Artificial intelligence helps predict performance of sugarcane in the field

October 19 2022, by Ricardo Muniz

A technique developed by Brazilian researchers enhances the efficiency of breeding programs, saving selection time and the cost of plant genotyping and characterization. Credit: IAC's Sugarcane Center in Ribeirão Preto

A Brazilian study published in *Scientific Reports* shows that artificial intelligence (AI) can be used to create efficient models for genomic selection of sugarcane and forage grass varieties and predict their performance in the field on the basis of their DNA.

In terms of accuracy compared with traditional breeding techniques, the proposed methodology improved predictive power by more than 50%. This is the first time a highly efficient genomic selection method based on machine learning has been proposed for polyploid plants (in which cells have more than two complete sets of chromosomes), including the grasses studied.

Machine learning is a branch of AI and computer science involving statistics and optimization, with countless applications. Its main goal is to create algorithms that automatically extract patterns from datasets. It can be used to predict the performance of a plant, including whether it will be resistant to or tolerant of biotic stresses such as pests and diseases caused by insects, nematodes, fungi or bacteria, and or abiotic stresses such as cold, drought, salinity or insufficient soil nutrients.

Crossing is the most widely used technique in traditional breeding programs. "You establish populations by crossing plants that are interesting. In the case of sugarcane, you cross a variety that produces a lot of sugar with another that's more resistant, for example. You cross them and then assess the performance of the resulting genotypes in the field," said computer scientist Alexandre Hild Aono, first author of the article on the study. Aono is a researcher at the State University of Campinas's Center for Molecular Biology and Genetic Engineering (CBMEG-UNICAMP). He graduated from the Federal University of São Paulo (UNIFESP).

"But this assessment process takes a long time and is very expensive. The method we propose can predict the performance of these plants even before they grow. We succeeded in predicting yield on the basis of the genetic material. This is significant because it saves many years of assessment," Aono explained.

In the case of sugarcane, the challenge is highly complex. Traditional

breeding techniques take between nine and 12 years and incur high costs, according to Anete Pereira de Souza, a professor of plant genetics at UNICAMP's Institute of Biology and Aono's Ph.D. supervisor at CBMEG.

"When breeders identify an interesting plant, they multiply it by cloning so that the genotype isn't lost, but this takes time and costs a great deal. An extreme example is the breeding of rubber trees, which can take as long as 30 years," Souza said. One way to surmount these difficulties is what she called "plant breeding 4.0," which makes intensive use of data analysis and highly efficient computational and statistical tools. Each genotyping-by-sequencing process can involve 1 billion sequences.

The main hurdle scientists face in trying to breed better varieties of polyploid plants such as sugarcane and forage grass is the complexity of their genomes. "In this case, we didn't even know if genomic selection would be possible, given the scarce resources and the difficulty of working with this complexity," Aono said.

Methods

The researchers began the genomic selection process with diploid plants (containing cells with two sets of chromosomes), as they have simpler genomes. "The problem is that high-value tropical plants like sugarcane aren't diploids but polyploids, which is a complication," Souza said.

While human beings and almost all animals are diploid, sugarcane may have as many as 12 copies of every chromosome. Any individual of the species Homo sapiens can have up to two variants of each gene, one inherited from the father and the other from the mother. Sugarcane is more complex because theoretically any gene can have many variants in the same individual. There are regions of its genome with six sets of chromosomes, others with eight, ten, and even 12 sets. "The genetics is

so complex that breeders work with [sugarcane](#) as if it were diploid," Souza said.

In 2001, Theodorus Meuwissen, a Dutch scientist who is currently a professor of animal breeding and genetics at the Norwegian University of Life Sciences (NMBU), proposed genomic selection to predict complex traits in animals and plants in association with their [phenotypes](#) (observable characteristics resulting from the interaction of their genotypes with the environment). The advantage of this approach to plant breeding is the link between the phenotypic traits of interest, such as yield, sugar level or precocity, and single nucleotide polymorphisms (SNPs). A "snip" (as SNP is pronounced) is a genomic variant at a single base position in the DNA, Souza explained.

"It's the difference in the genomes of any two individuals. For example, one may have an A (corresponding to the nucleotide adenine) that produces a little more than another with a G (guanine) at the same location in the genome. That changes everything," she said. "When you find an association with what you're looking for, like a high level of sugar production, and specific SNPs at different locations in the genome, you can sequence only the population on which your breeding work focuses."

The advances proposed by Aono and colleagues dispense with the need to plant and phenotype throughout the breeding cycle. "We do field experiments in the initial stages of the program to obtain the phenotype of interest for each clone," Souza said.

"In parallel, we sequence all the clones in the breeding population quite straightforwardly, without needing to have the whole genome for every clone. This is what's called genotyping-by-sequencing—partial sequencing in search of the differences and similarities in the base pairs for the clones, and their association with each clone's production. The

association between phenotype and genome shows which produces more and which SNPs are associated with higher production. In this manner, we can identify clones with a large proportion of the SNPs that contribute to the higher production observed in the initial experiments and obtain the most productive variety faster and more cheaply."

More information: Alexandre Hild Aono et al, A joint learning approach for genomic prediction in polyploid grasses, *Scientific Reports* (2022). DOI: 10.1038/s41598-022-16417-7

Provided by FAPESP