precisely equal to the respective ChIP-seq peak length + 100 (ie. extended 50 nucleotides in both the 5' and 3' direction), the not-bound pssm hits would need to be extended dynamically. To do this, we assumed that each of our background sequences was greater than or equal to pssm motif size (~15) + 100 and then used an O(1) algorithm⁸ to check the remaining space on each side of a pssm hit to extend as close to 50bp on each end as possible. To efficiently evaluate the flexibility of these regions while scanning for trinucleotide steps, we leveraged a Trie⁹ data structure to take advantage of common prefixes while maintaining a running count of observed substrings per sequence. Finally, we evaluated DNA flexibility using these counts via 2 separate procedures:

a) For 32 trinucleotide-steps, add the observed count for each trinucleotide to that of its reverse complement and take this as a single feature. So, we have 32 new features to train our classifier.

⁶ Choosing the minimally-scored not-bound sequence consistently resulted in 100% prediction accuracy of the classifier so we instead chose the maximally-scored to ensure that we could observe a relative difference in results.

⁷ Downloaded from the author's public ftp server <ftp://rohslab.usc.edu/hg19/>

⁸ Note that the implementation is not 100% working as-intended due to some off-by-one edge cases working with the indices. To resolve these, we simply set our returned indices to the bounds of the sequence from which the hit is derived in case our algorithm would cause it to extend too far.

⁹ Note that the Trie data structure was chosen for future extensibility to scanning variable-sized words other than trinucleotide steps (there has been some research as to the impact of other sub-sequences on DNA flexibility)

b) Associate each of the 64 possible trinucleotides to its bendability parameter (p_w) given by a previous experiment¹⁰. Use this bendability parameter (p_w) to weight the observed counts of each trinucleotide step in a single new feature:

$$flex_eval(seq) = \frac{\sum_{i=1}^{length(seq)} p_{w_i} * (count(w_i) + count(reverseComplement(w_i)))}{length(seq)}$$

Clarifying the above, note that the scale for these bendability parameters is as follows:

Table 1. Structure based DNA stiffness (Young's modulus) scale for trinucleotides

Trinucleotide	E (10^8 N/m ²)	Trinucleotide	E (10^8 N/m ²)
AAA/TTT	4.80	CAG/CTG	2.40
AAC/GTT	3.90	CCA/TGG	3.25
AAG/CTT	1.91	CCC/GGG	6.07
AAT/ATT	2.96	CCG/CGG	2.40
ACA/TGT	4.70	CGA/TCG	2.82
ACC/GGT	1.57	CGC/GCG	3.33
ACG/CGT	7.09	CTA/TAG	4.75
ACT/AGT	3.63	CTC/GAG	4.03
AGA/TCT	4.03	GAA/TTC	2.70
AGC/GCT	4.58	GAC/GTC	7.83
AGG/CCT	4.34	GCA/TGC	3.75
ATA/TAT	2.36	GCC/GGC	3.16
ATC/GAT	1.83	GGA/TCC	3.69
ATG/CAT	3.19	GTA/TAC	2.19
CAA/TTG	2.53	TAA/TTA	2.72
CAC/CTG	3.36	TCA/TGA	2.97

Gromiha MM. Structure Based Sequence Dependent Stiffness Scale for Trinucleotides: A Direct Method. *Journal of Biological Physics*. 2000;26(1):43-50. doi:10.1023/A:1005250718139.

Thus, our bendability parameter p_w for a trinucleotide step w is equal to its *Young's modulus* E_w and we have: $flex_eval(seq) = \frac{\sum_{i=1}^{length(seq)} E_i}{length(seq)} = E$ where E is the average *Young's modulus* of our sequence – a measure strongly correlated to binding free energy.(Gromiha, 2005)

¹⁰ See https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3456187/pdf/10867_2004_Article_241559.pdf

From this point on, note that each of our experiments leveraged a feature space formed from exactly one of the following one-dimensional feature vectors X_i :

$$X_0 = [Shape_1, Shape_2]$$

where $Shape_1$ and $Shape_2$ are first and second order DNA shape metrics at the pssm-motif s.t.

$Shape_1 = [HelT_0, \dots, HelT_n, ProT_0, \dots, ProT_n, MGW_0, \dots, MGW_n, Roll_0, \dots, Roll_n]$, and

$Shape_2 = [HelT2_0, \dots, HelT2_n, ProT2_0, \dots, ProT2_n, MGW2_0, \dots, MGW2_n, Roll2_0, \dots, Roll2_n]$

$$X_1 = [X_0, count(AAA | TTT), count(AAC | GTT), \dots, count(TCA | TGA)]$$

where *trinuc_counts* is a vector of 32 trinucleotide counts (procedure (a) from above) s.t.

$trinuc_counts = [count(AAA | TTT), count(AAC | GTT), \dots, count(TCA | TGA)]$

$$X_2 = [X_0, flex_eval(seq)] \text{ where } flex_eval(seq) \text{ is the single feature from procedure (b).}$$

Henceforth, for convenience, we may alias the feature vectors of our set $X = \{X_0, X_1, X_2\}$ as

$X = \{DNA_Shape_Only, DNA_Shape_and_Counts, DNA_Shape_and_Flex_Eval\}$

Objective 2

Below, we will discuss the methods for constructing our not-bound “background” datasets B_i (where B_i is the i – *th* background type investigated). Note that for each background type B_i , there is one corresponding dataset of not-bound training examples per protein.

We begin by detailing a general procedure for generating a set of “probably not bound” sequences from the regulatory region of a particular cell type (denoted CELL) for each protein.

Our first task was to isolate the bound “foreground” datasets. The initial input for this was the *factorbookMotifPos.txt* table containing data for positions of all canonical motifs for ~160 TF’s across hg19. Splitting the input set of genomic coordinates by transcription factor, we obtained one .bed file per protein. In order to limit this to the regulatory regions relevant for a particular cell line, we leveraged ENCODE’s experimentally-obtained ChIP-seq data aggregated in *wgEncodeRegTfbsClusteredWithCellsV3.CELL.merged.bed* to perform a bedtools intersect with

left outer join between the factorbook binding sites for a particular TF and ENCODE's bound regulatory regions along a particular cell line.

