Do androids dream of electric sorhgum?*

Predicting Phenotype from Multi-Scale Genomic and Environment Data using Neural Networks and Knowledge Graphs

Ryan P Bartelme† David S LeBauer‡ Tyson Lee Swetnam§

March 20, 2020

Introduction

During this period of unprecedented anthropogenic climate change, understanding genomic response to environmental variation and the influences on organismal phenotypes is critical. Predicting these responses is invaluable to maintain the innumerable services natural ecosystems provide. Multiscale effects over time and space combined with the non-linearity of natural systems Lorenz [1963], Ruel and Ayres [1999], West et al. [2009] obscures the signal of biological processes, their interactions with the environment, and resulting observable phenotypes.

Existing models for predicting phenotypes from genetic and environmental data focus on single species or single ecosystems. As we enter an era of Big Data in ecological research Balch et al. [2020], ecologists are shifting away from relatively small data sets toward national and global efforts such as the National Ecological Observatory Network (NEON) Keller et al. [2008] and airborne and space-based Earth Observation Systems (EOS). Both societal need and technical capabilities are moving toward addressing larger scale questions that require integration of multi-modal data and non-linear predictive models to understand the interactions between an organism's genetic potential and its environment and how those interactions result in observable phenotypes.

Challenges Dataset Interoperability

Incorporating genomic data into phenomics is challenging. Many recent studies have used only environmental features and machine learning to predict phenotypes of lilacs, honeysuckle, rice, and wheat Alderman and Stanfill [2017], Nissanka et al. [2015], Poor, 2019). There are a number of methods linking genomic data to environments or traits. For example, genome wide association studies (GWAS) enable researchers to examine the influence of single nucleotide polymorphisms (SNP) on phenotypes in both natural and controlled settings Beyer et al. [2019], Schläppi et al.

*Replication files are available on the corresponding author's Github account

†Corresponding Author rbartelme@arizona.edu, University of Arizona

‡University of Arizona

§University of Arizona

[2017], Spindel et al. [2016]. GWAS often provides generalized and mixed linear model associations between SNPs and environmental variables, roughly analogous to Genes + Environment = Phenotype (G+E=P). There are limitations in the assumptions made by existing methodologies, such as GWAS, that directly attribute plant phenotypes to environmental variables. These methods do not explicitly incorporate biological and molecular interactions, such as post-translational modification of macromolecules Running [2014], the importance of plant-microbe interactions Oyserman et al. [2019], or endogenous siRNA Katiyar-Agarwal et al. [2006]. However, our machine learning approach allows for these biological phenomena to be accounted for as latent variables while probing the interactions of Genomes+Environments=Phenomes in a multidimensional manner.

Conventional observations and statistical models are shifting toward remotely sensed observation and trait collection, which rely on machine learning (ML) and computer vision for measurements. For example, Bayesian Belief Networks Cooper [1990] may be implemented to associate environmental variables with traits, and Generative Adversarial Networks Radford et al. [2015] for classifying plant phenotype imaging. Rather than simply generating large quantities of machine readable data Hampton et al. [2013] and implementing ML methods ad-hoc Pichler et al. [2020], ecologists are now grappling with how to interpret the massive quantities of unstructured data that are available at scale.
ML predictions often rely on complex, "black box" methodologies to assess explanatory variables. Here we introduce the GenoPhenoEnvo project, an effort to predict phenotype from genetic and environmental data, while developing novel representations of the ML "black-box" internals.

Future Directions

Our project has the long-term goal of developing predictive analytics based on an organisms' genetic code and its associated phenomic response to environmental change. To design an initial analytical framework and workflow, we will first use phenomic, genomic, and environmental data about sorghum (Sorghum bicolor). These data are available through the TERRA-REF (Transportation Energy Resources from Renewable Agriculture Phenotyping Reference Platform) project LeBauer et al. [2020] Burnette et al. [2018]. After training the ML model on the highly controlled and thorough TERRA-REF data set, we aim to test and further develop the model with data from less controlled and lower resolution environments using data from sources such as NEON and the Global Phenology Network.

