2021

Representing, comparing, and querying phenotype descriptions in plants using computational methods

Ian Braun
Iowa State University

Recommended Citation

Braun, Ian, "Representing, comparing, and querying phenotype descriptions in plants using computational methods" (2021). *Graduate Theses and Dissertations*. 18462.
https://lib.dr.iastate.edu/etd/18462

Representing, comparing, and querying phenotype descriptions in plants using computational methods

by

Ian Robert Braun

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Bioinformatics and Computational Biology (Predictive Plant Phenomics)

Program of Study Committee:
Carolyn J. Lawrence-Dill, Major Professor
Iddo Friedberg
Marna Yandeau-Nelson
Baskar Ganapathysubramanian
Qi Li

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation. The Graduate College will ensure this dissertation is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2021

DEDICATION

This dissertation is dedicated to my incredible and supportive wife Allyson and to four amazing teachers: Sarah, Katie, Mary Jo, and Jim.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

ACKNOWLEDGMENTS

I would first like to thank my major professor and mentor Dr. Carolyn Lawrence-Dill, for all the support, guidance, and encouragement throughout the course of my studies. Your enthusiasm and willingness to help students including myself find their way through graduate school, research, and life has benefited me in so many ways, starting from my very first week in Ames. Being a part of your lab group has led me to so many incredible opportunities, from presenting at international conferences and honing my science communication skills to meeting and working with scientists with all kinds of expertise. Thank you for always giving me room to explore new research directions whether they were successful or not, always taking the time to collaborate on moving projects forwards, and also helping me to continue tackling new challenges. I'm incredibly grateful to have been able to do my PhD in the lab group of someone who places such a high level of importance on mentorship, and I can't thank you enough.

I also want to thank all the faculty members that served on my committee. Thank you Dr. Marna Yandeau-Nelson, Dr. Iddo Friedberg, Dr. Baskar Ganapathysubramanian, Dr. Qi Li, Dr. Annette O'Connor, and Dr. Sowmya Vajjala for your help, guidance, and feedback, as my projects continued to evolve.

I also want to especially thank all the other members of Dill-PICL and the BCB program at ISU who have made my time in Ames so rich, valuable, and fun. Thank you Darwin Campbell for your help with the countless issues and challenges you assisted me with over the years, no matter the topic. I admire your work ethic, breadth of knowledge, attention to detail, and sense of humor, and I've learned a lot from you. Thank you Scott Zarecor for all your assistance and for fielding all my questions. Having someone with your expertise as a part of the lab was incredibly valuable to me as a student. Thank you Colleen Yanarella and Leila Fattel for continuing to make the lab be such a wonderful place to do research, and for all the helpful discussions. Thank you to Carla

Mann, Gokul Wimalanathan, Ming He, Schuyler Smith, Naser Alkhalifah, and Paul Villanueva for all the fun times; I'm incredibly grateful that my time at ISU overlapped with all of yours, making it such a great place to work and live. Thank you Gokul for being there to help me hit the ground running as a graduate student, and helping me get started in the world of ontologies, and Paul for always encouraging me to keep learning.

ABSTRACT

Plant phenotype descriptions are abundant both in literature and in community datastores. Enabling basic aggregation, organization, and analyses over this data requires that phenotype descriptions be represented in a computable format. One successful approach to this challenge has been to develop standardized vocabularies and biological ontologies that can be used to annotate phenotypes, allowing for the sparsity of the data to be reduced by inferring implicit information about the annotated data, and enabling simple quantification of similarity between annotated data. This type of structured curation has shown promise for enabling dataset-wide analyses on plant phenotype descriptions, but the time and effort required for curation of individual phenotype descriptions is a limiting factor in how scalable this approach is in light of the increasing volume of available text data related to plant phenotypes. Computational approaches have the potential to alleviate this problem by providing methods for representing phenotype descriptions and allowing quantification of phenotype similarity. In this work, computational pipelines for representing and comparing phenotypes are presented, and evaluated for their ability to predict biological relationships between genes. Approaches from the natural language processing domain perform as well as similarity metrics over curated annotations for predicting shared phenotypes. These approaches also show promise both for helping curators organize large datasets as well as for enabling researchers to explore relationships among available phenotype descriptions. A web application for querying datasets of plant phenotype descriptions and identifying associated genes is also presented, and example use cases are discussed.