For each .bed file, genomic coordinates of peaks were then extended 50bp along both the 5' and 3' direction using the Bedtools slop function. Per transcription factor p and cell line C , we denote this dataset as $\text{bound}_{\text{bed}}(p, C)$.

In seeking a set of negative examples where a specific TF p does not bind, the first intuitive step was to remove the known binding sites $\text{bound}_{\text{bed}}(p)$ from consideration. For each protein p , we transformed *wgEncodeRegTfbsClusteredWithCellsV3.CELL.merged.bed* (denoted CELL_{bed}) to a "reduced" .bed file (denoted $*\text{CELL}_{\text{bed}}(p)$) which removes the overlap of $\text{bound}_{\text{bed}}(p, C)$ from CELL_{bed} using the Bedtools *subtract* function as follows:

$$*\text{CELL}_{\text{bed}}(p) = \text{CELL}_{\text{bed}} - \text{bound}_{\text{bed}}(p, C)$$

Supplementary FASTA files $\text{bound}_{\text{fasta}}(p, C)$ and $*\text{CELL}_{\text{fasta}}(p)$ were then derived (converted to sequences on the sense (+) strand of the human genome - assembly hg19) from $\text{bound}_{\text{bed}}(p, C)$ and $*\text{CELL}_{\text{bed}}$ respectively via Bedtools *getfasta* function.

Drawing from $*\text{CELL}_{\text{fasta}}(p)$, the next step was to generate a set of not-bound sequences for each transcription factor p . Ideally, we would match the number of background sequences generated to the foreground $\text{bound}_{\text{fasta}}(p, C)$ and leverage some prior knowledge in our choice.

With these tenets in mind, 5 different not-bound data sets (B_i) were generated per protein p .

$$B_0 : \%GC \text{ Matched from the regulatory region of GM12878}$$

Here, we chose to build off a result that found success by matching the %GC content of foreground & background sequences (Worsley Hunt, et al., 2014). Taking as input the foreground ($\text{bound}_{\text{fasta}}(p, \text{GM12878})$) and sequences to match against ($*\text{GM12878}_{\text{fasta}}(p)$), the tool BiasAway was executed to extract a subset of sequences with matched %GC, sequence length, and total sequence count from $*\text{GM12878}_{\text{fasta}}(p)$.

B_1 : %GC Matched from the regulatory region of K562

Same as B_0 except using ($\text{bound}_{\text{fasta}}(\text{p}, \text{K562})$) and ($*\text{K562}_{\text{fasta}}(\text{p})$) in lieu of GM12878.

B_2 : GM12878 foreground %GC matched against random sequences

Same as B_0 except using ($\text{bound}_{\text{fasta}}(\text{p}, \text{GM12878})$) as the foreground and %GC matching against random hg19 sequences.

B_3 : Di-nucleotide shuffled GM12878 background

The BiasAway tool was re-executed per $*\text{GM12878}_{\text{fasta}}(\text{p})$ to generate dinucleotide shuffled sequences matching the bp-length and total sequence count for each ($\text{bound}_{\text{fasta}}(\text{p}, \text{GM12878})$).

B_4 : Random %GC distribution from $*\text{GM12878}_{\text{fasta}}(\text{p})$

In this case, we generated background sequences extracted from $*\text{GM12878}_{\text{fasta}}(\text{p})$ by repeatedly choosing half from a random progression of %GC bins until our sequence count matched the foreground.

In all cases above, if the foreground count was still higher, we trimmed ($\text{bound}_{\text{fasta}}(\text{p}, \text{C})$) to the size of the corresponding background. Lastly, some final data manipulation via custom scripts curated our background files to one .bed & one .fasta per protein per background type: $\text{notBound}_{\text{bed}}(\text{p}, B_i)$ and $\text{notBound}_{\text{fasta}}(\text{p}, B_i)$.

Objective 3

Evidently, the setup described above gives many combinations of experiments to investigate. Each experiment $\text{Expm}(X, B, \text{p})$ ¹¹ is a combination of 3 factors: the feature vector (X), background type (B), and transcription factor (p). We seek to group these experiments in a meaningful way to extract information about TF family-wise trends and protein-specific intramolecular interactions.

¹¹ Each background type is of course mapped to the foreground of a particular protein and cell line regulatory region

First, note that for each experiment $Exp_m(X, B, p)$, we produce the following output: (i) 5-fold aggregated PRC & ROC curves, (ii) reusable classifier (fitted to entire input data), (iii) feature importance ranking on fitted classifier, (iv) feature vectors and predicted binding probability for each input data instance, (v) Friedman's H statistic¹² to assess the relative strength of interaction effects between each pair of classification features. We proceed by discussing our intended setup for analyzing these outputs.

Due to severe limitations on time, data, and computational resources, we elected to focus primarily on backgrounds B_0 and B_3 as well as two major families of transcription factor: ETS-family and Zinc Fingers (testing 3 proteins per family). These factors are consistent across our analyses but other independent proteins and families (e.g. leucine zippers) show up intermittently.

To evaluate the performance of our classifier, we take feature $X_0 = DNA_Shape_Only$ as the control and aim for an improvement using flexibility feature vectors X_1 and X_2 . To compare the predictive power of the two flexibility experiments vs. shape only per TF, we computed a record per protein for the difference in area under the 5-fold averaged PRC curves (AUPRC's) and computed a p-value for the significance of the difference between the areas via an online tool¹³ implementing a one-tailed statistical test for comparing ROC curves (Lowry, 2016). This technique was leveraged for 9 main proteins (3 families) on backgrounds B_0 and B_3 .

