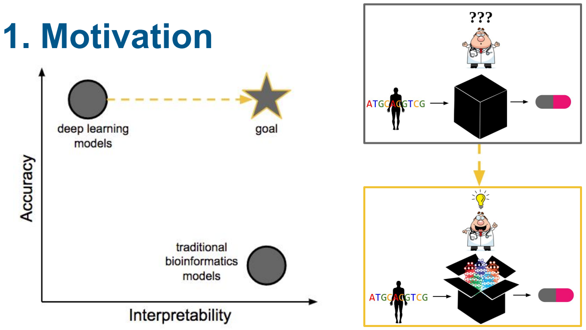
UNIVERSITY of VIRGINIA

Deep Motif Dashboard: Visualizing and Understanding Genomic Sequences Using Deep Neural Networks

Jack Lanchantin, Ritambhara Singh, Beilun Wang, Yanjun Qi
University of Virginia, Department of Computer Science

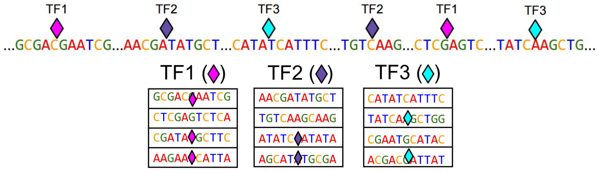
PSB 2017

1. Motivation

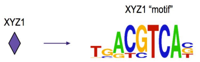


2. Genomic Classification Task

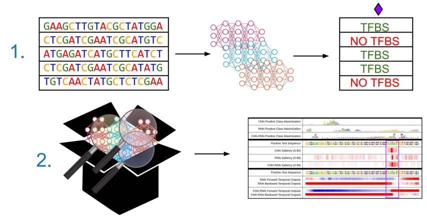
Transcription Factors (TFs) are proteins which bind to DNA and regulate gene expression. Predicting and understanding the **Transcription Factor Binding Sites** (TFBSs), or subsequences where TFs bind is important to biologists.



The binding of a TF is triggered by local sequential patterns within TFBSs, known as "**motifs**". Previous methods predicted TFBSs by constructing motifs using **position weight matrices** which best represented the binding sites.



3. Deep Motif Dashboard



4. NN Models for TFBS Prediction

1. **Convolutional (CNN)**

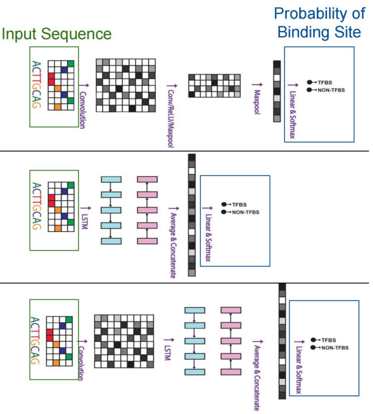
(short local patterns, or motifs)

2. **Recurrent (RNN)**

(long term dependencies)

3. **Convolutional-Recurrent (CNN-RNN)**

(long term dependencies among motifs)

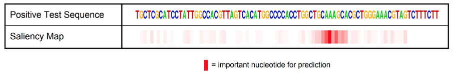


5. Visualization Methods

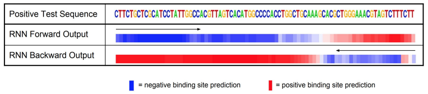
1. Saliency Maps: which nucleotides are most important for classification?

$$S_+(X) \approx w^T X + b = \sum_{i=1}^{|X|} w_i x_i$$

$$w = \left.\frac{\partial S_+}{\partial X}\right|_{X_0} = \text{“saliency map”}$$



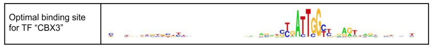
2. Temporal Output Scores: what are the model's predictions at each timestep of the DNA sequence?



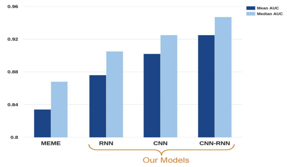
3. Class Optimization: for a particular TF, what does the optimal binding site sequence look like?

$$\arg\max_X S_+(X) + \lambda \|X\|_2^2$$

Where X is the input sequence and the score S_+ is probability of sequence X being a positive binding site

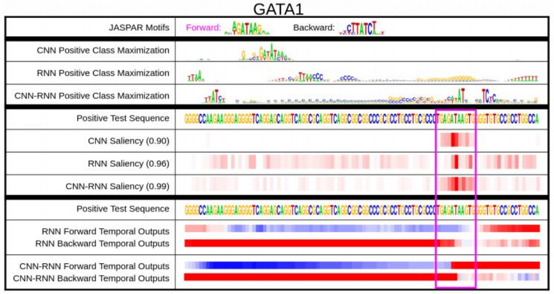


6. Results

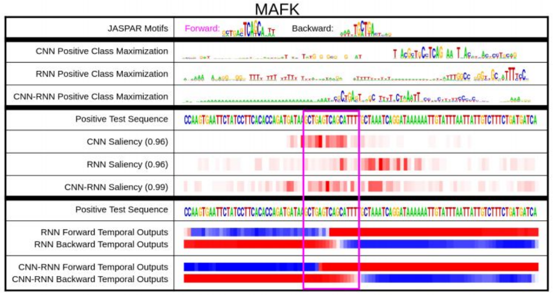


AUC on TFBS dataset from Alipanahi et al., 2015

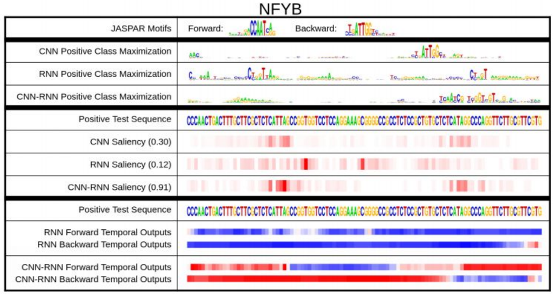
GATA1



MAFK



NFYB


JASPAR Motifs
Forward:
Backward:
CNN Positive Class Maximization
RNN Positive Class Maximization
CNN-RNN Positive Class Maximization
Positive Test Sequence
CNN Saliency (0.30)
RNN Saliency (0.12)
CNN-RNN Saliency (0.91)
Positive Test Sequence
RNN Forward Temporal Outputs
RNN Backward Temporal Outputs
CNN-RNN Forward Temporal Outputs
CNN-RNN Backward Temporal Outputs