CHAPTER 1. GENERAL INTRODUCTION

1.1 Introduction

Phenotypes can be defined as observable characteristics of living things that result from the combination of, and interactions between, genetics and environment. Phenotypes accounts for an incredibly diverse set of properties, including those crucial to understanding important biological systems like diseases in humans, and agriculturally crucial traits like plant biomass and crop species' ability to resist attack by pathogens. Phenotypes can also be both described in a quantitative sense (e.g., height is x meters) or qualitatively (e.g., the mutant line is shorter than the wildtype). The range of information encompassed by phenotypes makes it a challenge to organize this information in a way that is broadly usable for large-scale analyses, especially with respect to already-existing data. This is not necessarily a problem for focused studies, such as a genome-wide association study (GWAS) to identify loci related to one phenotype (e.g., plant height) that was measured for the purpose of that study, but it becomes a profound challenge when trying to provide phenomic data in a generally accessible manner, where the ways in which the data will be used are not immediately known at the point that it is recorded. Being able to compare phenotypes with one another in a large-scale way that is repeatable across different datasets and over different types of analyses is the primary challenge that results from the fact that phenotypic data is so diverse in how it is collected, described, curated, and stored.

Biological ontologies have been crucially important towards the goal of representing biological data in standardized ways, not just for phenotypes specifically, but also chemical and compound names, cellular components, gene functions, crop traits, and other types of biological entities (Ashburner et al. (2000); Hastings et al. (2012); Cooper et al. (2013); Cooper et al. (2018); Gkoutos et al. (2005)) The hierarchical nature of ontologies allows for the reduction in sparsity of datasets by allowing a single annotation to inherit other related information (e.g., annotating a phenotype with the term *leaf senescence* implies that this phenotype is also related more generally to *aging*). The shared use of common ontologies across studies and datasets allows for observations or predictions made in one time and location to be used or referenced later by others in a meaningful and unambiguous way.

In the case of plant phenotypes in particular, curating phenotype descriptions with ontology terms has shown promise in that these annotations can be used to generate similarity values between genes that are representative of known biological relationships, such as orthology or membership in a biochemical pathway (Oellrich et al. (2015)). However, the reliance on curators to produce these annotations from descriptions of phenotypes is a limiting factor, given the volume of phenotype descriptions available. This fact emphasizes the importance of finding computational solutions for representing and comparing text descriptions of phenotypes, and the importance of understanding what biological relationships can or cannot be reliably captured using these representations.

The work described in this dissertation is focused on describing this problem, describing computational approaches that address this problem with and without using biological ontologies, and characterizing these approaches in terms of where they are effective and where they are not. This work builds on the work of many others, especially community databases and research efforts that have organized or categorized phenotype descriptions of plants and annotations of those data, which provides both the datasets used here to develop computational pipelines, and also curated representations as a point of comparison in the analyses presented (notably including the work of Oellrich et al. (2015); Lloyd and Meinke (2012); Berardini et al. (2015); Portwood et al. (2019); Fernandez-Pozo et al. (2015); Cooper et al. (2018)). The computational pipelines presented also build on the work of many others in the fields of semantic annotation, text mining, natural language processing, machine learning, and artificial intelligence with respect to methods for representing and modelling text in ways that can be used for comparing phenotype descriptions and making predictions about gene relatedness (notably including the work of Hoehndorf et al. (2011); Mikolov et al. (2013); Le and Mikolov (2014); Tseytlin et al. (2016)).

1.2 Research Goals

The first research goal was to determine if the work of Oellrich et al. (2015) in producing a dataset of phenotype annotations for comparing phenotypes to one another could be reproduced using a computational pipeline rather than curation. Work towards this goal involved developing a computational pipeline for assigning annotations to text, evaluating the ability of this pipeline to produce similarity values between plant genes that were useful in recovering known biological relationships in the same manner that was explored in the original work (Oellrich et al. (2015)), and comparing against simple natural language processing approaches for representing and comparing text.