To investigate the relative specificity of family-wise intramolecular interactions, we analyzed H-statistics averaged from all DNA shape – flexibility interactions above a varying threshold for several Zinc Finger and ETS-family proteins. Where possible, we compared these statistics to energy Z-scores from experimental readouts¹⁴ (Ahmad, et al., 2006) on the relative contribution of

¹² Described in <https://arxiv.org/pdf/0811.1679.pdf>

¹³ See http://vassarstats.net/roc_comp.html

¹⁴ See <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkl104> and <http://14.139.186.108/jspui/bitstream/123456789/25603/1/21.pdf><http://14.139.186.108/jspui/bitstream/123456789/25603/1/21.pdf>

direct and indirect – contact binding in the DNA-protein complexes formed by these transcription factors.

Delving deeper into proteins of particular interest, feature importance measures were computed by the Python scikit-learn module (Pedregosa, et al., 2011) and analyzed across families.

Results

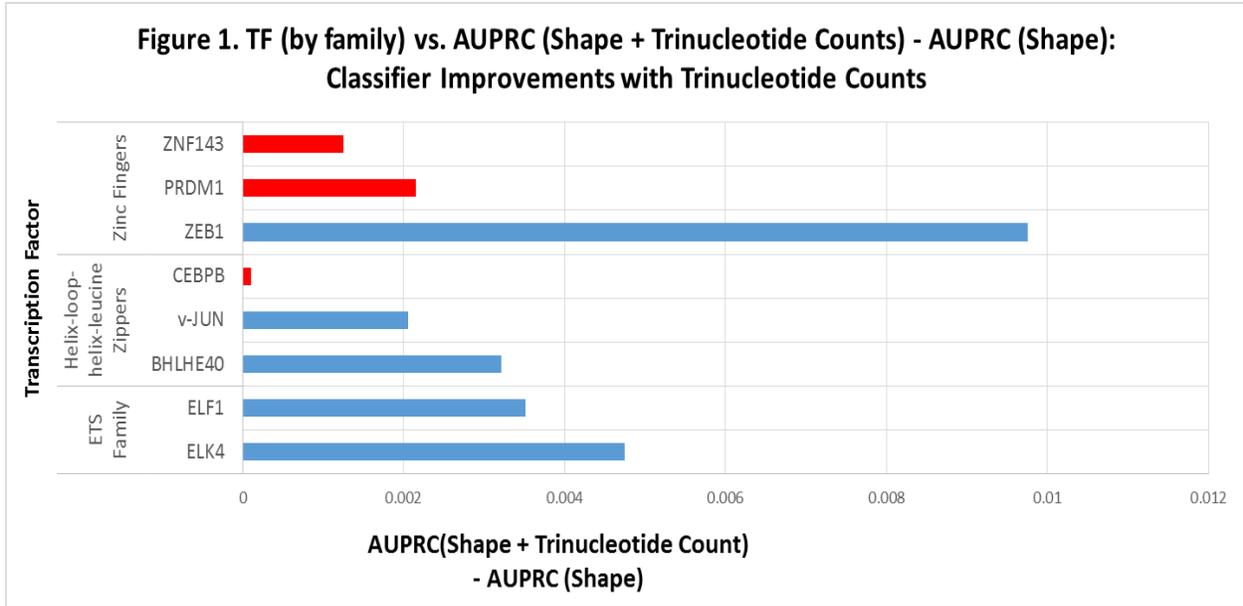
This section outlines the key findings in varying X, B and p across experiments $Expm(X, B, p)$.

We begin by discussing the overall performance of our flexibility feature vectors compared to *DNA_Shape_Only* across backgrounds B_0 and B_3 .

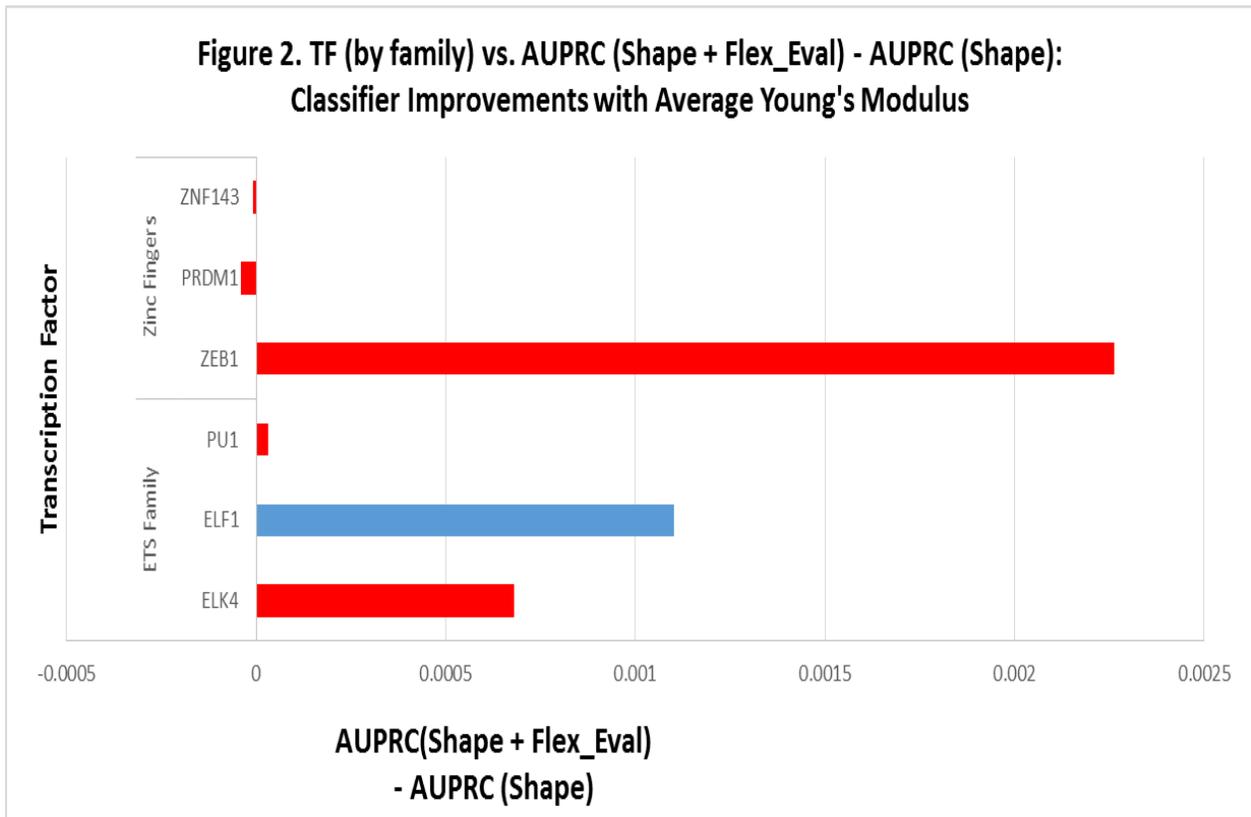
In the case of the %GC Matched background B_0 , the differences in AUPRC's across all proteins performed generally worse when flexibility measures were included compared to the *DNA_Shape_Only* feature space