The first challenge is to prepare these data for use as input into ML models. In addition to empirical data, ontologically-supported knowledge graphs can be used to inform the ML Mungall et al. [2017]. Knowledge graphs (KG) are directed acyclic graphs that represent knowledge in a computational format and are an integral part of Google's Answer Box and IBM's Watson. KGs can help constrain and prioritize results, provide quality control, fill in data gaps with inferencing, and integrate heterogeneous data. In this project, KGs provide an easier way to integrate data from established plant phenome databases such as Planteome Cooper et al. [2018], Gramene (Jaiswal, 2011 in Pereira [2011]), and TAIR (Poole, 2007). The true power of ontologies and KGs is a formal logical structure, enabling inferential and similarity analyses Mungall et al. [2017], Washington et al. [2009]. In particular, it is this latter feature that will enable data set interoperability. As we begin to make predictions in more complicated systems with multiple species and heterogeneous data, the knowledge graphs will be critical for managing phenotype data.

The GenoPhenoEnvo (GPE) project aims to predict phenotype with genomic and environmental data using a multimodal approach to training ML models. We are actively developing a visualization tool to increase understanding of why the model gave a particular result. In this way, the GPE project will work toward phenotype predictions and an increased understanding in the biological and molecular processes that translate genotype to phenotype. In addition to increased awareness of molecular effects, the ML models could enable specific ecological hypothesis testing or predicting long-term speciation events driven by environmental factors. Ultimately, we believe this combination of methods will generate a new scientific subdiscipline, one we have called "Computational Ecogenomics."

Author contributions

The ordering of authors following RB is alphabetical.

References

Phillip D. Alderman and Bryan Stanfill. Quantifying model-structure- and parameter-driven uncertainties in spring wheat phenology prediction with bayesian analysis. *European Journal of Agronomy*, 88:1 – 9, 2017. ISSN 1161-0301. doi: https://doi.org/10.1016/j.eja.2016.09.016. URL http://www.sciencedirect.com/science/article/pii/S1161030116301800. Uncertainty in crop model predictions.

Jennifer K Balch, R Chelsea Nagy, and Benjamin S Halpern. Neon is seeding the next revolution in ecology. *Frontiers in Ecology and the Environment*, 18(1):3–3, 2020.

Savannah Beyer, Sintayehu Daba, Priyanka Tyagi, Harold Bockelman, Gina Brown-Guedira, Mohsen Mohammadi, et al. Loci and candidate genes controlling root traits in wheat seedlings—a wheat root gwas. *Functional & integrative genomics*, 19(1):91–107, 2019.

Maxwell Burnette, Rob Kooper, JD Maloney, Gareth S Rohde, Jeffrey A Terstriep, Craig Willis, Noah Fahlgren, Todd Mockler, Maria Newcomb, Vasit Sagan, Nadia Shakoor, Sidke Paheding, Rick Ward, and David LeBauer. Terra-ref data processing infrastructure. In *Proceedings of the Practice and Experience on Advanced Research Computing*, pages 1–7. 2018.

Gregory F Cooper. The computational complexity of probabilistic inference using bayesian belief networks. *Artificial intelligence*, 42(2-3):393–405, 1990.

Laurel Cooper, Austin Meier, Marie-Angélique Laporte, Justin L Elser, Chris Mungall, Brandon T Sinn, Dario Cavaliere, Seth Carbon, Nathan A Dunn, Barry Smith, et al. The planteome database: an integrated resource for reference ontologies, plant genomics and phenomics. *Nucleic acids research*, 46(D1):D1168–D1180, 2018.

Stephanie E Hampton, Carly A Strasser, Joshua J Tewksbury, Wendy K Gram, Amber E Budden, Archer L Batcheller, Clifford S Duke, and John H Porter. Big data and the future of ecology. *Frontiers in Ecology and the Environment*, 11(3):156–162, 2013.

