Deep Learning Modeling of Transcription Factor Binding Specificity Using DNA Biophysical Properties

Manish Bhattarai[1], Raviteja Vangara[1], Sheng Qian[2], Tsu-Pei Chiu[3], Kim O. Rasmussen[1], Alan R. Bishop[1], Jubao Duan [2,4], Remo Rohs[3], Xin He[2], Boian S. Alexandrov[1]

[1]Los Alamos National Laboratory, [2] University of Chicago, [3]University of Southern California, [4]NorthShore University HealthSystem

ASHG VIRTUAL MEETING 2021 OCTOBER 18-22

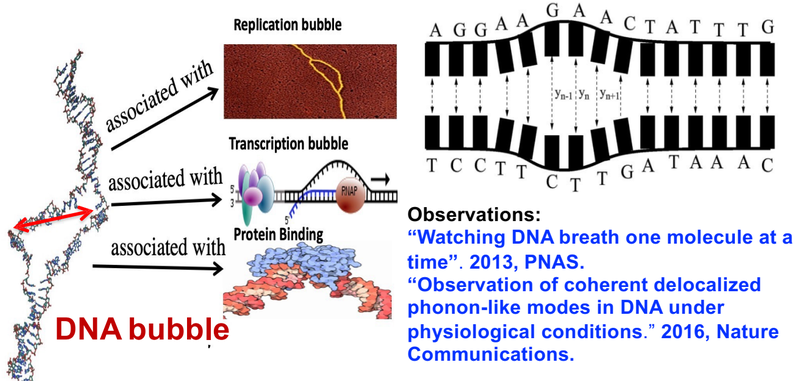
1. Background and Motivation

Active transcription is initiated and assisted by transcription factors (TF) binding to DNA.

- Important for TF-DNA binding are DNA sequence and local biophysical properties of DNA such as DNA breathing and local 3D structure (shape) of DNA.
- We integrate biophysical properties with sequence into modeling of TF binding specificity in a **Deep Attention DNA machine learning model** (**DADm**).
- We demonstrate that DADm, outperforms Support Vector Regressor (SVR), Deep-Sea, and DeepBind. and that DNA breathing and shape characteristics-augmented models compared favorably to models solely based on sequence.

2. DNA breathing: local transient bubbles

DNA Bubble Modeling: "Statistical mechanics of a nonlinear model for DNA denaturation", 1989, PRL



Observations:
"Watching DNA breath one molecule at a time". 2013, PNAS.
"Observation of coherent delocalized phonon-like modes in DNA under physiological conditions." 2016, Nature Communications.

3. Data: HT-SELEX; gcPBM; Chip-Seq

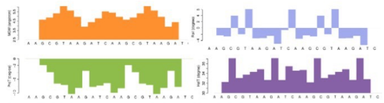
"High-throughput SELEX determination of DNA sequences bound by transcription factors in vitro ", 2012, Methods Mo. Biol. **(HT-Selex: 215 TFs)**
"Quantitative modeling of transcription factor binding specificities using DNA shape".PNAS,2015 **(gcPBM:3 TFs)**
"Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning." 2015, Nature Bio-technology **(Chip-Seq:506 experiments; Encode)**

4. Input Features to DADm

DNA Sequence: A,T,G,C

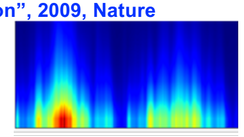
TTGTTGCCCAGGCTGGAGCGCAAGGGTACAATCTCAGCTCACCGCAACCTCCGCCTCCCGGGTTCAA
GTGATTCTCCTGCCTCAGCCTTCCAAGTAGCTAGGATT
TTGTTGCCCAGGCTGGAGCGCAAGGGTACAATCTCAGCTCACCGCAACCTCCGCCTCCCGGGTTCAA
GTGATTCTCCTGCCTCAGCCTTCCAAGTAGCTAGGATT

DNA Shape: MGW, Roll, ProT, HelT, and Electrostatic potential.