The second research goal was to determine to what extent natural language processing methods can be used to produce similarity values between plant genes that are reflective of biological relationships in a generalized way, rather than with specific examples from a limited set of plant genes or particular orthologous genes or pathways. Work towards this goal involved analyzing the similarity values produced by a diverse set of natural language processing approaches for representing and comparing text with respect to a number of existing resources about gene relationships in terms of orthology, protein interactions, pathway membership, and phenotypic category.

The third research goal was to provide researchers with a webtool that leverages the finding that direct computation on natural language phenotype descriptions is useful for representing and querying phenotypes so that it can be utilized to identify genes or groups of genes matching particular descriptions. Work towards this goal involved developing the application itself, and describing its utility with respect to the second research goal and specific examples of queries that return related genes.

1.3 Dissertation Organization

Following this general introduction (Chapter 1), Chapters 2-6 contain either reformatted published works or papers in preparation. Chapter 2 is a published book chapter that contains a general discussion on the utility of comparing phenotypes within and across species, and introduces how computational approaches both have already and could in the future continue to improve the scope and utility of these types of comparisons and analyses. Chapter 3 is a brief conference proceedings publication that discusses progress on using computational approaches to annotate phenotype descriptions with ontology terms. Chapter 4 is a published research article presenting a pipeline for annotating phenotype descriptions with ontology terms, as well as comparing phenotypes using simple NLP approaches. The utility of these methods in comparing phenotypes both within and across species in comparison to using curated datasets for these tasks is explored. Chapter 5 is a published perspectives paper discussing the potential for computational methods to enable phenotype comparison, and the potential to use these approaches for an alternative to traditional definitions of traits as an input to genome-wide association studies. Chapter 6 is a researcher paper in preparation, which builds on the results of the paper presented in Chapter 4 to more fully explore how generalizable these results are to a larger dataset of phenotype descriptions, a broad range of methods for representing and comparing text, and what types of biological relationships can be predicted from the assessed phenotype similarities. Chapter 7 is a brief conclusion to this dissertation, including a summary of the findings and a discussion of future research in this field.

1.4 References

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. (2000). Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29.

Berardini, T. Z., Reiser, L., Li, D., Mezheritsky, Y., Muller, R., Strait, E., and Huala, E. (2015). The arabidopsis information resource: making and mining the “gold standard” annotated reference plant genome. *genesis*, 53(8):474–485.

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

Cooper, L., Walls, R. L., Elser, J., Gandolfo, M. A., Stevenson, D. W., Smith, B., Preece, J., Athreya, B., Mungall, C. J., Rensing, S., et al. (2013). The plant ontology as a tool for comparative plant anatomy and genomic analyses. *Plant and Cell Physiology*, 54(2):e1–e1.

Fernandez-Pozo, N., Menda, N., Edwards, J. D., Saha, S., Tecle, I. Y., Strickler, S. R., Bombarely, A., Fisher-York, T., Pujar, A., Foerster, H., et al. (2015). The sol genomics network (sgn)—from genotype to phenotype to breeding. *Nucleic acids research*, 43(D1):D1036–D1041.

Gkoutos, G. V., Green, E. C., Mallon, A.-M., Hancock, J. M., and Davidson, D. (2005). Using ontologies to describe mouse phenotypes. *Genome biology*, 6(1):R8.

Hastings, J., de Matos, P., Dekker, A., Ennis, M., Harsha, B., Kale, N., Muthukrishnan, V., Owen, G., Turner, S., Williams, M., et al. (2012). The chebi reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic acids research*, 41(D1):D456–D463.

Hoehndorf, R., Schofield, P. N., and Gkoutos, G. V. (2011). Phenomenet: a whole-phenome approach to disease gene discovery. *Nucleic acids research*, 39(18):e119–e119.

Le, Q. V. and Mikolov, T. (2014). Distributed representations of sentences and documents. corr abs/1405.4053 (2014). *arXiv preprint arXiv:1405.4053*.

