Annabelle Martin

6/10/2025

CSE 428, Computational Biology Capstone

**Generalizing Plant Gene Expression Models Across 15 Species**

Interpreting the DNA‐encoded cis‐regulatory code that governs plant gene expression across diverse lineages remains a key challenge. I expanded a convolutional neural network (CNN) framework from four to 15 model plants, spanning angiosperms and early‐diverging non‐flowering taxa, by extracting 1 kb upstream and 1 kb downstream per gene, one‐hot encoding these regions, and training binary classifiers on RNA‐Seq quartiles. Chromosome‐wise cross‐validation across four benchmark species yielded mean leaf accuracies 0.3–0.9 percentage point improvement over prior models. Shuffle controls remained near random (~33% accuracy). These results demonstrate that a broadly trained CNN captures evolutionarily conserved regulatory principles, laying the groundwork for robust cross-species annotation of plant cis-elements.

Gene expression in plants is tightly controlled by cis-regulatory elements, which are short DNA motifs located in promoter regions, untranslated regions (UTRs), and transcription terminators that recruit transcription factors and influence mRNA synthesis, processing, and stability. Experimental assays such as chromatin immunoprecipitation (ChIP-seq) or DNase hypersensitivity mapping can identify these elements but are labor-intensive, species-specific, and often miss context-dependent interactions. Convolutional neural networks (CNNs) offer a powerful, data-driven alternative, automatically learning sequence features predictive of expression directly from genomic and transcriptomic datasets. Recent CNN models trained on four well-studied plants (Arabidopsis thaliana, Solanum lycopersicum, Sorghum bicolor, Zea mays) achieved leaf expression prediction accuracies of 80–86% (auROC 0.85–0.92), highlighting proximal sequences’ predictive value, but their species-specific scope may limit discovery of universally conserved regulatory grammar.

The previous CNNs may capture species-specific features without fully revealing the cis-regulatory grammar conserved across broader evolutionary distances. This limited scope constrains the models’ applicability to less–studied crops and wild relatives. To address this, I trained a single CNN on 15 diverse plant genomes, using standardized leaf RNA-Seq labels and chromosome-wise cross-validation. My aim is to build a robust, generalizable model that uncovers fundamental regulatory principles shared across plants and improves expression prediction in underrepresented species.

I compiled a uniform data resource for 11 plant species by downloading their reference genomes and GTF annotations (Ensembl Plants v52) and publicly available leaf RNA-Seq experiments under comparable growth conditions. For each species, I generated a transcriptome FASTA with gffread and built a kallisto index. I then retrieved raw paired-end FASTQ from the SRA, or Expression Atlas, and quantified transcript abundance with kallisto quant. Sample-level TPMs were imported using the tximport R package, and per-gene expression was summarized as the log₁₊ₜₚₘ of the maximum TPM across replicates (logMaxTPM). To derive binary labels, I computed the 25th and 75th percentiles of the logMaxTPM distribution for each species. Genes with logMaxTPM ≤ 25th percentile were labeled “low” (0), those ≥ 75th percentile “high” (1), and intermediate values assigned to a “medium” class, which was excluded from binary training. This ensqured balanced, high-confidence training sets for the promoter\_vs\_terminator CNN.

I represented each gene by concatenating 1 kb of upstream sequence (covering promoter and 5′ UTR) with 1 kb of downstream sequence (3′ UTR plus terminator), yielding a 2 kb region per gene. These sequences are extracted via FastaSequenceLoader, which uses the reference genome FASTA and GTF annotation to pull precisely the TSS/UTR windows. Each nucleotide is then one-hot encoded into a 4-channel vector, producing an input tensor of shape (N\_genes × 3000 × 4). To avoid gene family bias, the data was split by chromosome: for each fold, one chromosome’s genes form the validation set, while all other chromosomes’ genes form the training set. The CNN architecture comprises three convolutional blocks followed by two dense layers. Each block contains two 1D convolutional layers (kernel size = 8) with filter counts that increase from 64 to 128 to 256, interleaved with max-pooling and 20% dropout for regularization. The convolutional output is flattened and fed into two fully connected layers of 512 and 128 units, each with 50% dropout, ending in a single sigmoid unit that outputs the probability of “high” expression.

