Predicting DNA Bendability from Sequences

Samin Rahman Khan

Sadman Sakib Ononnyo

Undergraduate Students

CSE, BUET

Supervised by

Md. Abul Hassan Samee, Ph.D.

Assistant Professor

Department of Molecular Physiology and Biophysics

Baylor College of Medicine

Dr. M. Sohel Rahman

Professor

Depart of Computer Science and Engineering

Bangladesh University of Engineering and Technology

Outline

- DNA Bendability and Loop-Seq
- Problem at hand
  - A better predictive model
    - Current NN Architectures and Results
  - Newer Interpretations
- Current models and Results
- Correlation of DNA binding sites, breakage sites with their bendability across species
- References

DNA Bendability and Loop-Seq[1]

Article

Measuring DNA mechanics on the genome scale

https://doi.org/10.1038/s41586-020-03052-3

Received: 11 April 2020

Accepted: 21 October 2020

Published online: 16 December 2020

Check for updates

Aakash Basu[1,2], Dmitriy G. Bobrovnikov[1], Zan Qureshi[3], Tunc Kayikcioglu[2,4], Thuy T. M. Ngo[2,4], Anand Ranjan[5], Sebastian Eustermann[6,7], Basilio Cieza[3], Michael T. Morgan[1], Miroslav Hejna[2,8], H. Tomas Rube[2,8], Karl-Peter Hopfner[6,7], Cynthia Wolberger[1], Jun S. Song[2,8,9] & Taekjip Ha[1,2,3,10,11]✉

Mechanical deformations of DNA such as bending are ubiquitous and have been implicated in diverse cellular functions[1]. However, the lack of high-throughput tools to measure the mechanical properties of DNA has limited our understanding of how DNA mechanics influence chromatin transactions across the genome. Here we

