Predicting Transcription Factor Binding Sites Using a CNN

P-value > 0.5

Yves-Langston Mays
Wyatt Lamberth
Michael Carreno
Pierre Ingram
Alexander Rosales
Victoria Vu

UNIVERSITY OF HOUSTON

Contents

1. Project Timeline*

Milestone	Due Date	Task
Proposal Submission	March 21, 2025	Submit Proposal
Progress Report	April 11, 2025	Process Dataset and train CNN
Final Report	April 28, 2025	Evaluate Moel, Handle Errors

1.1. Objective

The goal of this project is to predict transcription factor binding sites in eukaryotic DNA using a 1 dimensional CNN. Our goal is to determine whether a 200 bp DNA sequence contains a binding site for a single transcription factor.

2. Background

Transcription factors are proteins that regulate gene expression by binding to specific DNA motifs in promoter regions. Accurately predicting these binding sites can be useful for:

Synthetic biology – Designing custom promoters for gene circuits
Gene therapy – Targeted activation or repression of genes
Functional genomics – Understanding regulatory networks in eukaryotic cells

3. Task Definition

Input: A 200 bp DNA sequence represented using one hot encoding
Output: Binary classification (1 = TF binding site, 0 = no TF binding site)
Approach: Train a CNN model to learn DNA sequence motifs associated with TF binding

3.1. Example Inputs

DNA Sequence	Label
ATGCCGTTAGCGTAC…	1 (Binding Site)
CGTATAGGCCGCTAA…	0 (No Binding Site)

3.2. Example Outputs

DNA Sequence	Probability	Label
ATGCCGTTAGCGTAC…	0.96	1 (Binding Site)
CGTATAGGCCGCTAA…	0.18	0 (No Binding Site)

4. Data Source

JASPAR Database (jaspar.genereg.net) – Provides validated TF binding sites for transcription factors

Synthetic Negative Samples – Generated by shuffling non-binding sequences from background genomic DNA to balance the dataset. This prevents the model from becoming biased towards positive samples and reduces overfitting, thus decreasing false positives.

5. Preprocessing Steps

One hot encode DNA sequences

Standardize all sequences to 200 bp

Introduce small mutations to sequences through swapping or deleting bases to simulate natural variation to provide more natural data.

Improve model generalization by applying techniques such as dropout, normalization, and data shuffling to improve model generalization.

Could potentially utilize positional information to account for perferencial positioning of binding sites and use secondary structure prediction to identify secondary structure features that may influence transcription factor binding.

Data could benefit from utilizing reverse complement sequences to create better generalization through learning binding sites from both strands of DNA.

6. Evaluating Metric

Accuracy – Overall model performance
- **AUC-ROC (Area Under the Curve - Receiver Operating Characteristic)**
- Assesses the model's ability to distinguish between binding vs. non-binding sites
- ROC Curve plots TPR vs. FPR across different thresholds
- AUC Score closer to 1 indicates better classification performance
- **Baseline Comparison** – The CNN will be compared against a random classifier (50% accuracy baseline) and a logistic regression baseline.

7. Model Selection

1D Convolutional Neural Networks (CNNs) are particularly effective for identifying sequence motifs in DNA because they can capture spatial hierarchies in sequential data. CNNs efficiently detect local patterns (such as binding motifs) through convolutional layers, making them suitable for biological sequence classification tasks where positional information is critical.

8. CNN Design

Input Layer: One hot encoded DNA sequence

Conv Layer 1: 16 filters, kernel size = 8, ReLU activation

Conv Layer 2: 32 filters, kernel size = 4, ReLU activation

Pooling Layers: Max-pooling layers (pool size = 2) after each convolutional layer to reduce dimensionality and control overfitting. Regularization Techniques: Dropout (rate = 0.5) applied after the Dense Layer to prevent overfitting. Optimization Algorithm: Adam optimizer due to its efficiency and adaptive learning rate capabilities. Learning Rate: Initial learning rate of 0.001, with adaptive adjustments using learning rate scheduling if necessary.

Flatten Layer

Dense Layer: Fully connected layer with 32 neurons

Output Layer: Sigmoid activation for binary classification

9. Small Mutations to Improve Model Generalization

Small mutations, such as random substitutions (e.g., A → G) and insertion/deletion of single nucleotides at random positions, introduce realistic variability. This improves the CNN's generalization ability, allowing it to perform better on unseen biological sequences.

10. Potential Areas for Model Improvement (Exploration Phase)

As part of early experimentation and optional exploration by team members, a few potential improvements have been considered for future iterations of the model. These ideas are **not finalized**, but may help inform model tuning discussions later in the project.

- **Global MaxPooling Layers**: May help the CNN detect motifs regardless of their position in the sequence.
- **Reduced Regularization**: Lower dropout or L2 penalty could preserve meaningful motif representations that are otherwise suppressed.
- **Fixed-Position Motif Embedding**: Embedding motifs at a known sequence position during dataset generation may improve learnability.
- **Filter Visualization**: Extracting convolutional filters to interpret what patterns the CNN is learning.
- **Logistic Regression Baseline**: Comparing CNN performance to a baseline logistic regression model using k-mer frequency or TF-IDF features.
- **Class Weighting Adjustments**: Exploring how weighting the loss function by class frequency might reduce class prediction bias.

These improvements are **currently exploratory** and can be incorporated during model tuning stages if the group agrees they are beneficial.