Do androids dream of electric sorhgum?: Predicting Phenotype from Multi-Scale Genomic and Environment Data using Neural Networks and Knowledge Graphs *

Ryan P Bartelme *University of Arizona*

David S LeBauer *University of Arizona*

Tyson L Swetnam *University of Arizona*

Michael Behrisch *Utrecht University*

Emily J Cain *Univerisity of Arizona*

Ishita Debnath *Michigan State University*

Pankaj Jaiswal *Oregon State University*

Ab Mosca *Tufts University*

Monica Munoz-Torres *Oregon State University*

Kent Shefchek *Oregon State University*

P. Bryan Heidorn *University of Arizona*

Remco Chang *Tufts University*

Arun Ross *Michigan State University*

Anne E Thessen *Oregon State University*

The interplay between an organism's genes, its environment, and the expressed phenotype is dynamic. These interactions within ecosystems are shaped by non-linear multiscale effects that are difficult to disentangle into discrete components. Advances in the fields of ontolgoies, unsuper-

*Corresponding Author: Ryan Bartelme rbartelme@arizona.edu

vised learning, and genomics are currently disparate. The interoperability of these different data schema is difficult. However, here we present a framework utilizing phenotypic data, knoweldge graphing, and deep learning, to better link phenotypes, environments, and genotypes.

Keywords: machine learning, genomics, phenotype

Introduction

Unprecedented anthropogenic climate change necessitates us to have the ability to adapt crops and modify ecosystems. Understanding genomic responses of plants and animals to environmental variation allows prediction [1, 2]. Environmental factors that influence organismal phenotype, fecundity, morbidity, and mortality in turn affect agricultural and natural ecosystems (Fig. 1). However, multi-scale effects over time and space combined with the non-linearity of natural systems [3–5] obscures the signal of biological processes, interactions with the environment, and the resulting (observable) phenotypes. Maintaining the innumerable benefits and services agronomic and natural systems provides is therefore critical to our survival and prosperity.

Existing models for predicting phenotypes from genetic and environmental data focus on single species or single ecosystems. However, both societal need and technical capabilities are moving toward addressing larger scale questions that require integration of multi-modal data. In addition to relatively small and heterogeneous data sets, researchers are relying on national and global data collection efforts [6] such as the National Ecological Observatory Network (NEON) [7] and airborne and space-based Earth Observation Systems (EOS). These efforts produce large quantities of homogeneous "born-digital" data, but a significant interdisciplinary data-integration task remains. Ontologies and knowledge graphs using semantic similarity to cope with problems of granularity and terminology (e.g., [8, 9]) are now available to facilitate data integration at scales beyond a single taxon or single ecosystem. In addition, as the data sets and questions become more complicated, more non-linear predictive models are needed to address them.

Challenges of 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 [10–12]. 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 [13–15]. 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 [16], the importance of plant-microbe interactions [17], or endogenous siRNA [18]. However, a machine learning approach allows for these biological phenomena to be accounted for as latent variables while probing the interactions of genomes, environments, and phenotypes 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 [19] may be implemented to associate environmental variables with traits, and Generative Adversarial Networks [20] for classifying plant phenotype imaging. Rather than simply generating large quantities of machine readable data [21] and implementing ML methods ad-hoc [22], ecologists are now grappling with how to interpret the massive quantities of unstructured data that are available at scale. Unfortunately, the ML that provides a scalable, non-linear method for using these data, relies on complex, "black box" methodologies to assess explanatory variables, which does not aid interpretation. 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 phenotypic 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 [23, 24]. 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 National 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 [25]. 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 [26], Gramene [27], and TAIR [28]. The true power of ontologies and KGs is a formal logical structure, enabling inferential and similarity analyses [25, 29]. 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 sub-discipline, one we have called "*Computational Ecogenomics*."

Author contributions

The ordering of authors following RB is alphabetical.

RPB: Conceptualization, Writing-original draft DSL: Conceptualization, Funding Acquisition, Writing-review&editing TLS: Conceptualization, Funding Acquisition, Writing-review&editing MB: Conceptualization EC: Conceptualization ID: Conceptualization PJ: Conceptualization, Funding Acquisition AM: Conceptualization MMT: Conceptualization, Funding Acquisition, Project Administration KS: Conceptualization PBH: Conceptualization, Funding Acquisition RC: Conceptualization, Funding Acquisition AR: Conceptualization, Funding Acquisition AET: Conceptualization, Funding Acquisition, Supervision, Writing-review&editing

References

[1] Ungerer, M. C., L. C. Johnson, and M. A. Herman (2008). Ecological genomics: understanding gene and genome function in the natural environment. *Heredity 100*(2), 178–183.

[2] Des Marais, D. L., K. M. Hernandez, and T. E. Juenger (2013). Genotype-by-environment interaction and plasticity: exploring genomic responses of plants to the abiotic environment. *Annual Review of Ecology, Evolution, and Systematics 44*, 5–29.

[3] Lorenz, E. (1963). Deterministic Nonperiodic Flow. *Journal of the Atmospheric Sciences 20*, 130–141.

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

[5] West, G. B., B. J. Enquist, and J. H. Brown (2009). A general quantitative theory of forest structure and dynamics. *Proceedings of the National Academy of Sciences 106*(17), 7040–7045.

[6] Balch, J. K., R. C. Nagy, and B. S. Halpern (2020). Neon is seeding the next revolution in ecology. *Frontiers in Ecology and the Environment 18*(1), 3–3.