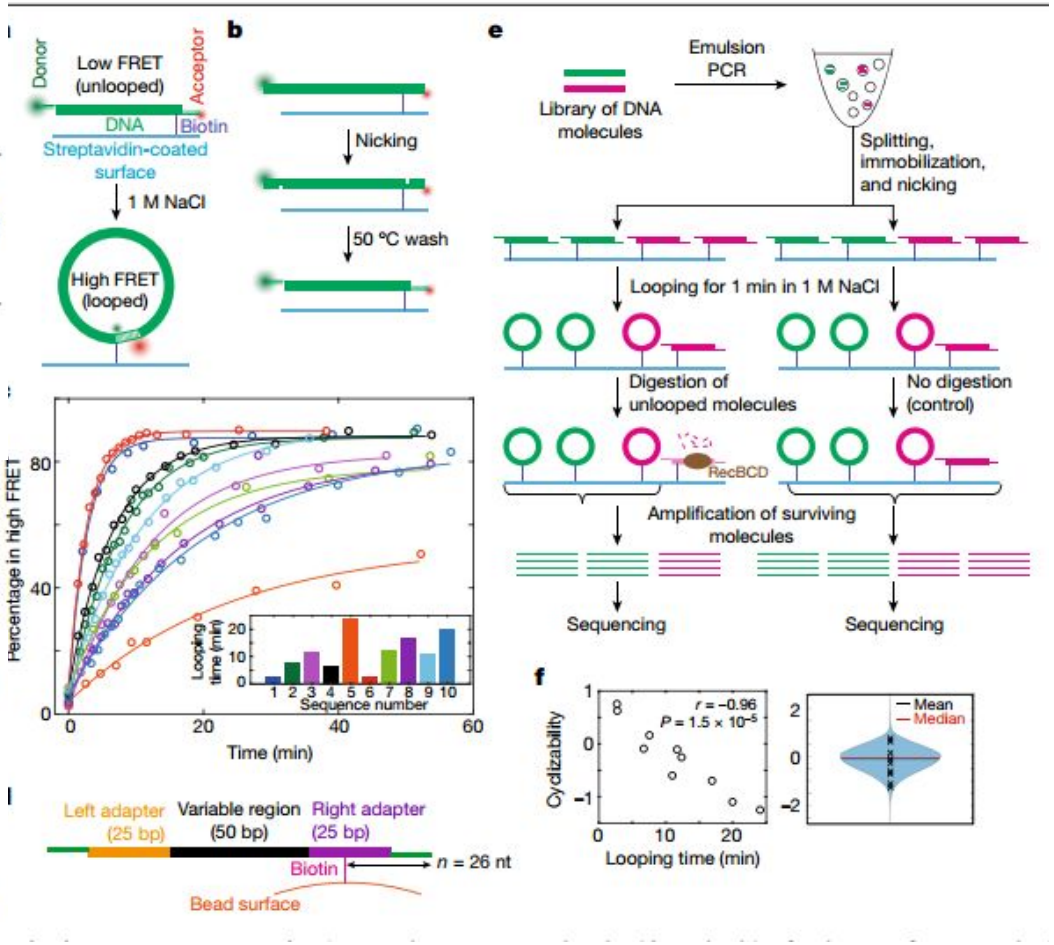


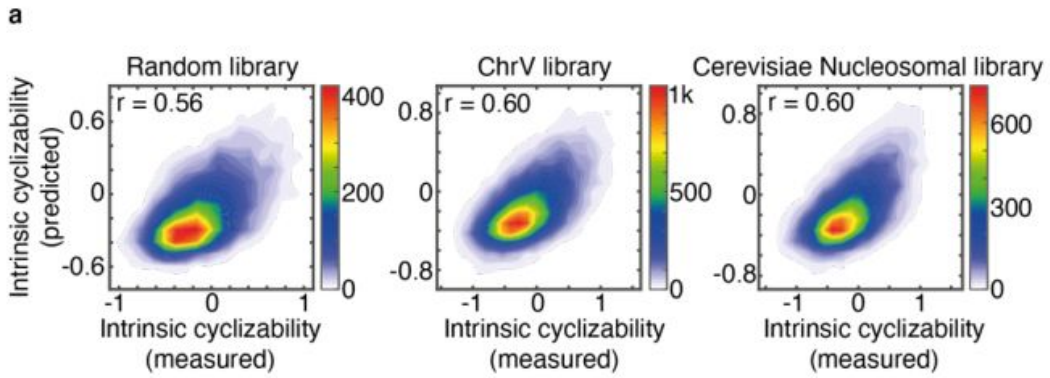
Fig. 1 | A high-throughput method to measure DNA mechanics. a, Schematic of the single-molecule looping assay. **b**, In situ nicking of 120-bp duplex DNA 10 nucleotides from either end, followed by washing with buffer at 50 °C, results in the formation of 100-bp duplex molecules flanked by 10-nucleotide single-stranded overhangs (Supplementary Note 1). **c**, Percentage of DNA molecules in the high-FRET (looped) state as a function of time after adding high-salt buffer, for ten DNA sequences (Supplementary Note 1). Inset, looping times (time constants of exponential decay fits (solid lines)). **d**, Schematic of a typical DNA fragment in a library just before looping. *n* denotes the distance in nucleotides (nt) of the biotin tether from the end of the molecule. **e**, Schematic of loop-seq performed on a hypothetical library comprising only two sequences: green and pink. The library is amplified, immobilized on beads via biotin–streptavidin interactions and nicked in situ to generate loopable molecules (**d**, Methods). After looping for 1 min in high-salt buffer followed by the digestion of unlooped molecules and amplification of surviving molecules, the relative populations of green and pink in the digested fraction (left) are two-thirds and one-third, respectively. In the control fraction (right), the corresponding values are half and half. The cyclizability of green and pink are thus: $\log_e\left(\frac{2/3}{1/2}\right)$ and $\log_e\left(\frac{1/3}{1/2}\right)$, respectively. **f**, Left, cyclizabilities of two sequences (listed in Supplementary Note 1) that were part of the 'cerevisiae nucleosomal library' (see Supplementary Note 4) versus their looping times obtained via smFRET (**c**). 95% confidence interval (CI) = −0.99, −0.81. Pearson's *r* value is shown. *P* value determined by two-sided *t*-test. Right, violin plot of the cyclizabilities of all 19,907 sequences in the cerevisiae nucleosomal library. 'x' denotes looping times of the ten sequences measured by smFRET (**c**).

