AI deciphers new gene regulatory code in plants and makes accurate predictions for newly sequenced genomes

Press Release

Scientific Contact
Dr. Jedrzej Jakub Szymanski
Tel.: +49 39482 5753
szymanski@ipk-gatersleben.de

Media Contact
Christian Schafmeister
Phone: +49 39482 5461
schafmeister@ipk-gatersleben.de

Gatersleben, 25.04.2024 **Elucidating the relationship between the sequences of non-coding regulatory elements and their target genes is key to understanding gene regulation and its variation between plant species and ecotypes. Now, an international research team led by IPK Leibniz Institute and with the participation of Forschungszentrum Jülich developed deep learning models that link gene sequence data with mRNA copy number for several plant species and predicted the regulatory effect of gene sequence variation. The results were published in the journal "Nature Communications".**

Genome sequencing technology provides thousands of new plant genomes annually. In agriculture, researchers merge this genomic information with observational data (measuring various plant traits) to identify correlations between genetic variants and crop traits like seed count, resistance to fungal infections, fruit color, or flavor. However, the grasp of how genetic variation influences gene activity at the molecular level is quite limited. This gap in knowledge hinders the breeding of "smart crops" with enhanced quality and reduced negative environmental impact achieved by combination of specific gene variants of known function.

Researchers from the IPK Leibniz Institute and Forschungszentrum Jülich (FZ) have made a significant breakthrough to tackle this challenge. Led by Dr. Jedrzej Jakub Szymanski, the international research team trained interpretable deep learning models, a subset of AI algorithms, on a vast dataset of genomic information from various plant species. "These models not only were able to accurately predict gene activity from sequences but also pinpoint which sequence parts contribute to these predictions", explains the head of IPK's research group "Network Analysis and Modelling". The AI technology which the researchers applied is akin to that used in computer vision, which involves recognizing facial features in images and inferring emotions.

In contrast to previous approaches based on statistical enrichment, here the researchers combined identification of sequence features with determination of the mRNA copy number in the frame of a mathematical model that has been trained accounting for biological information on gene model structure and sequence homology, thus gene evolution.

"We were truly amazed by the effectiveness. Within a few days of training, we rediscovered many known regulatory sequences and found that about 50% of the features identified were entirely new. These models excellently generalized across plant species they were not trained on, making them valuable for analyzing newly sequenced genomes", says Dr. Jedrzej Jakub Szymanski. "And we specifically demonstrated their application in diverse tomato cultivars with long-read sequencing data. We pinpointed specific regulatory sequence variations that explained observed differences in gene activity and, consequently, variations in shape, color, and robustness. This is a remarkable improvement over classically used statistical associations of single nucleotide polymorphisms.”

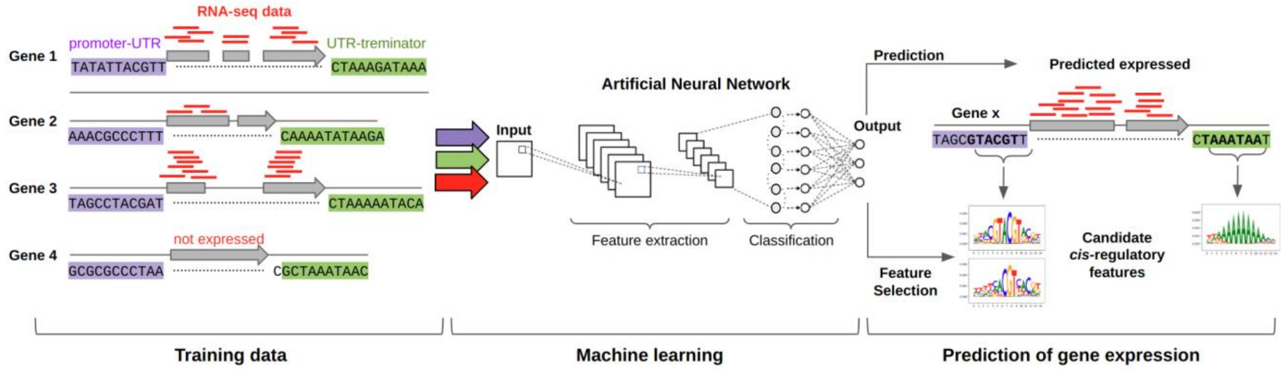
The team has openly shared their models and provided a web interface for their use. "Interestingly, much effort went into degrading our model's performance. To avoid overly optimistic results due to AI finding shortcuts required from me a deep dive into gene regulation biology to eliminate any potential bias, reduce data leakage and overfitting”, says Fritz Forbang Peleke, the lead machine learning researcher and first author of the study, which was published in the journal “Nature Communications”.

Dr. Simon Zumkeller, a co-author and evolutionary biologist from FZ Jülich, remarked, “With the presented analyses we can investigate and compare gene regulation in plants and infer its evolution. For practical applications, the method provides a new foundation, too. We are approaching the routine identification of gene regulatory elements in known and newly sequenced plant genomes, in various tissues, and under different environmental conditions."

Original publication:

Peleke *et al.* (2024): Deep learning the cis-regulatory code for gene expression in selected model plants. Nature Communications. DOI: 10.1038/s41467-024-47744-0

Figure:


RNA-seq data
promoter-UTR
UTR-treminator
Gene 1
TATATTACGTT
CTAAAGATAAA
Gene 2
AAACGCCCTTT
CAAAATATAAGA
Gene 3
TAGCCTACGAT
CTAAAAATACA
Gene 4
not expressed
GCGCGCCCTAA
CGCTAAATAAC
Input
Artificial Neural Network
Feature extraction
Classification
Output
Prediction
Predicted expressed
Gene x
TAGCGTACGTT
CTAAATAAT
Feature Selection
Candidate *cis*-regulatory features
Training data
Machine learning
Prediction of gene expression