[7] Keller, M., D. S. Schimel, W. W. Hargrove, and F. M. Hoffman (2008). A continental strategy for the national ecological observatory network. *Frontiers in Ecology and the Environment 6*(5), 282–284.

[8] Mungall, C. J., G. V. Gkoutos, C. L. Smith, M. A. Haendel, S. E. Lewis, and M. Ashburner (2010). Integrating phenotype ontologies across multiple species. *Genome Biology 11*.

[9] Stucky, B. J., R. Guralnick, J. Deck, E. G. Denny, K. Bolmgren, and R. Walls (2018). The Plant Phenology Ontology: A New Informatics Resource for Large-Scale Integration of Plant Phenology Data. *Frontiers in Plant Science 9*.

[10] Alderman, P. D. and B. Stanfill (2017). Quantifying model-structure- and parameter-driven uncertainties in spring wheat phenology prediction with bayesian analysis. *European Journal of Agronomy 88*, 1–9.

[11] Nissanka, S. P., A. S. Karunaratne, R. Perera, W. Weerakoon, P. J. Thorburn, and D. Wallach (2015). Calibration of the phenology sub-model of apsim-oryza: going beyond goodness of fit. *Environmental Modelling & Software 70*, 128–137.

[12] Mehdipoor, H. (2019). *Geocomputational Workflows for Analysing Spring Plant Phenology in Space and Time*. Ph. D. thesis.

[13] Beyer, S., S. Daba, P. Tyagi, H. Bockelman, G. Brown-Guedira, M. Mohammadi, et al. (2019). Loci and candidate genes controlling root traits in wheat seedlings—a wheat root gwas. *Functional & integrative genomics 19*(1), 91–107.

[14] Schläppi, M. R., A. K. Jackson, G. C. Eizenga, A. Wang, C. Chu, Y. Shi, N. Shimoyama, and D. L. Boykin (2017). Assessment of five chilling tolerance traits and gwas mapping in rice using the usda mini-core collection. *Frontiers in plant science 8*, 957.

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

[16] Running, M. P. (2014). The role of lipid post–translational modification in plant developmental processes. *Frontiers in plant science 5*.

[17] Oyserman, B. O., V. Cordovez, S. W. S. Flores, H. Nijveen, M. H. Medema, and J. M. Raaijmakers (2019). Extracting the gems: Genotype, environment and microbiome interactions shaping host phenotypes. *bioRxiv*.

[18] Katiyar-Agarwal, S., R. Morgan, D. Dahlbeck, O. Borsani, A. Villegas, J.-K. Zhu, B. J. Staskawicz, and H. Jin (2006). A pathogen-inducible endogenous sirna in plant immunity. *Proceedings of the National Academy of Sciences 103*(47), 18002–18007.

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

[20] Radford, A., L. Metz, and S. Chintala (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

[21] Hampton, S. E., C. A. Strasser, J. J. Tewksbury, W. K. Gram, A. E. Budden, A. L. Batcheller, C. S. Duke, and J. H. Porter (2013). Big data and the future of ecology. *Frontiers in Ecology and the Environment 11*(3), 156–162.

[22] Pichler, M., V. Boreux, A.-M. Klein, M. Schleuning, and F. Hartig (2020). Machine learning algorithms to infer trait-matching and predict species interactions in ecological networks. *Methods in Ecology and Evolution 11*(2), 281–293.

[23] LeBauer, D. S., M. A. Burnette, R. Kooper, C. Willis, P. Andrade-Sanchez, N. Fahlgren, Z. Li, S. Marshall, G. Morris, T. Mockler, M. Newcomb, R. Pless, N. Shakoor, R. Ward, J. White, and M. M. Others (2020). 'data from: Terra ref, an open reference data set from high resolution genomics, phenomics, and imaging sensors'.

[24] Burnette, M., R. Kooper, J. Maloney, G. S. Rohde, J. A. Terstriep, C. Willis, N. Fahlgren, T. Mockler, M. Newcomb, V. Sagan, N. Shakoor, S. Paheding, R. Ward, and D. LeBauer (2018). Terra-ref data processing infrastructure. In *Proceedings of the Practice and Experience on Advanced Research Computing*, pp. 1–7.

[25] Mungall, C. J., J. A. McMurry, S. Köhler, J. P. Balhoff, C. Borromeo, M. Brush, S. Carbon, T. Conlin, N. Dunn, M. Engelstad, et al. (2017). The monarch initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic acids research 45*(D1), D712–D722.

[26] Cooper, L., A. Meier, M.-A. Laporte, J. L. Elser, C. Mungall, B. T. Sinn, D. Cavaliere, S. Carbon, N. A. Dunn, B. Smith, et al. (2018). The planteome database: an integrated resource for reference ontologies, plant genomics and phenomics. *Nucleic acids research 46*(D1), D1168–D1180.

[27] Jaiswal, P. (2011). Gramene Database: A Hub for Comparative Plant Genomics. In A. Pereira (Ed.), *Plant Reverse Genetics: Methods and Protocols*, Methods in Molecular Biology, pp. 247–275. Totowa, NJ: Humana Press.

[28] Poole, R. L. (2007). The TAIR Database. In D. Edwards (Ed.), *Plant Bioinformatics: Methods and Protocols*, Methods in Molecular Biology™, pp. 179–212. Totowa, NJ: Humana Press.

[29] Washington, N. L., M. A. Haendel, C. J. Mungall, M. Ashburner, M. Westerfield, and S. E. Lewis (2009). Linking human diseases to animal models using ontology-based phenotype annotation. *PLoS biology 7*(11).