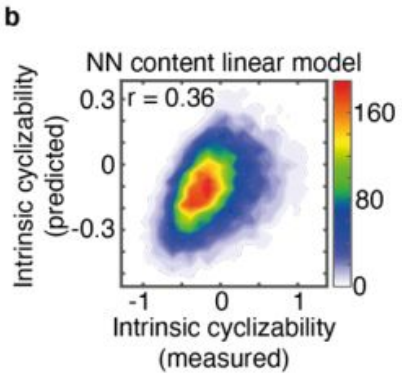
DNA Bendability and Loop-Seq[2]

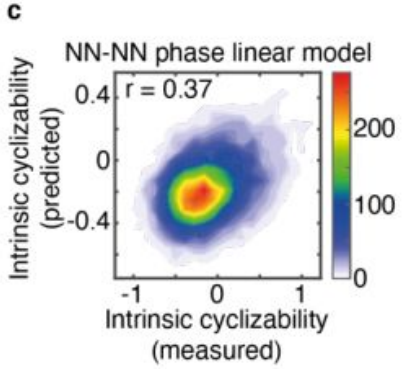
Deciphering the mechanical code of genome and epigenome

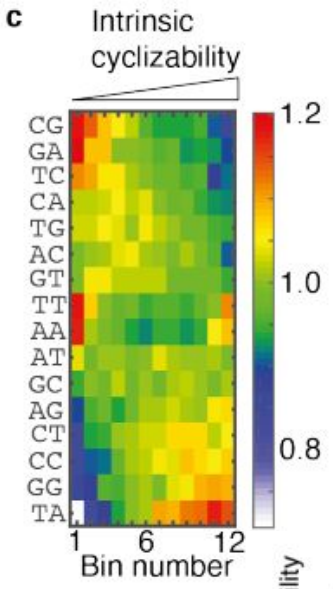
Aakash Basu[1], Dmitriy G. Bobrovnikov[1], Basilio Cieza[2], Zan Qureshi[2], and Taekjip Ha[1,2,3,4,a].

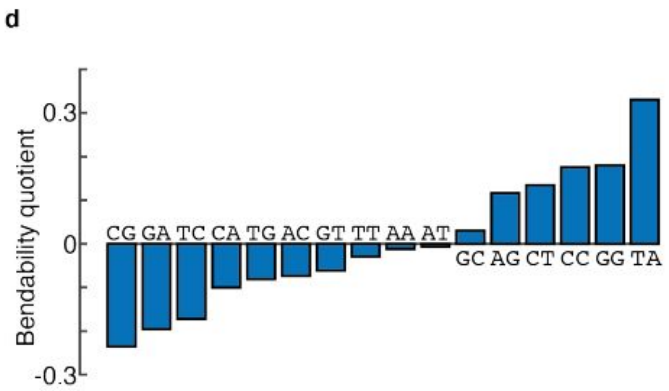
[1]Department of Biophysics and Biophysical Chemistry, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA; [2]Department of Biophysics, Johns Hopkins University, Baltimore, MD 21218, USA; [3]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21205; [4]Howard Hughes Medical Institute, Baltimore, MD 21205, USA.

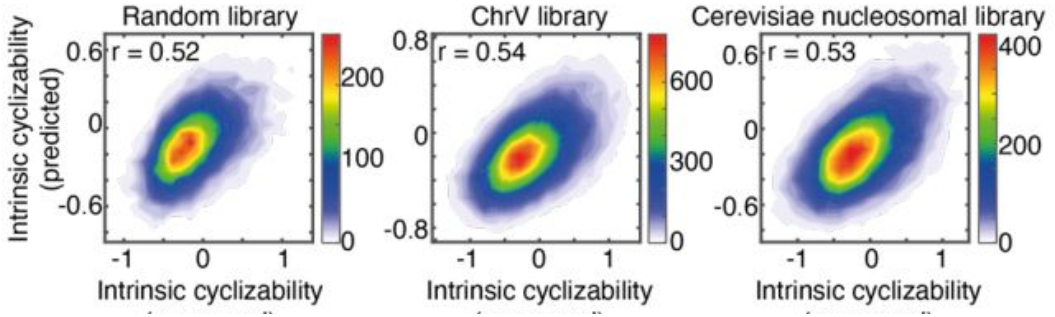


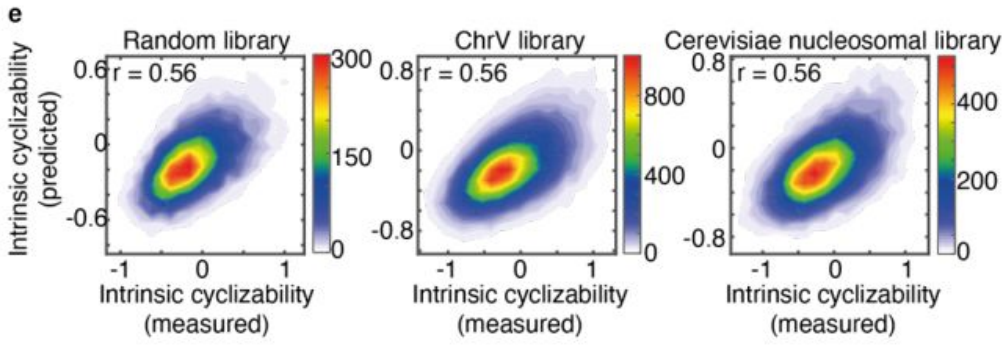












DNA Bendability and Loop-Seq[2]

Deciphering the mechanical code of genome and epigenome

Aakash Basu[1], Dmitriy G. Bobrovnikov[1], Basilio Cieza[2], Zan Qureshi[2], and Taekjip Ha[1,2,3,4,a].