In contrast, differences in AUPRC's across all proteins with dinucleotide-shuffled background B_3 generally performed better than *DNA_Shape_Only* with other variables held constant. To visualize this improvement, we plotted the differences in AUPRC per protein for each of *DNA_Shape_and_Counts* and *DNA_Shape_and_Flex_Eval* vs. *DNA_Shape_Only* .

Statistically significant results ($(p_{\text{one-tailed}} < 0.05)$.) are denoted as blue bars on figures 1 & 2, while red ones denote statistically insignificant differences. All p-values are listed per protein for these in our appendixes.



Above, our most improved (and statistically significant) classification was for the protein ZEB1.



Above, our most improved (albeit not statistically significant) classification was again for ZEB1.

Though the evidence is far from conclusive, ETS-family proteins¹⁵ typically showed greater improvements on both flexibility-trained classifiers compared to Zinc Fingers (with the notable exception of our most-improved protein ZEB1).

To further characterize these results, we summarize our attempt to quantify the relative specificity of direct vs. indirect binding preference per transcription factor. In the table below, lower (ie. more negative) Z-scores indicate stronger energetic preferences for a category of binding interaction.

Table 2. Direct and Indirect Readout Energy Z-scores for Intra- vs. Inter- Molecular Protein-DNA Binding Affinity by Transcription Factor

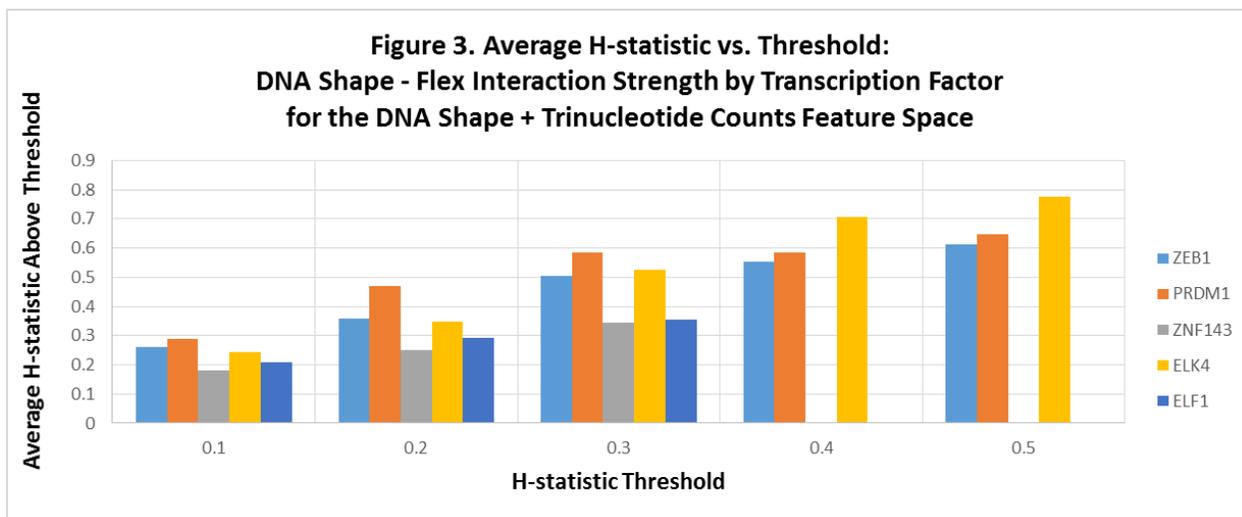
TF Family	Protein	Source Alias	Intramolecular Readout Z-score	Intermolecular Readout Z-score	PDB Code for Analysis	Source
Zinc_Finger	ZEB1	NA	-2.37	NA	2e19	http://www.abren.net/readout/
Zinc_Finger	PRDM1	NA	-2.37	NA	3dal	http://www.abren.net/readout/
Zinc_Finger	ZNF143	NA	NA	NA	NA	NA
ETS_Family	ELK4	NA	-1.04	-2.41	1bc8	http://www.abren.net/readout/
ETS_Family	ELK4	SAP-1	-3.1	-0.5	1hbx	Gromiha, et al., 2005
ETS_Family	ELF1	NA	NA	NA	NA	NA
ETS_Family	PU1	PU.1/SPI1	-2.7	-1.1	1pue	Gromiha, et al., 2005

Evidently, the above data is regrettably incomplete and indecisive - at least one protein (ELK4) had contradicting results depending on the source and/or pdb structure under consideration.

To further explore the relative specificity of intramolecular interactions in protein-DNA binding, we present averaged H-statistics for interaction strength of DNA shape – flexibility feature pairs of the *DNA_Shape_and_Counts* and *DNA_Shape_and_Flex_Eval* experiments by protein at varying thresholds. Here, the Friedman H-statistic¹⁶ is a measure (ranging from 0 to 1) of interaction strength between features fitted to a gradient boosting ensemble classifier where a value of 0 indicates no interaction and larger values indicate stronger interaction effects.

¹⁵ Note: PU1 was excluded from the *DNA_Shape_and_Counts* execution due to RAM limitations causing crashes during execution. Other proteins excluded intermittently were due to similar time & space resource limitations.

¹⁶ See <https://arxiv.org/pdf/0811.1679.pdf>



Despite some variance within families, the contribution of interaction effects in ZEB1 (our TF with the best classification improvement) is consistent. We investigate ZEB1 further below.

