Prediction of gene expression under environmental stress in *Orzya sativa*

Ketan Jog[1], Christina Azodi[2], Dr. Shin-Han Shiu[2]
Columbia University[1], Michigan State University[2]

Background

When a plant is exposed to **environmental stresses** like drought or cold, it triggers the production of regulatory proteins, called **transcription factors**

These regulators attach to **binding sites** on/near genes, which leads the gene to be transcribed into mRNA. This mRNA gets translated into proteins that help the plant combat the stress. This is called **gene expression**

Genes without stress induced transcription factor binding sites will be less responsive to that stress

Objectives

We build a deep learning model that **predicts** whether a gene is expressed in **response** to cold, drought or salt stress in both root and shoot tissue, using rice data

We interpret the trained model to **discover** putative novel transcription factor **binding sites**

Method

- Raw **RNA-Seq** reads were downloaded from the NCBI Sequence Read Archive (SRA)
- Reads were **mapped** to the rice genome
- Levels of **expression** were calculated
- Genes with a **log fold change** in **expression** between the control and each stress/tissue > 1 were considered up-regulated.

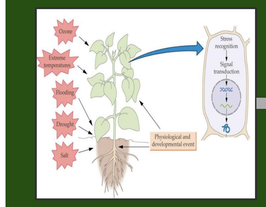
3000-bp of DNA sequence (1kb downstream and 0.5kb upstream from both ends of a gene) for each gene was one-hot encoded into a **binary matrix**

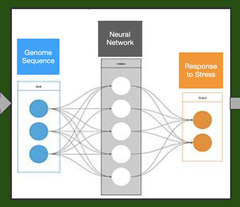
Model Layers

Genes	One Gene equals one sample	A C G A G C … T T A G A A T **Promoter < - > Terminator**
CNN	Identifies potential binding sites	A C G A G C T T A G A A T C T C G A T G G
Attention	Identifies higher order interactions	**G A G C T and T C G A T are both present in the PROMOTER region.**
Dense	Maps sites & interactions to stress	**G A G C T in promoter implies possibility for response to SALT STRESS**

The Big Picture

DNA Sequence → Prediction Model → Stress Response

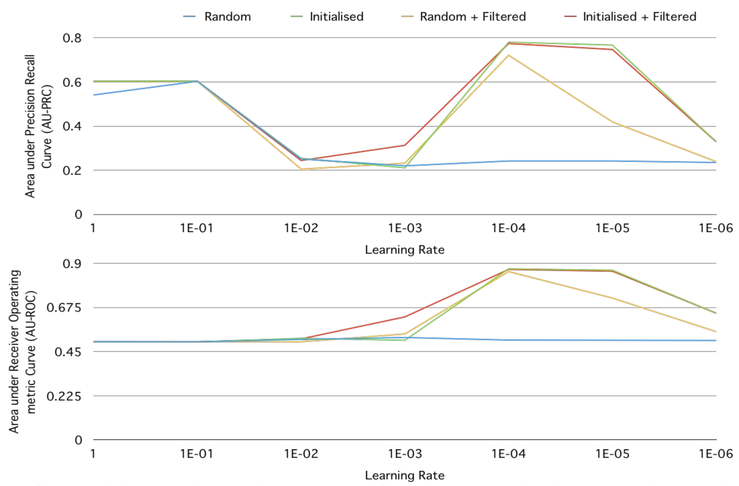




Example: **Rice variety B55 from China is tolerant to low temperatures during germination**

Results: Model Performance

We experimented to find the optimum **learning rate (LR)**:



We found that a learning rate of **1e-04** tended to produce the best results.

We experimented to find optimum **initial parameters (i.e. kernels)** and **input dataset**:

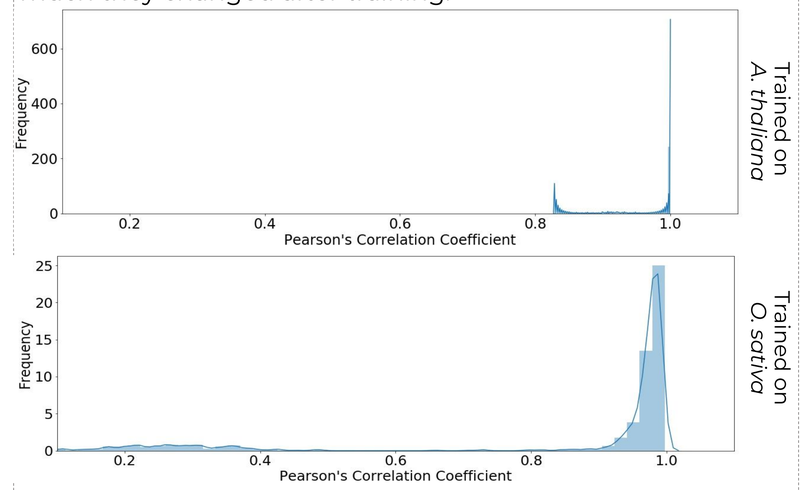
In another set of experiments, the dataset was **filtered** to improve loss and train time by removing genes **non-responsive** under all 6 conditions

All kernels were initialized by: (1) **randomly** drawing from a uniform Xavier distribution (2) from **known** transcription factor **binding sites** (TFBMs) from A. thaliana

LR = 1e-04	Random	Initialised	Random + Filtered	Initialised + Filtered
AU-PRC	0.242	***0.78***	0.721	0.774
AU-ROC	0.509	***0.872***	0.858	0.869

Results: Finding novel motifs

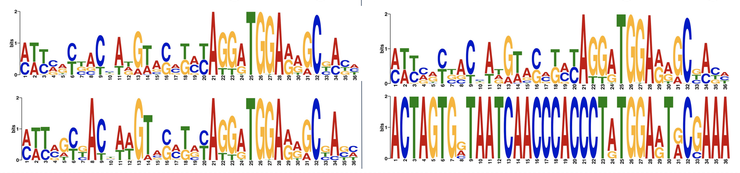
We extracted the **kernels** from the **CNN** layers, to see how much they changed after training:



Some known A. thaliana TFBM kernels were conserved after training on *O. sativa*, while others changed substantially.

A thaliana: pre and post training change

O sativa: pre and post training change



Takeaways

We were able to develop model that predicted environmental stress response for *O sativa*

We were able to use this model to discover novel putative binding site motifs in *O sativa*, as well as observe binding sites conserved from *A thaliana*

Acknowledgements

- NSF ACRES REU – OAC1560168
- Shiu Lab
- Our collaborators Yuning Hao and Yuying Xie in CMSE.