The networks were trained to distinguish the top versus bottom quartile of expression using binary cross-entropy loss and the Adam optimizer (learning rate = 1e-3). Batches of 128 samples are processed for up to 50 epochs with early stopping based on validation loss. To confirm that the models learn genuine sequence features rather than nucleotide composition, I ran a parallel “shuffle” control where each training sequence is randomly permuted prior to encoding. All training is implemented in TensorFlow 2.x with GPU acceleration and memory growth enabled.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Plant** | **Original Accuracy (%)** | **15-Species Accuracy (%) (mean ± SD)** | **Original auROC** | **15-Species auROC (mean ± SD)** |
| **Arabidopsis thaliana** | 85.59 | 85.64 ± 2.10 | 0.92 | 0.928 ± 0.014 |
| **Solanum lycopersicum** | 83.55 | 84.48 ± 2.30 | 0.89 | 0.901 ± 0.025 |
| **Sorghum bicolor** | 79.70 | 80.55 ± 2.62 | 0.85 | 0.855 ± 0.056 |
| **Zea mays** | 81.16 | 81.48 ± 3.33 | 0.87 | 0.881 ± 0.041 |

The 15‐species SSR models yield slight but consistent accuracy gains over the original four‐species CNN. Across the four benchmark plants, mean leaf-model accuracy improved by 0.05 percentage points (ppt) in Arabidopsis, by 0.93 ppt in tomato, by 0.85 ppt in sorghum, and by 0.32 ppt in maize. Corresponding mean auROC values also rose across the board, demonstrating that taxonomic expansion marginally sharpens classification performance. Shuffle‐sequence controls remained near random (~ 33 % accuracy), confirming that learned features reflect genuine sequence signals. Overall, the combined four‐species average accuracy of about 83.04 % matches or slightly exceeds the original benchmarks, indicating that incorporating eleven additional genomes does not dilute but in fact enhances model robustness.

Expanding the CNN from four to 15 plant species yielded modest yet consistent improvements in leaf‐model performance across all benchmarks, indicating that broadening taxonomic scope strengthens the model’s ability to learn core regulatory features rather than overfitting species‐specific signals. The retention of high accuracy (80–86 %) and auROC (0.85–0.93) alongside low shuffle‐control performance confirms that the network remains sensitive to genuine cis‐regulatory grammar. Notably, increased standard deviations in some species (e.g., maize ±3.3 ppt) reflect variable chromosome counts and assembly quality, underscoring the importance of assembly completeness. By demonstrating improved robustness on well‐studied species, the 15‐species framework establishes a scalable foundation for comparative regulatory genomics across the plant kingdom.

Future steps would include detailed model interpretation via nucleotide‐resolution importance scoring (DeepLIFT) and motif discovery (TF‐MoDISco) to catalogue conserved cis‐elements across all 15 lineages. Incorporating additional tissues (roots, flowers) and environmental conditions (stress treatments) might improve the model’s generality. Architectural enhancements—such as attention mechanisms or multi‐task learning for simultaneous tissue predictions—and benchmarking against fast, interpretable models (e.g., random forests with k-mer features) can further refine performance and biological insight. Ultimately, extending this approach to non‐plant eukaryotes will probe the universality of cis‐regulatory codes.

Training a single CNN on 15 diverse plant genomes modestly improves leaf‐expression prediction in key model species and underscores the value of broad taxonomic representation in capturing conserved regulatory grammar. My results validate that core cis-regulatory signals in promoters and UTRs are somewhat generalizable across varying plant species. This unified framework paves the way for comprehensive cross-species annotation of regulatory elements and deeper biological interpretation in future work.

Citations

Peleke, F.F., Zumkeller, S.M., Gültas, M. *et al.* Deep learning the *cis*-regulatory code for gene expression in selected model plants. *Nat Commun* **15**, 3488 (2024). https://doi.org/10.1038/s41467-024-47744-0