Table 3. Feature Importance for TF ZEB1 in TFBS Classification with DNA shape +Trinucleotide Count Features (Top 20).

Protein	Feature_Name	Importance_Value
ZEB1	MGW - 5	0.169208857
ZEB1	ProT - 4	0.126592317
ZEB1	HelT2 - 6	0.058187697
ZEB1	AAC/GTT	0.031465782
ZEB1	Roll2 - 2	0.028179742
ZEB1	HelT2 - 5	0.026213114
ZEB1	Roll2 - 5	0.02522105
ZEB1	MGW - 3	0.024257215
ZEB1	CTC/GAG	0.019570234
ZEB1	Roll2 - 3	0.018170797
ZEB1	ProT - 2	0.018107358
ZEB1	Roll2 - 0	0.017741068
ZEB1	CGC/GCG	0.017461104
ZEB1	Roll2 - 6	0.016883172
ZEB1	GAA/TTC	0.016618564
ZEB1	AGG/CCT	0.016296746
ZEB1	GAC/GTC	0.016023872
ZEB1	CCC/GGG	0.015882917
ZEB1	ACC/GGT	0.015075608

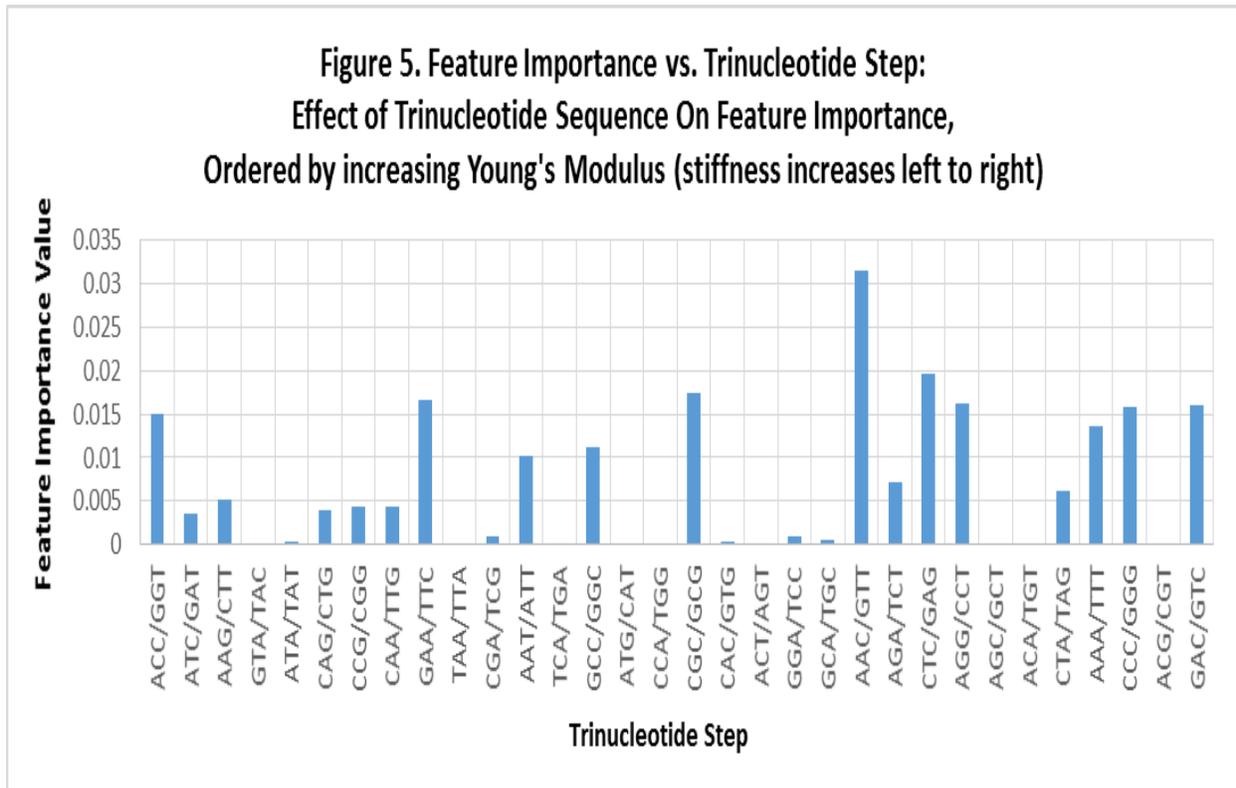


Figure 5 (above) illustrates the feature importance of each trinucleotide-step in classifying TFBS for ZEB1 (reverse complement is equivalent per step). The steps are listed in increasing order of bendability parameters (ie. ordered most to least flexible from left to right on x-axis).

Above, mass is very concentrated towards stiffer trinucleotide steps (with occasional peaks at the low- and mid- stiffness regions). Note also that – of the 8 trinucleotide steps ranked in the top 20 features, the stiffest sequence (CCC/GGG) was also a feature in the 6-th most significant interaction pair (HelT at motif position 2, count(CCC | GGG)) with an H-statistic of 0.53879 .

Finally, to further validate the improvement in ZEB1 classification, we reserved backgrounds B_1 and B_2 as held-out test sets and demonstrated improvements in AUPRC with flexibility features.

