The use of shallow convolutional neural networks in predicting promoter strength in *Escherichia coli*

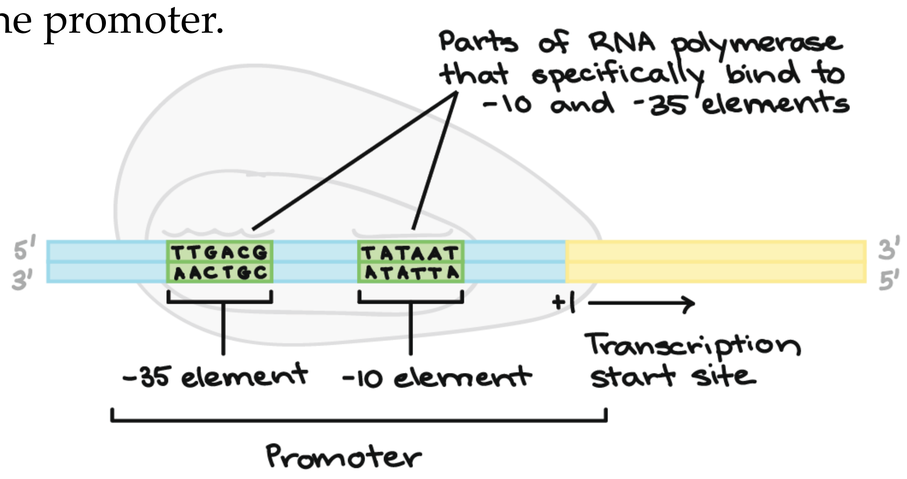
Jim Clauwaert[1,2], Michiel Stock[1], Marjan De Mey[2], Willem Waegeman[1]

Introduction

Aim: Predicting the rate of transcription (promoter strength) in microorganisms from a DNA sequence.

The promoter strength is correlated to the binding affinity between the RNA-polymerase (RNAP) protein and the promoter sequence.

Methodology: Using convolutional neural networks to map complex interactions of the DNA nucleotides (nt) upon the strength of the promoter.



https://www.khanacademy.org/science/biology/gene-expression-central-dogma/transcription-of-dna-into-rna/a/stages-of-transcription

Data

No large promoter libraries available

→ ***Use of ChIP-Chip data as a workaround to train the model***

- Chromatin Immunoprecipitation-chip (ChIP-Chip)
- Affinity measures between RNAP and DNA sequences
- Large datasets (**388468 samples**)
- DNA sequences (probes) of 50 nt

As ChIP-Chip data is very noisy, a **binary classification** approach was taken to train the model, separating labels into 0 (not a promoter) and 1 (promoter) values.

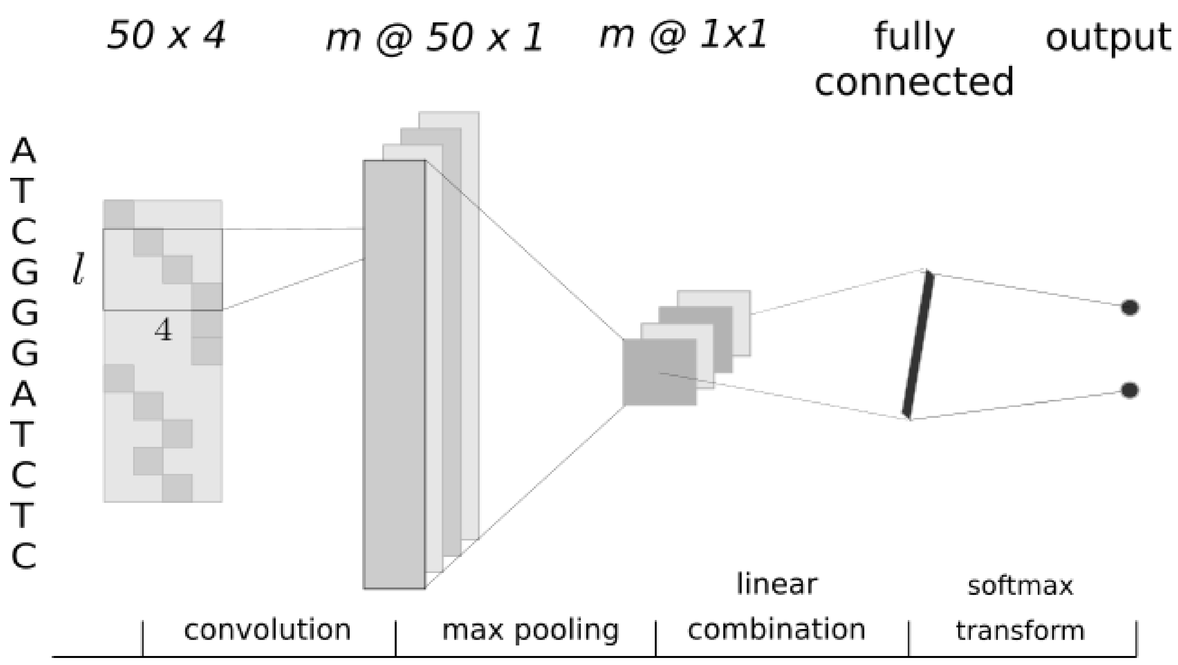
DNA-sequences are transformed into **4 x 50 binary images**