Surekha Katiyar-Agarwal, Rebekah Morgan, Douglas Dahlbeck, Omar Borsani, Andy Villegas, Jian-Kang Zhu, Brian J Staskawicz, and Hailing Jin. A pathogen-inducible endogenous sirna in plant immunity. *Proceedings of the National Academy of Sciences*, 103(47):18002–18007, 2006.

Michael Keller, David S Schimel, William W Hargrove, and Forrest M Hoffman. A continental strategy for the national ecological observatory network. *Frontiers in Ecology and the Environment*, 6(5):282–284, 2008.

David S. LeBauer, Max A Burnette, Rob Kooper, Craig Willis, Pedro Andrade-Sanchez, Noah Fahlgren, Zongyang Li, Stewart Marshall, Geoff Morris, Todd Mockler, Maria Newcomb, Robert Pless, Nadia Shakoor, Rick Ward, Jeff White, and Many Many Others. 'data from: Terra ref, an open reference data set from high resolution genomics, phenomics, and imaging sensors', 2020.

Edward N Lorenz. Deterministic nonperiodic flow, journal of the atmospheric sciences, vol. 20, no, 1963.

Christopher J Mungall, Julie A McMurry, Sebastian Köhler, James P Balhoff, Charles Borromeo, Matthew Brush, Seth Carbon, Tom Conlin, Nathan Dunn, Mark Engelstad, et al. The monarch initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic acids research*, 45(D1):D712–D722, 2017.

Sarath P Nissanka, Asha S Karunaratne, Ruchika Perera, WMW Weerakoon, Peter J Thorburn, and Daniel Wallach. Calibration of the phenology sub-model of apsim-oryza: going beyond goodness of fit. *Environmental Modelling & Software*, 70:128–137, 2015.

Ben O Oyserman, Viviane Cordovez, Stalin W Sarango Flores, Harm Nijveen, Marnix H Medema, and Jos M Raaijmakers. Extracting the gems: Genotype, environment and microbiome interactions shaping host phenotypes. *bioRxiv*, page 863399, 2019.

Andy Pereira. *Plant reverse genetics: methods and protocols*. Springer, 2011.

Maximilian Pichler, Virginie Boreux, Alexandra-Maria Klein, Matthias Schleuning, and Florian Hartig. Machine learning algorithms to infer trait-matching and predict species interactions in ecological networks. *Methods in Ecology and Evolution*, 11(2):281–293, 2020.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Jonathan J Ruel and Matthew P Ayres. Jensen's inequality predicts effects of environmental variation. *Trends in Ecology & Evolution*, 14(9):361–366, 1999.

Mark Paul Running. The role of lipid post–translational modification in plant developmental processes. *Frontiers in plant science*, 5:50, 2014.

Michael R Schläppi, Aaron K Jackson, Georgia C Eizenga, Aiju Wang, Chengcai Chu, Yao Shi, Naoki Shimoyama, and Debbie L Boykin. Assessment of five chilling tolerance traits and gwas mapping in rice using the usda mini-core collection. *Frontiers in plant science*, 8:957, 2017.

JE Spindel, H Begum, D Akdemir, B Collard, E Redoña, JL Jannink, and S McCouch. Genome-wide prediction models that incorporate de novo gwas are a powerful new tool for tropical rice improvement. *Heredity*, 116(4):395–408, 2016.

Nicole L Washington, Melissa A Haendel, Christopher J Mungall, Michael Ashburner, Monte Westerfield, and Suzanna E Lewis. Linking human diseases to animal models using ontology-based phenotype annotation. *PLoS biology*, 7(11), 2009.

Geoffrey B West, Brian J Enquist, and James H Brown. A general quantitative theory of forest structure and dynamics. *Proceedings of the National Academy of Sciences*, 106(17):7040–7045, 2009.