[1]Department of Biophysics and Biophysical Chemistry, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA; [2]Department of Biophysics, Johns Hopkins University, Baltimore, MD 21218, USA; [3]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21205; [4]Howard Hughes Medical Institute, Baltimore, MD 21205, USA.

- How the total contents of various dinucleotides in a sequence influences its intrinsic cyclizability, ignoring for now how the dinucleotides are distributed along the sequence
- Found that TpA has the highest bendability quotient, implying that, on average, sequences with high intrinsic cyclizability tend to have more TA steps. CpG has the lowest bendability
- quotient
- Certain dinucleotides such as TT and AA are represented highly in both the bins of extremely high and the bins of extremely low intrinsic cyclizability, suggesting that their distribution along the DNA segment could play a greater role in determining their contribution to intrinsic cyclizability than their overall content alone.
- Intrinsic cyclizability is related to DNA shape
  - DNA shape has been described by various features that define the local geometry of base pairing, such as helical twist, propeller twist, roll, shift, etc
- DNA mechanics and amino acid sequence are linked
- CpG Methylations influences local DNA bendability
- Applications of the predictive model for the sequence-dependence of intrinsic cyclizability
  - around TSS(Transription Start Sites)s of genes
  - around Nucleosome attach sites

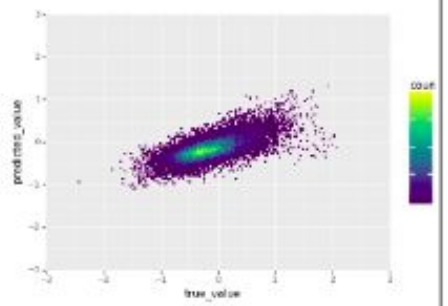
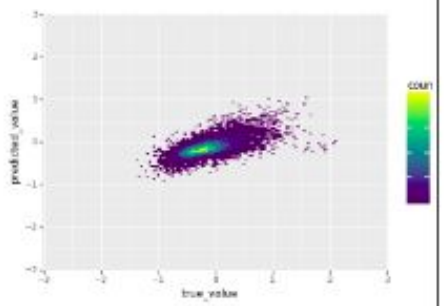
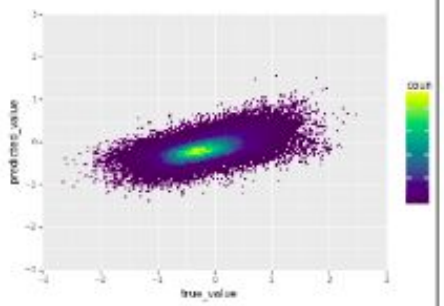
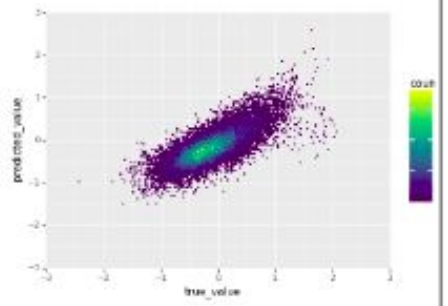
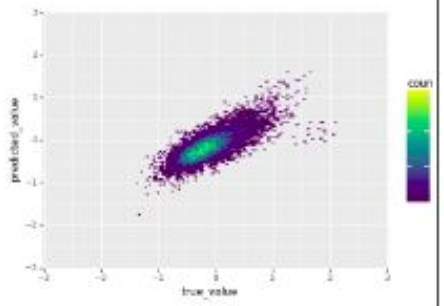
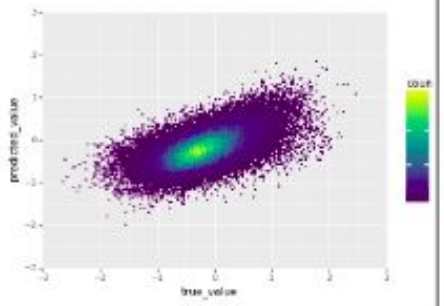
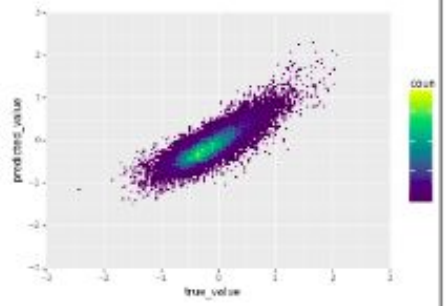
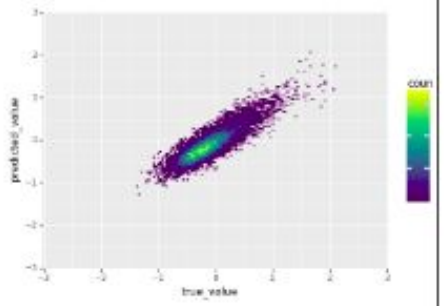
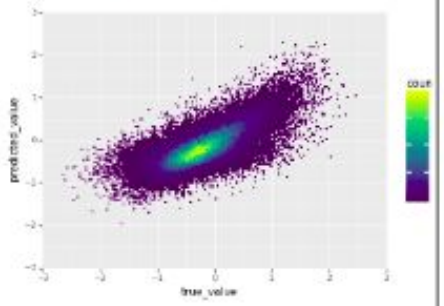
Predicting bendability from sequence