Models are trained solving a binary classification problem (not a promoter/promoter).

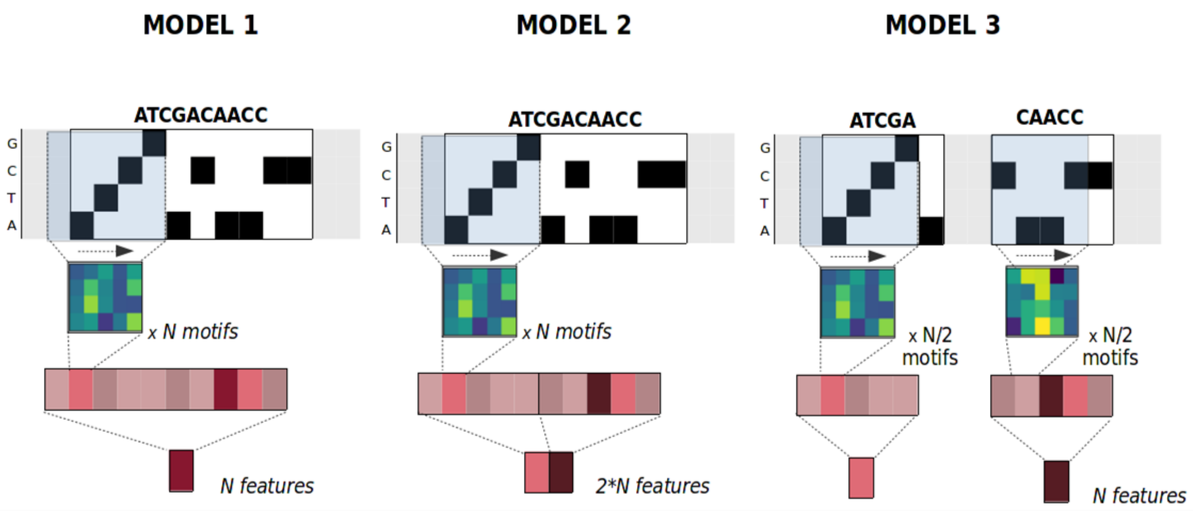
The model is tweaked to improve the Spearman rank correlation coefficient score obtained by ranking existing promoter libraries

Models

Architectures: Shallow convolutional neural networks are trained to give a probability prediction for each class.