Figure 6. PRC Curves of ZEB1 classifier trained on GM12878 Sequences using DNA Shape + Trinucleotide Count vs. DNA shape only when applied to classification on held-out genomic sequences of K562

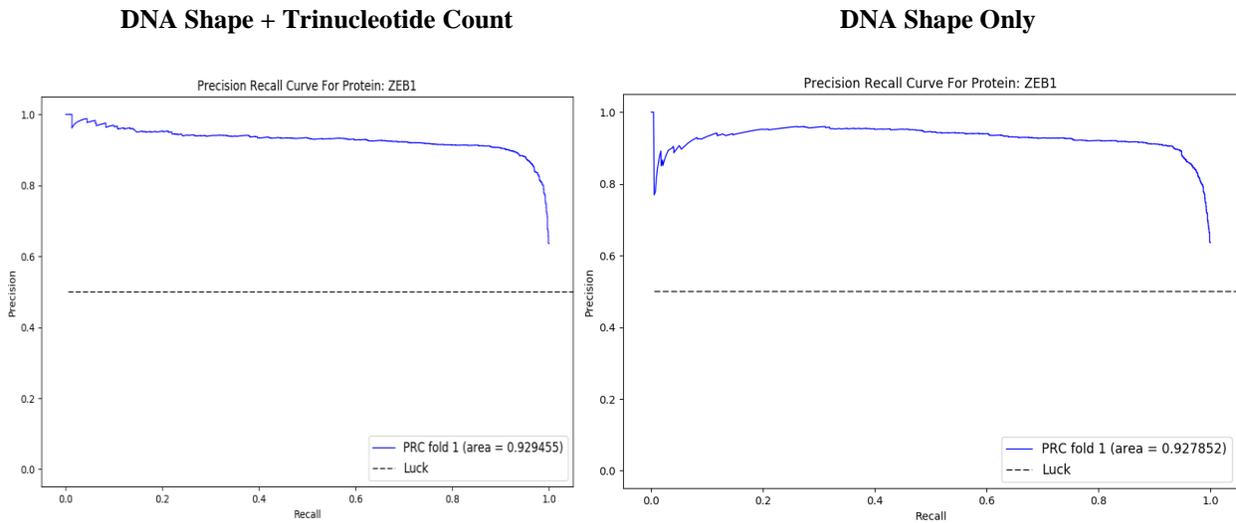
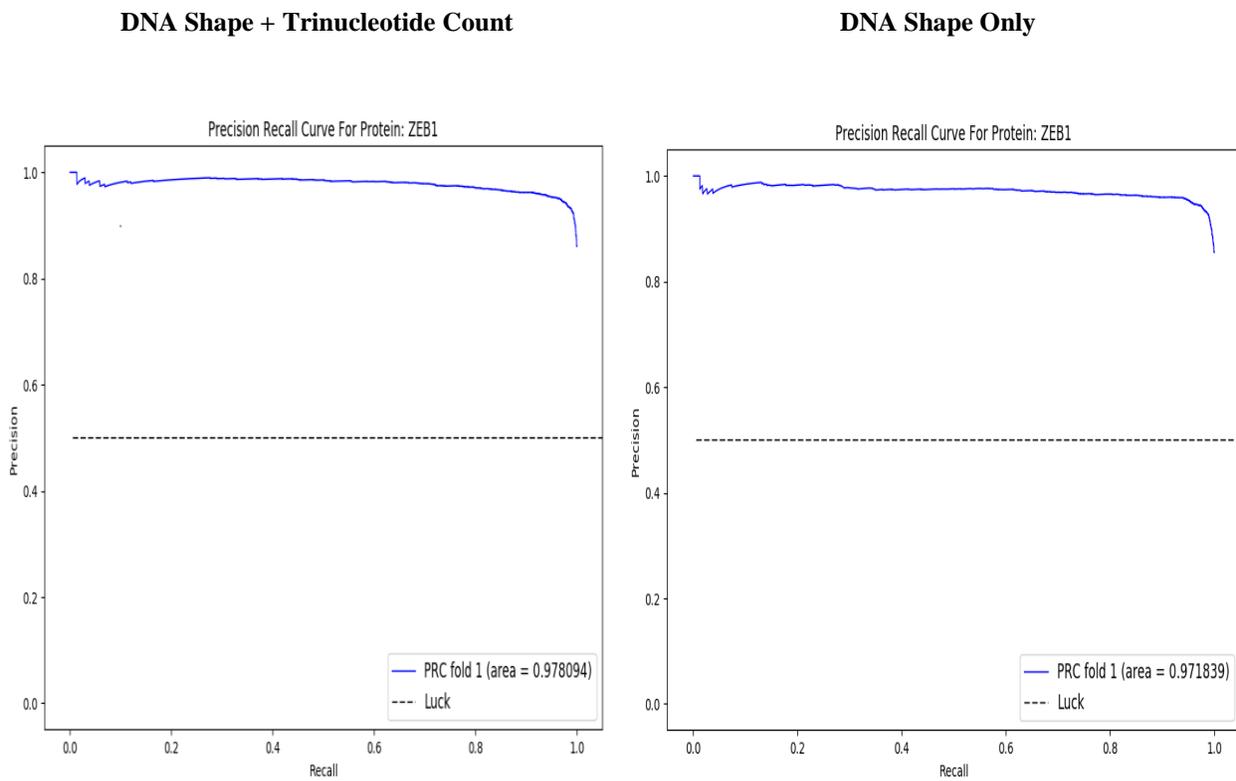


Figure 7. PRC Curves of ZEB1 classifier trained on GM12878 Sequences using DNA Shape + Trinucleotide Count vs. DNA shape only when applied to classification on held-out random hg19 sequences



Discussion and Future Work

The above results support our claim that flexibility information about the region surrounding transcription factor binding sites improves the performance of a machine learning classifier in discriminating between bound and not-bound sequences for some transcription factors.

Two approaches supported this claim, each one augmenting the feature space of DNA shape metrics. Of our two approaches, results suggest that the one directly using counts of 32 trinucleotide steps effected a greater improvement to our TFBS classification than our attempt to encapsulate this information via a single weighted evaluation function (*flex_eval*).

Interpreting these results, our focus is primarily to compare relative contributions of the flexibility features between ETS-family and Zinc Finger transcription factors. Previously, the literature has indicated generally stronger intramolecular contributions to binding affinity in ETS-family proteins compared to Zinc Fingers (Gromiha, 2005). In some regard, our results fit these observations in that the ETS-family proteins PU1, ELF1, and ELK4 demonstrated greater (more often significant) classifier improvements in feature spaces that included flexibility metrics compared to the Zinc Finger proteins ZNF143 and PRDM1. Among other indicators, ELF1 was in fact the only protein which experienced a statistically significant improvement in the *flex_eval* feature space and ELK4 exhibited the strongest DNA Shape – Flex interaction effect in the *trinucleotide_counts* feature space.