Input : DNA Sequence

Output: Intrinsic Cyclizability (C0) of the Sequence

Sequence	C0
ACTCATGCGGGTGCTATGATACAATTATATCTTATTTCCATTCCCATATG	0.061849
GGATCTATAACGAAATGTCAAATAATTTTACGGTAATATAACTTATCAGC	-0.285112
ATCAATAATTTATGTTCTTAACCTAACATTTGATGACCTTTGATGCGTTG	1.1147800
AAAATAATTAGAAAGTAGCACAATTTTTACAGTAATGTAGCACGCGTAAC	0.414965
CTTGAGAACCAAAAAAAAAAAAAAAAAATACTGATCCTTACAGGTTTTA	0.353268
GAGCTTGTTCAGGTAAATCAAAAATGGCAATTGAAATATAATATTACCAA	-0.649921
CTAAAAGGACATCCAATTTAATTAATAATTTTTTTTCTTTGAAAGAGGTA	-1.170588
GGTCGCGGAAGCCGTCTGTGTTTCAGCATGATTGAATCTTGAAATTGAAG	-0.940166
TCTTCATTCGTAGTCTGTGGCCTCCATGTTGGATAGACCGTAACAACATC	-0.3658

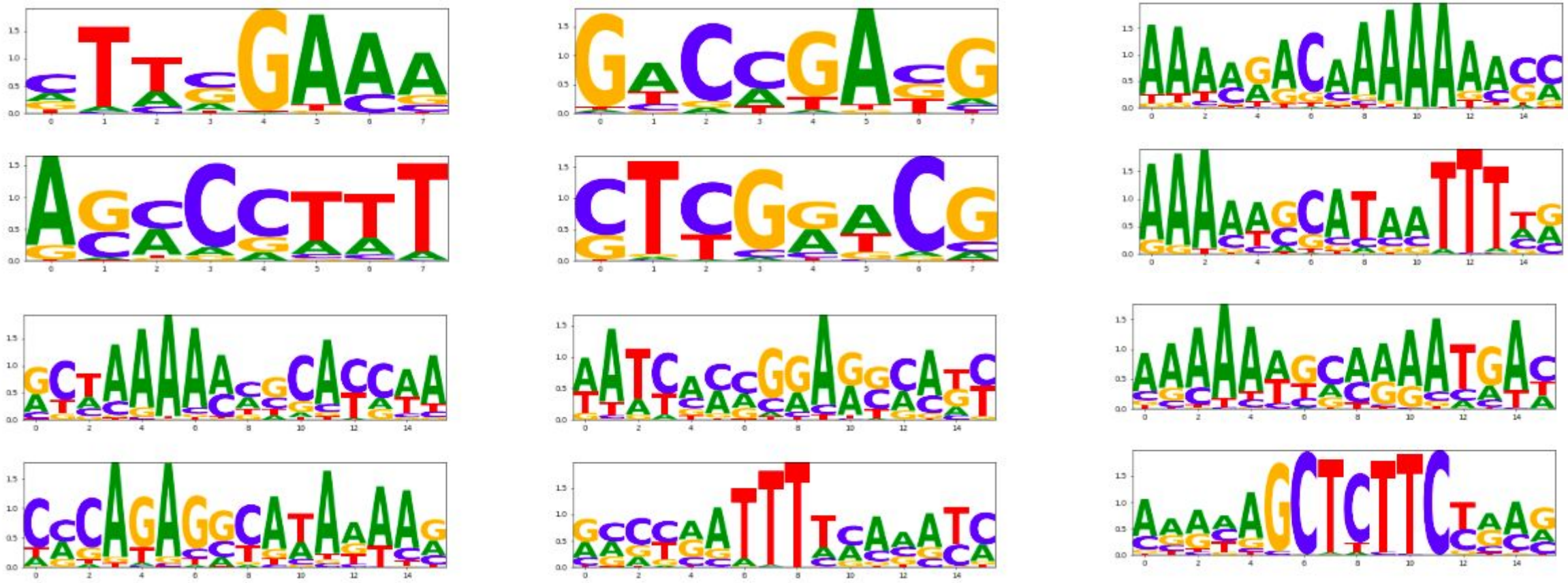
Current Models and Results

	Cerevisiae Nucleosomal Library	Random Library	ChrV Library
Single Kernel Size (24)	r = 0.64 	r = 0.6156 	r = 0.4606 
Multiple Kernel Size (8, 16, 32, 50)	r = 0.743 	r = 0.743 	r = 0.571 
2 Convolution layers (Model 4)	r = 0.837, ρ = 0.807 	r = 0.854, ρ = 0.834 	r = 0.663, ρ = 0.659 

Current Models and Results

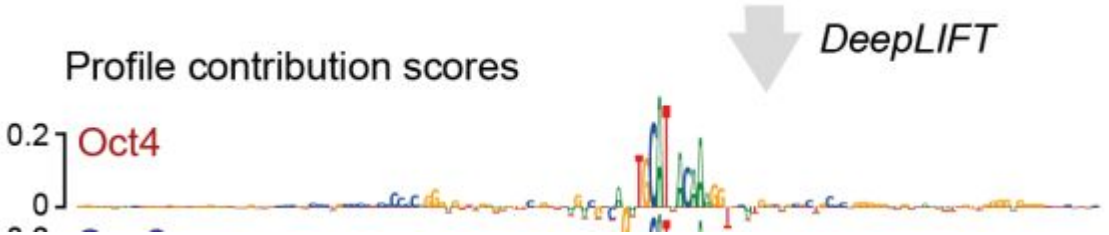
Motif Discovery from the Meuseum Model [*] (Multiple Sized Kernels) [all motifs]:

- Did not compare the relative importance of these motifs yet.
- Have not systematically checked these motifs against known motifs. Need to use tomtom from meme-suite to check against known motifs in Yeast DNA.