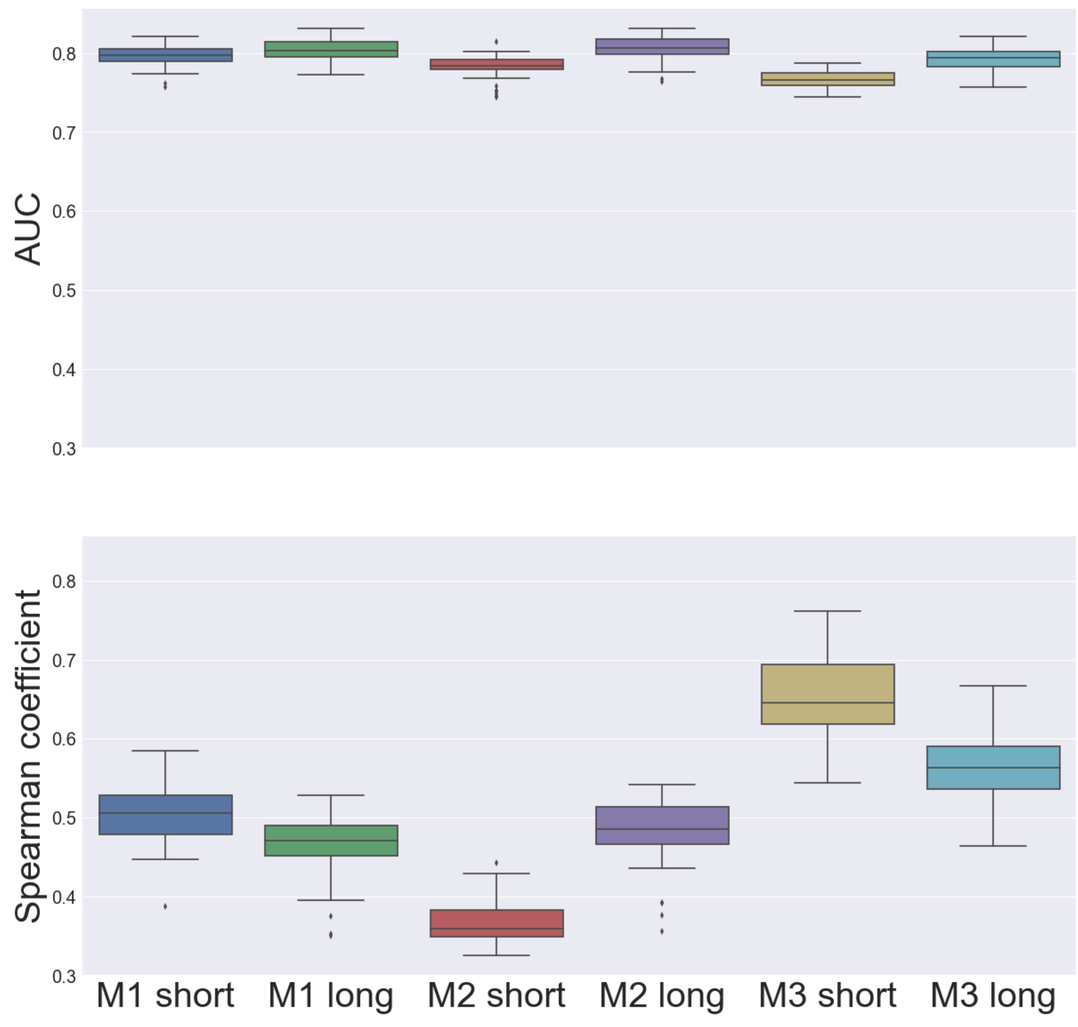
Three model architectures are considered:



Model 1 pools over the complete output of every kernel. **Model 2** adapts the size of the pooling layer to the length of the convolutional kernel to retain positional data of high-scoring subsequences.
Model 3 first splits the sequence into subsequences according to the kernel length. This reduces the number of outputs created by the pooling layer.

Results

Experimental setup: Short (10nt) and long (25 nt) kernels were used to evaluate all three models. The AUC score is obtained by classifying the test set. The spearman coefficient score is obtained by ranking existing promoter libraries.



Contact
jim.clauwaert@ugent.be

[1]**Kermit**
http://www.kermit.ugent.be

[2]**Centre for Synthetic Biology**
http://www.csb.ugent.be