A deep learning model reveals sequence signatures associated with DNA bendability and links bendability-altering mutations with aberrant chromosomal conformation

Samin Rahman Khan[1], Sadman Sakib[1], M. Sohel Rahman[1], Md. Abul Hassan Samee[2]

[1]Bangladesh University of Engineering and Technology, Computer Science and Engineering, Dhaka, Bangladesh,

[2]Baylor College of Medicine, Molecular Physiology and Biophysics, Houston, TX

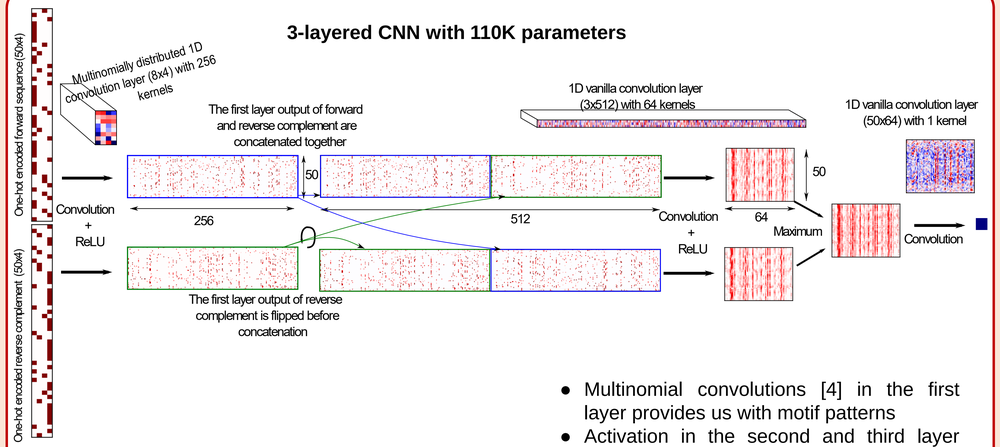
Introduction

Motivation: Bendability is a critical mechanical property of the genomic DNA with implications for its structure and function. An accurate model of the sequence features that determine DNA bendability can reveal how the genome structurally organizes into loops and domains.

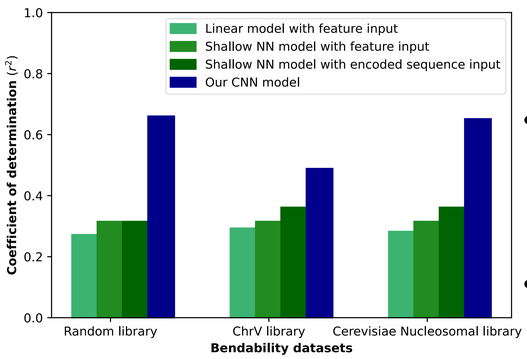
State of the art: Thanks to technological advances (loop-seq; [1]), it was recently possible to assay bendability across the yeast genome at high resolution (nearly 200,000 sequences, each 50 bps long). Current models of loop-seq data have confirmed the role of classically known features, such as distributions of dinucleotides and dinucleotide pairs, in determining bendability [2, 3]. However, the models leave ~60% of the variance unexplained.

Aim: To improve the performance and interpretability of bendability prediction models to pinpoint the sequence features that determine DNA bendability and, in turn, affect chromatin conformation.

A Deep Convolutional Neural Network Model of DNA Bendability

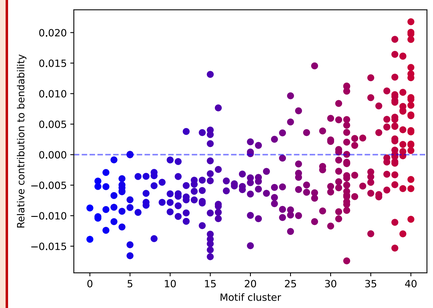


- Multinomial convolutions [4] in the first layer provides us with motif patterns
- Activation in the second and third layer shows the motif grammar
- We ranked motifs according to their contribution to bendability by turning off the activation of the corresponding kernel and observing changes in predicted bendability. [5]

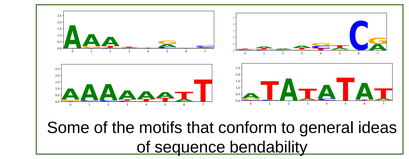


- Models by Basu [2] use different features including conventional ones like frequencies of dinucleotides and dinucleotide pairs. These individual models predict sequence bendability with a coefficient of determination of around 0.36-0.40.
- Our deeplearning model trained on bendability of segments across the whole yeast genome has an average coefficient of determination of ~0.61 on held-out data.