Newer interpretations

- We would want to interpret our model and find out motifs that effect dependability of DNA
- Confirmatory results and preferably newer motifs and interactions
  - Integrated Gradients [link], DeepLift [3] , DeepExplain, Shap



[4]

[Have recently started to look at deeplift papers. Have not got any results yet]

- Finding correlation with other DNA features and functions
  - Gene regulation
    - Comparing with known motifs of Yeast and also humans
  - Mutation, SNB

[Have not started work here yet]

References

[1] Measuring DNA mechanics on the genome scale; Aakash Basu, Dmitriy G. Bobrovnikov, Zan Qureshi, Tunc Kayikcioglu, Thuy T. M. Ngo, Anand Ranjan, Sebastian Eustermann, Basilio Cieza, Michael T. Morgan, Miroslav Hejna, H. Tomas Rube, Karl-Peter Hopfner, Cynthia Wolberger, Jun S. Song & Taekjip Ha

[2] Deciphering the mechanical code of genome and epigenome; Aakash Basu, Dmitriy G. Bobrovnikov, Basilio Cieza, Zan Qureshi, and Taekjip Ha

[3] Learning Important Features Through Propagating Activation Differences; Avanti Shrikumar, Peyton Greenside, Anshul Kundaje

[4] Deep learning at base-resolution reveals cisregulatory motif syntax; Žiga Avsec, Melanie Weilert, Avanti Shrikumar, Sabrina Krueger, Amr Alexandari, Khyati Dalal, Robin Fropf, Charles McAnany, Julien Gagneur, Anshul Kundaje and Julia Zeitlinger