At odds with this argument, the Zinc Finger transcription factor ZEB1 is juxtaposed against prior expectations regarding these two families. Across the board, our results consistently favored ZEB1 as a strong example of classifier improvement with evidence of intramolecular feature interactions in the space of flexibility measures. ZEB1 demonstrated the greatest raw increase in AUPRC value across *flex_eval* and *trinucleotide_counts* experiments. Its fitted classifier also evidenced several non-negligible DNA Shape – Flex interaction effects (measured by the H-statistics) and a distribution of trinucleotide-step feature importance values with considerable mass concentrated towards stiffer trinucleotide steps (as measured by *Young's Modulus*). ZEB1's classification improvements in flexibility feature spaces was even shown to out-perform DNA

shape alone when cross-validated against two held-out datasets (random hg19 sequences and K562 cell line regulatory regions).

Attempting to explain this outcome, it is essential to recognize that – relative intramolecular binding affinities notwithstanding – the binding affinities of both ETS-family and Zinc finger proteins have been strongly correlated to DNA stiffness (coefficient in the range of 0.65–0.97) by previous research (Gromiha, 2005). For this reason, it is not entirely surprising that inter-family differences in the magnitude of indirect binding energetics might affect our results even in the ideal case of constant within-family preferences for direct or indirect binding. To adjust for this factor, we had intended to map our results to per-protein energy Z-scores for direct vs. indirect binding affinities obtained from DNA-protein crystallization data (see **Table 2**). However, inconsistencies in available data (e.g. ELK4’s contradictory indirect binding Z-scores less than or greater than ZEB1 depending on the source) make it difficult to relate classifier improvements to relative binding specificities.

Fortunately, some trends were consistent across our experimental setups. On all background-protein combinations, the *trinucleotide_counts* feature space has outperformed *flex_eval*. In stark contrast to an amalgamated evaluation function, using the trinucleotide counts directly as features permits the classifier to autonomously learn the contributions of trinucleotide steps to transcription factor binding (without imposing an arbitrary weighting on significance). In our most-highly improved TFBS classification (for ZEB1), 8 out of 20 of our most highly ranked features (by importance) were counts of trinucleotide steps. Most interestingly, 6/8 of these features had a *Young’s Modulus* in the top 50% for DNA stiffness (of the remaining 2, one was ACC/GGT – the most flexible trinucleotide step). Complementarily, our interaction test also revealed 26 trinucleotide step – DNA shape feature pairs with an H-statistic greater than 0.1 for interaction strength (3 such pairs greater than 0.5) in ZEB1.

From the above observations, our hypothesis on promoter-region flexibility as an informative complement to pssm-hit DNA shape is supported in that trinucleotide counts as features seemed to improve performance most when linked to DNA flexibility. We see this especially because of the relatively high feature importance of several stiff trinucleotide steps (in addition to several feature importance peaks at otherwise-sparse regions – such as the CGC/GCG step). Intuitively,

it makes sense that features with stiff bendability parameters are the most informative since high or low counts for these parameters approximate whether the bending of DNA around a protein can stabilize the bound complex in DNA-protein docking (van der Vliet and Verrijzer, 1993). A rigid and conformationally valid DNA structure would favor binding by reducing entropic costs.(Harteis and Schneider, 2014). Rigid DNA will also often help stabilize the approaching protein through hydrogen-bonding and Van der Waals interactions between the two (Luscombe, et al., 2001).

From a protein-DNA docking perspective, this physical notion of DNA flexibility has been experimentally shown to greatly improve prediction of binding sites (van Dijk, et al., 2006). The issue with using only static structure of DNA (measured chiefly by DNA shape features) is that it weakly captures the dynamic conformational properties of the supramolecular interactions. The bendability of DNA is a key factor in sterically permitting a protein to enter the binding region and remain there during binding. Consequently, promoter region flexibility certainly contributes to facilitating promoter interactions as the TF approaches the binding site. Moreover, DNA in fact often shows a large conformational change upon binding to a protein (van Dijk, et al., 2006). This can greatly alter the shape of the interaction surface and influence the potential for binding. These conformational fluctuations may lead to favorable changes in enthalpy but come with an inherent entropic cost. (Harteis and Schneider, 2014). Weighing these entropic & enthalpic factors across the flanking regions is near-intractable for supramolecular analysis but appears to have been successfully inferred from training examples by our classifier.

Bearing this physical justification in mind, we propose that proteins with less-improved classification probably do not depend as much on the short-range or long-range DNA-protein stabilization intrinsic to DNA flexibility. The fact that certain proteins benefited more from trinucleotide counts than others has been linked to previous evidence on family-specific roles of inter- and intra- molecular interactions and suggests that TFBS classification with structural properties should be catered more highly to the protein at hand as dictated by its size and structure – with an emphasis on family-wise binding affinities.

aLimitations: Unfortunately, due to constraints on time & computational resources, there were severe limits on what we were able to investigate. The size of our dataset was restricted by the PSSM's available on JASPAR, well-characterized TF families, our ability to tweak input data was discouraged by the extensive logistics involved in transforming it (parsing, removing characters, etc.), and the number/size of proteins we were able to test was limited by available computation power. Most notably, our intended analysis was bounded by sparse data on direct- and indirect- binding energy readouts (even for the canonical ETS and Zinc Finger families).

Future Work & Improvements: Our foremost suggestion for future research would be to develop on techniques¹⁷ for analyzing the relative contributions of intra- and inter-molecular interactions in protein-DNA binding across TF families. A well-defined procedure for comparing these aspects of binding-specificity would be a major step towards catering classifiers to family-specific protein structural properties.

Given sufficient resources, we would have liked to run even more experiments to observe the effect of minor changes to our algorithm. For example, what is the effect of combining non-structural features (e.g. pssm-score, string encoding, etc.) in tandem with the structural features of our input data? What could we discover if we executed feature & model selection to choose the best setup for our machine-learning classifier?