"The role of DNA shape in protein–DNA recognition", 2009, Nature

DNA bubbles: Probability for flipping, average opening , variation of the openings



"DNA dynamically directs its own transcription initiation", 2004, NAR

5. Schematics of DADm

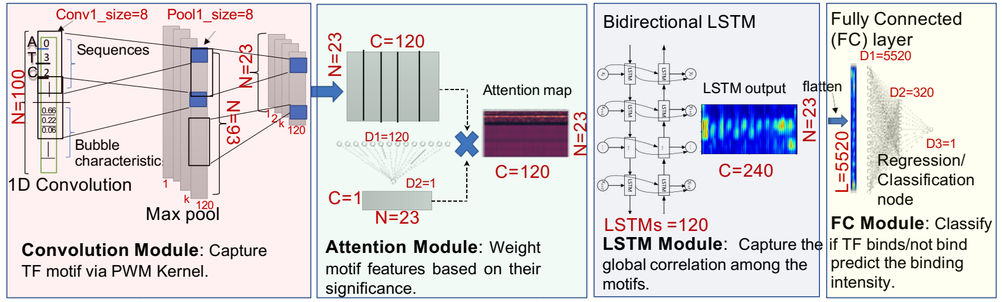
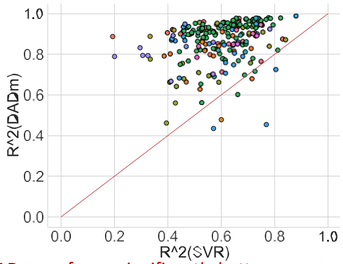


Fig: Proposed DADm model for short sequence(N=100). The model and data parameters are shown and vary for different input lengths.

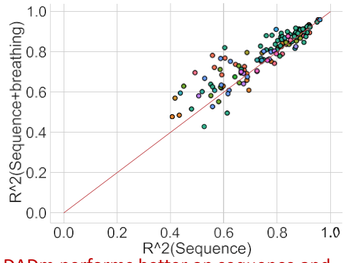
6. HT-Selex: 215 TFs

6a. Sequence only DADm vs SVR:



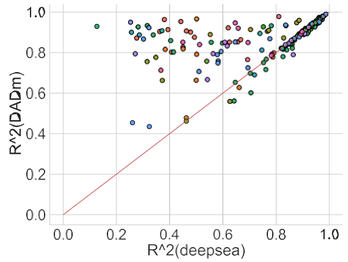
DADm performs significantly better compared to SVR on sequences only.

6b. DADm: Sequence + Breathing vs Sequence:



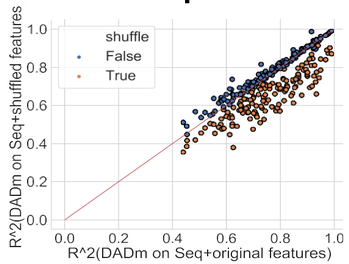
DADm performs better on sequence and breathing features compared to sequences only.

6c. DADm vs Deep-Sea:



DADm performs significantly better compared to modified DeepSea on sequences only.

6d. DADm: Seq. + Shuffled Breath. vs Seq. + Breath.



Demonstrate feature significance as DADm performance degrades with shuffled features compared against original features.

7. SVR: gcPBM

Augmentation of DNA features with sequences improve the prediction for SVR.

Data	Seq(R^2)	seq+breath.(R^2)	seq+shape(R^2)	seq+shape+breath.(R^2)
Mad	.90	.915	.93	.94
Max	.84	.88	.92	.93
Myc	.83	.87	.88	.89

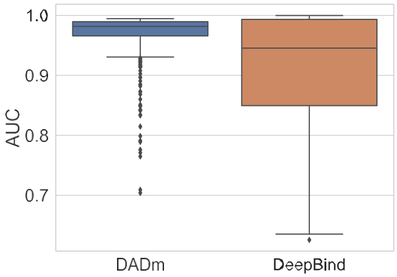
8. DADm: gcPBM

Augmentation of DNA features improve With sequences improve the prediction for DADm.

Data	Seq(R^2)	seq+breath.(R^2)	seq+shape(R^2)	seq+shape+breath.(R^2)
Mad	.95	.96	.945	.95
Max	.92	.93	.922	.94
Myc	.89	.91	.895	.914

R^2/AUC are used for regression/classification scores

9. Chip-Seq:



Comparison of performance between DeepBind and DADm on held-out Encode Chip-Seq data taken from Deep-Bind supplementary materials. For Chip-Seq, DADm achieves better overall classification AUC compared to Deep-Bind.

Los Alamos NATIONAL LABORATORY
National Nuclear Security Administration
Managed by Triad National Security, LLC, for the U.S. Department of Energy's NNSA.

Funding: NIH R01MH116281 (to J.D., X.H and B.S.A)