FEATURE ARTICLE

15-NOV-2019

AI for Plant Breeding in an Ever-Changing Climate

ORNL's Dan Jacobson and team design algorithms for climate-resilient crops

DOE/OAK RIDGE NATIONAL LABORATORY

How might artificial intelligence (AI) impact agriculture, the food industry, and the field of bioengineering? Dan Jacobson, a research and development staff member in the Biosciences Division at the US Department of Energy's (DOE's) Oak Ridge National Laboratory (ORNL), has a few ideas.

For the past 5 years, Jacobson and his team have studied plants to understand the genetic variables and patterns that make them adaptable to changing environments and climates. As a computational biologist, Jacobson uses some of the world's most powerful supercomputers for his work--including the recently decommissioned Cray XK7 Titan and the world's most powerful and smartest supercomputer for open science, the IBM AC922 Summit supercomputer, both located at the Oak Ridge Leadership Computing Facility (OLCF), a DOE Office of Science User Facility at ORNL.

Last year, Jacobson and his team won an Association for Computing Machinery Gordon Bell Prize after using a special computing technique known as "mixed precision" on Summit to become the first group to reach exascale speed--approximately a quintillion calculations per second.

Jacobson's team is currently working on numerous projects that form an integrated roadmap for the future of AI in plant breeding and bioenergy. The team's work was featured in Trends in Biotechnology in October.

In this Q&A, Jacobson talks about his team's work on a genomic selection algorithm, his vision for the future of environmental genomics, and the space where simulation meets AI.

What has your team been working on over the past year?

Jacobson: We've been working on a couple of things. Recently, we've developed new ways to do what's called "genomic selection," or designing an organism for breeding purposes. We've developed a new genomic selection algorithm that's driven by emerging machine learning methods collectively called "explainable AI," which is a field that improves on black box classifier AI methods by attempting to understand how these algorithms make decisions.

This algorithm helps us determine which variations in a genome we need to combine to produce plants that can adapt to their environments. This informs breeding efforts, gene editing efforts, or combinations of those, depending on what sort of bioengineering strategy you want to take.

Last year you earned a Gordon Bell Prize after breaking the exascale barrier with a code that allows you to study combinatorial interactions between organisms and their environments. How does this algorithm fit into that research?

Jacobson: We're still using the model we used last year, but now, we've introduced this AI-driven genomic selection algorithm to our Combinatorial Metrics [CoMet] code and we're feeding it environmental information across every day of a year, so we can do genome-wide association studies across climate time.

Additionally, we've expanded to a global scale our efforts in what we're calling "climatypes"--the climate and environmental information that plants are adapting to. With the help of ORNL's Peter Thornton and his group's expertise in biogeography and climate, we built models of every square kilometer of land on the planet and encoded 50 years of environmental and climate data into these models, ranging from the soil, up through light spectral quality, and everything in between.

To understand all the relationships between different environments, we compared these environments to each other on Summit using a new algorithm called Duo that we added to our CoMet code base. To our knowledge, this is the largest scientific calculation ever done.

That sounds like a pretty hefty accomplishment. What kind of information can these comparisons give you?

Jacobson: These comparisons can help us determine exactly where we can target certain environments and what gene mutations and alleles we need to include to help these plants adapt to different environments. We can look at an environment and say, "For this environment, this is what we're going to need to have in this plant's genome for it to thrive as well as it can."

Is this the future of environmental genomics?

Jacobson: The integrated vision that we see is the connection of all the "-omics" layers, from genomics (gene expression), proteomics (protein expression), and metabolomics (metabolite expression) all the way up through phenotypes--observable traits; so, from genome to phenome and everything in between.

Ideally, we'd like to have a combination of genotype data with climate and environmental data in an integrated model, from single nucleotides--the molecular structures that make up DNA--up to environment and climate at the planetary scale. Such comprehensive integrated models are now possible because we've actually calculated the light spectral scale of every point on the planet--that's an astrophysical phenotype that comes from our nearest star, the Sun.

Jacobson and his team study plants to understand the genetic variables and patterns that make them adaptable to changing environments and climates. Image Credit: Jason Richards, ORNL

First, we need to look at the combinatorial interactions in such models to see how they lead to the emergent properties that we're trying to optimize in plants for future productivity and sustainability. Then, we can connect that with how plants have historically adapted to environments in order to design new ideal genotypes for bioenergy or food production that are optimized to thrive in specific environments.

Is this something that will be required in agriculture in the future?

Jacobson: As the world changes, there is increasing pressure to utilize "marginal land," that is land that's often not currently used for agriculture or isn't efficiently used for agriculture. So, if we design genotypes that thrive in these marginal environments, we'll be able to increase our food production in addition to our energy production. This is a dual-use technology.

We're also really concerned about overfertilization of the land because it can lead to runoff that has large ecological consequences. If we can optimize plants to use the nutrients that are there with little additional fertilizer, that's a big benefit for sustainability as well. So, we're really trying to look at this holistically and build as much of these adaptations as we possibly can in the model so that we'll know the effects in certain environments.

What are you working on next?

Jacobson: The next step is to look at the historical data and all these relationships and then project forward so that we can actually design genotypes that will not only thrive in the current environmental zones but continue to thrive in the future as the global network changes. The ability to project forward, both for annual crops as well as long-term perennial crops, is really important.

What are some remaining challenges?

Jacobson: Everything we're doing is a heavy lift, but we're looking at how we can design this new approach on Summit and the OLCF's future exascale system, Frontier, so that we can really understand all these relationships. Also, now that we have this data at all these "-omics" layers, we have to run these combinations of layers--called polytopes--thousands or tens of thousands or hundreds of thousands of times. The next set of algorithms we're building is to find all possible relationships and associations within and across all polytopes. That's the next frontier.

Will your work intersect with traditional climate simulation models at all?

Jacobson: This is a data- and AI-driven view of climate information, which is different from a simulation approach. Over time, it will be interesting to see where they intersect, and there may be things we learn here that are very informative for climate models and vice versa. We also know that this same sort of explainable AI technology can help out a lot with simulation studies. Ideally, we could develop explainable AI-driven models that can help simulation models with some of their bottlenecks. If we can learn the patterns simulation models use and replace some of their bottlenecks with a learned outcome, then those models can do more creative things. That's really where we might see some of this space intersecting in the future.

This work was funded by the Center for Bioenergy Innovation, a DOE Bioenergy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science, and by DOE's Laboratory Directed Research and Development program (ORNL AI Initiative ProjectID 9613) at ORNL. This work used OLCF resources.

Related Publication: A. L. Harfouche, D. A. Jacobson, D. Kainer, J. C. Romero, A. H. Harfouche, G. S. Mugnozza, M. Moshelion, G. A. Tuskan, J. J. B. Keurentjes, and A. Altman, "Accelerating Climate Resilient Plant Breeding by Applying Next-Generation Artificial Intelligence." Trends in Biotechnology (2019), doi:10.1016/j.tibtech.2019.05.007.