Machine learning uncovers 'genes of importance' in agriculture and medicine

September 24 2021

Corn (maize) growing in the NYU Rose Sohn Zegar Greenhouse on the roof of the NYU Center for Genomics & Systems Biology. Credit: NYU Coruzzi Lab

Machine learning can pinpoint "genes of importance" that help crops to grow with less fertilizer, according to a new study published in *Nature Communications*. It can also predict additional traits in plants and disease outcomes in animals, illustrating its applications beyond agriculture.

Using genomic data to predict outcomes in agriculture and medicine is both a promise and challenge for systems biology. Researchers have been working to determine how to best use the vast amount of genomic data available to predict how organisms respond to changes in nutrition, toxins, and pathogen exposure—which in turn would inform crop improvement, disease prognosis, epidemiology, and public health.

However, accurately predicting such complex outcomes in agriculture and medicine from genome-scale information remains a significant challenge.

In the *Nature Communications* study, NYU researchers and collaborators in the U.S. and Taiwan tackled this challenge using machine learning, a type of artificial intelligence used to detect patterns in data.

"We show that focusing on genes whose expression patterns are evolutionarily conserved across species enhances our ability to learn and predict 'genes of importance' to growth performance for staple crops, as well as disease outcomes in animals," explained Gloria Coruzzi, Carroll & Milton Petrie Professor in NYU's Department of Biology and Center for Genomics and Systems Biology and the paper's senior author.

"Our approach exploits the natural variation of genome-wide expression and related phenotypes within or across species," added Chia-Yi Cheng of NYU's Center for Genomics and Systems Biology and National Taiwan University, the lead author of this study. "We show that paring down our genomic input to genes whose expression patterns are conserved within and across species is a biologically principled way to reduce dimensionality of the genomic data, which significantly improves the ability of our machine learning models to identify which genes are important to a trait."

Corn (maize) growing in the NYU Rose Sohn Zegar Greenhouse on the roof of the NYU Center for Genomics & Systems Biology. Credit: NYU Coruzzi Lab

As a proof-of-concept, the researchers demonstrated that genes whose responsiveness to nitrogen are evolutionarily conserved between two diverse plant species—Arabidopsis, a small flowering plant widely used as a [model organism](#) in plant biology, and varieties of corn, America's largest crop—significantly improved the ability of machine learning models to predict genes of importance for how efficiently plants use nitrogen. Nitrogen is a crucial nutrient for plants and the main component of fertilizer; crops that use nitrogen more efficiently grow better and require less fertilizer, which has economic and environmental

benefits.

The researchers conducted experiments that validated eight master transcription factors as genes of importance to nitrogen use efficiency. They showed that altered gene expression in Arabidopsis or corn could increase plant growth in low nitrogen soils, which they tested both in the lab at NYU and in cornfields at the University of Illinois.

"Now that we can more accurately predict which corn hybrids are better at using nitrogen fertilizer in the field, we can rapidly improve this trait. Increasing nitrogen use efficiency in corn and other crops offers three key benefits by lowering farmer costs, reducing environmental pollution, and mitigating greenhouse gas emissions from agriculture," said study author Stephen Moose, Alexander Professor of Crop Sciences at the University of Illinois at Urbana-Champaign.

Moreover, the researchers proved that this evolutionarily informed machine learning approach can be applied to other traits and species by predicting additional traits in plants, including biomass and yield in both Arabidopsis and corn. They also showed that this approach can predict genes of importance to drought resistance in another staple crop, rice, as well as disease outcomes in animals through studying mouse models.

"Because we showed that our evolutionarily informed pipeline can also be applied in animals, this underlines its potential to uncover genes of importance for any physiological or clinical traits of interest across biology, agriculture, or medicine," said Coruzzi.

"Many key traits of agronomic or clinical importance are genetically complex and hence it's difficult to pin down their control and inheritance. Our success proves that big data and systems level thinking can make these notoriously difficult challenges tractable," said study author Ying Li, faculty in the Department of Horticulture and Landscape

Architecture at Purdue University.

More information: Evolutionarily informed machine learning enhances the power of predictive gene-to-phenotype relationships, *Nature Communications* (2021). [DOI: 10.1038/s41467-021-25893-w](#)

Provided by New York University