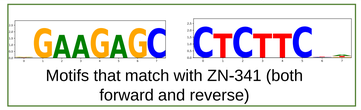
Model Confirms Classically Known Sequence Features, Discovers Transcription Factor Binding Motifs and Quantifies Each Feature's Contribution to Bendability



We obtained 256 motifs and grouped them into 41 clusters and ranked all the motifs according to their contribution to bendability



Some of the motifs that conform to general ideas of sequence bendability



Motifs that match with ZN-341 (both forward and reverse)

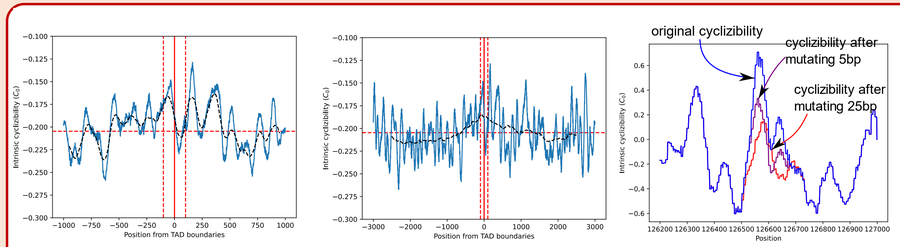
Motif that match with M0524_1.02(RIM101)

Motif-139 Motif-192

Potential new bendability motif

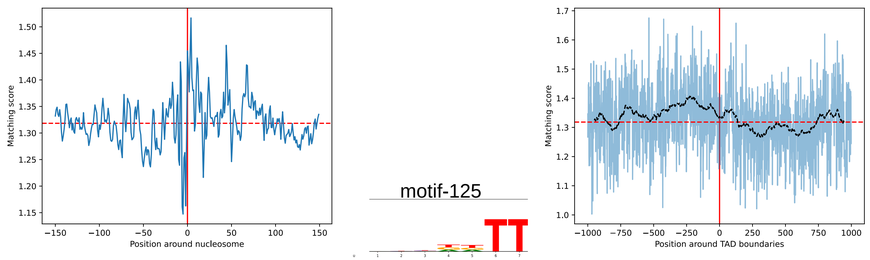
- Motifs that belonging to same clusters may contribute differently to bendability.
- Confirmatory patterns:
  - AT rich regions and AA and TT dinucleotides contribute positively to bendability
  - Long dA:dT, CG contributes negatively
- Transcription binding site motifs:
  - M0524_1.02(RIM101), M2109_1.02(GCR1), M4289_1.02 (GTS1) and more matches in CIS BP yeast database
  - Matches with ZN-341, HoXA10 and other zinc finger and homeobox TFBS families from HOCOMOCOv11
- Interesting novel motifs:
  - Motifs like motif-139 and 192 have no match in TFBS libraries and are potential novel bendability controlling motifs

Influence of DNA Bendability on Chromatin Conformation



Consistent patterns at TAD boundaries: high bendability (1000 bp) and periodically altering bendability (100bp). Sequence mutations causing changes in such patterns

- **Hypothesis based on prior research:**
  - High bendable regions indicate TFs and nucleosome occupancy
  - Densely packed TFs and nucleosome make the chromatin in these region more rigid causing them to become boundaries [6]
  - Disruption of these bendability patterns will cause abnormal chromatin conformation.



We find that, nucleosome and TAD boundaries are more enriched with our patterns. Here is an example of enrichment of motif-125 around nucleosome and boundaries in ChrV

Conclusion

- Our model predicts intrinsic bendability of a sequence more accurately
- Motif patterns have been learnt by our model from the data itself.
- From the model we can determine which regions in a sequence are most important for bendability at bp resolution.
- Mutations of these important region bring changes in bendability patterns which could potentially impair chromatin conformation

References

[1] Basu, A., Bobrovnikov, D.G., Qureshi, Z. *et al.* Measuring DNA mechanics on the genome scale. *Nature* **589,** 462–467 (2021). https://doi.org/10.1038/s41586-020-03052-3

[2] Basu, Aakash & Bobrovnikov, Dmitriy & Cieza, Basilio & Qureshi, Zan & Ha, Taekjip. (2020). Deciphering the mechanical code of genome and epigenome. 10.1101/2020.08.22.262352

[3] Zhang, Y., Basu, A., Ha, T., & Bialek, W. (2020). Searching for sequence features that control DNA flexibility. *arXiv: Biological Physics*.

[4] Multinomial Convolutions for Joint Modeling of Sequence Motifs and Enhancer Activities; Minjun Park, Salvi Singh, Francisco Jose Grisanti Canozo, Md. Abul Hassan Samee; bioRxiv 2020.07.28.224212; doi: https://doi.org/10.1101/2020.07.28.224212

[5] Global importance analysis: An interpretability method to quantify importance of genomic features in deep neural networks Koo PK, Majdandzic A, Ploenzke M, Anand P, Paul SB (2021) Global importance analysis: An interpretability method to quantify importance of genomic features in deep neural networks. PLOS Computational Biology 17(5): e1008925. https://doi.org/10.1371/journal.pcbi.1008925

[6] Dixon, J. R., Gorkin, D. U., & Ren, B. (2016). Chromatin Domains: The Unit of Chromosome Organization. *Molecular cell*, *62*(5), 668–680. https://doi.org/10.1016/j.molcel.2016.05.018

Contacts: saminrahmankhan97@gmail.com, saadsakib3@gmail.com, sohel.kcl@gmail.com, Md.AbulHassan.Samee@bcm.edu