With regards to structural properties of the promoter region, there are many others worth investigating. As noted in previous research (Bansal, et al., 2014), curvature and nucleosome positioning preference would be promising complements to DNA bendability. Moving further in this direction, there is room in the machine-learning space to explore protein structural features (e.g. training classifiers on protein families) in future experiments to predict DNA-protein binding. Finally, it may be a worthwhile endeavor to investigate larger regions surrounding ChIP-seq peaks. Stabilizing effects from distant regions may help lower the overall Gibbs free energy of the bound protein-DNA complex, thermodynamically favoring a bound state between the two (Zhou and Gilson, 2009).

¹⁷ In the vein of tools such as <http://www.abren.net/readout/>

References (Bibliography)

- Ahmad, S., *et al.* ReadOut: structure-based calculation of direct and indirect readout energies and specificities for protein–DNA recognition. *Nucleic acids research* 2006;34(Web Server issue):W124-W127.
- Ahmad, S., *et al.* ReadOut: structure-based calculation of direct and indirect readout energies and specificities for protein–DNA recognition. *Nucleic acids research* 2006;34(suppl_2):W124-W127.
- Bansal, M., Kumar, A. and Yella, V.R. Role of DNA sequence based structural features of promoters in transcription initiation and gene expression. *Current Opinion in Structural Biology* 2014;25:77-85.
- Barabasi, A.-L. and Oltvai, Z.N. Network biology: understanding the cell's functional organization. *Nat Rev Genet* 2004;5(2):101-113.
- Brukner, I., *et al.* Sequence-dependent bending propensity of DNA as revealed by DNase I: parameters for trinucleotides. *The EMBO Journal* 1995;14(8):1812-1818.
- Dror, I., Rohs, R. and Mandel-Gutfreund, Y. How motif environment influences transcription factor search dynamics: Finding a needle in a haystack. *BioEssays* 2016;38(7):605-612.
- Friedman, J.H. and Popescu, B.E. Predictive learning via rule ensembles. 2008:916-954.
- Gabrielian, A., *et al.* Prediction of bendability and curvature in genomic DNA. In, *Structure, Motion, Interaction and Expression of Biological Macromolecules, Volume 1: Proceedings of the Tenth Conversation*. Adenine press; 1998.
- Geggier, S. and Vologodskii, A. Sequence dependence of DNA bending rigidity. *Proceedings of the National Academy of Sciences* 2010;107(35):15421-15426.
- Godoy, M., *et al.* Improved protein-binding microarrays for the identification of DNA-binding specificities of transcription factors. *The Plant journal : for cell and molecular biology* 2011;66(4):700-711.
- Gromiha, M.M. Structure Based Sequence Dependent Stiffness Scale for Trinucleotides: A Direct Method. *Journal of Biological Physics* 2000;26(1):43-50.
- Gromiha, M.M. Influence of DNA stiffness in protein–DNA recognition. *Journal of Biotechnology* 2005;117(2):137-145.
- Hannenhalli, S. Eukaryotic transcription factor binding sites—modeling and integrative search methods. *Bioinformatics* 2008;24(11):1325-1331.
- Harteis, S. and Schneider, S. Making the Bend: DNA Tertiary Structure and Protein-DNA Interactions. *International Journal of Molecular Sciences* 2014;15(7):12335-12363.
- Jacob, F. and Monod, J. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology* 1961;3(3):318-356.
- Kitano, H. Systems Biology: A Brief Overview. *Science* 2002;295(5560):1662-1664.
- Krissinel, E. and Henrick, K. Inference of Macromolecular Assemblies from Crystalline State. *Journal of Molecular Biology* 2007;372(3):774-797.
- Libbrecht, M.W. and Noble, W.S. Machine learning applications in genetics and genomics. *Nat Rev Genet* 2015;16(6):321-332.
- Lowry, R. In.; 2016.
- Luscombe, N.M., Laskowski, R.A. and Thornton, J.M. Amino acid–base interactions: a three-dimensional analysis of protein–DNA interactions at an atomic level. *Nucleic acids research* 2001;29(13):2860-2874.

- Mathelier, A., *et al.* JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic acids research* 2015.
- Mathelier, A., *et al.* DNA Shape Features Improve Transcription Factor Binding Site Predictions In Vivo. *Cell Systems*;3(3):278-286.e274.
- Ortiz, V. and de Pablo, J.J. Molecular Origins of DNA Flexibility: Sequence Effects on Conformational and Mechanical Properties. *Physical review letters* 2011;106(23):238107-238107.
- Pedregosa, F., *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* 2011;12:2825-2830.
- Ptashne, M. and Gann, A. Transcriptional activation by recruitment. *Nature* 1997;386(6625):569-577.
- Quinlan, A.R. and Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010;26(6):841-842.
- Sarai, A., *et al.* Sequence dependence of DNA conformational flexibility. *Biochemistry* 1989;28(19):7842-7849.
- Sharan, R. Analysis of Biological Networks: Transcriptional Networks - Promoter Sequence Analysis. In.
- van der Vliet, P.C. and Verrijzer, C.P. Bending of DNA by transcription factors. *BioEssays* 1993;15(1):25-32.
- van Dijk, M., *et al.* Information-driven protein–DNA docking using HADDOCK: it is a matter of flexibility. *Nucleic acids research* 2006;34(11):3317-3325.
- Wasserman, W.W. and Sandelin, A. Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genet* 2004;5(4):276-287.
- Worsley Hunt, R., *et al.* Improving analysis of transcription factor binding sites within ChIP-Seq data based on topological motif enrichment. *BMC Genomics* 2014;15(1):472.
- Yang, L., *et al.* Transcription factor family-specific DNA shape readout revealed by quantitative specificity models. *Molecular Systems Biology* 2017;13(2):910.
- Zhou, H.-X. and Gilson, M.K. Theory of Free Energy and Entropy in Noncovalent Binding. *Chemical Reviews* 2009;109(9):4092-4107.
- Zhou, T., *et al.* DNashape: a method for the high-throughput prediction of DNA structural features on a genomic scale. *Nucleic acids research* 2013;41(Web Server issue):W56-62.