Lloyd, J. and Meinke, D. (2012). A comprehensive dataset of genes with a loss-of-function mutant phenotype in arabidopsis. *Plant physiology*, 158(3):1115–1129.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Oellrich, A., Walls, R. L., Cannon, E. K., Cannon, S. B., Cooper, L., Gardiner, J., Gkoutos, G. V., Harper, L., He, M., Hoehndorf, R., et al. (2015). An ontology approach to comparative phenomics in plants. *Plant methods*, 11(1):1–15.

Portwood, J. L., Woodhouse, M. R., Cannon, E. K., Gardiner, J. M., Harper, L. C., Schaeffer, M. L., Walsh, J. R., Sen, T. Z., Cho, K. T., Schott, D. A., et al. (2019). Maizegdb 2018: the maize multi-genome genetics and genomics database. *Nucleic acids research*, 47(D1):D1146–D1154.

Tseytlin, E., Mitchell, K., Legowski, E., Corrigan, J., Chavan, G., and Jacobson, R. S. (2016). Noble–flexible concept recognition for large-scale biomedical natural language processing. *BMC bioinformatics*, 17(1):32.

CHAPTER 2. COMPUTABLE PHENOTYPES ENABLE COMPARATIVE AND PREDICTIVE PHENOMICS AMONG PLANT SPECIES AND ACROSS DOMAINS OF LIFE

Ian BRAUN[a], James P. BALHOFF[b**], Tanya Z. BERARDINI[c**], Laurel COOPER[d**], Georgios GKOUTOS[e**], Lisa HARPER[f,g**], Eva HUALA[c**], Pankaj JAISWAL[d**], Toni KAZIC[h**], Hilmar LAPP[i**], James A. MACKLIN[j**], Chelsea D. SPECHT[k**], Todd VISION[l**], Ramona L. WALLS[m**], and Carolyn J. LAWRENCE-DILL[a,n*]

[a]Department of Genetics, Development and Cell Biology and Interdepartmental Bioinformatics and Computational Biology, Iowa State University, Ames, Iowa, USA

[b]Renaissance Computing Institute, University of North Carolina, Chapel Hill, North Carolina, USA

[c]Phoenix Bioinformatics, Redwood City, California, USA

[d]Department of Botany and Plant Pathology, Oregon State University, Corvallis, Oregon, USA

[e]Institute of Cancer and Genomic Sciences, University of Birmingham, Birmingham, UK

[f]USDA-ARS Corn Insects and Crop Genetics Research Unit, Ames, Iowa, USA

[g]USDA-ARS Plant Gene Expression Center, Albany, California, USA

[h]Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, Missouri, USA

[i]Center for Genomic and Computational Biology, Duke University, Durham, North Carolina, USA

[j]Agriculture and Agri-Food Canada, Ottawa, Ontario, Canada

[k]School of Integrative Plant Sciences, Cornell University, Ithaca, New York, USA

[l]Biology Department, University of North Carolina, Chapel Hill, North Carolina, USA

[m]CyVerse, Bio5 Institute, University of Arizona, Tucson, Arizona, USA

[n]Department of Agronomy, Iowa State University, Ames, Iowa, USA

*communication triffid@iastate.edu

**alphabetical

Modified from a chapter published in *Application of Semantic Technology in Biodiversity Science; Anne E. Thessen (Ed.); ISBN 978-3-89838-733-0; Copyright 2018 AKA Verlag Berlin*

2.1 Abstract

Scientists are adept at comparing genomic sequences. The collection of more such data promises to increase our ability to determine gene function, discover and describe biological processes, and prioritize causative variants of interest that underlie disease response. Yet the question remains: Can we compare phenotypes or traits of interest across disciplines in a manner similar to how we compare genomic sequences? Here we present examples of 'semantic reasoning' computational methodologies that enable computation across organized formal phenotypic representations. These methods facilitate the analysis of phenotype information across species, domains of knowledge, people, and computers. We review representative examples of successful semantic reasoning to recover known biological phenomena in medical and agricultural applications. Necessary changes in how we collect, analyze, and share data to enable such computations are presented, and database and analytic tool suites for these sorts of analyses are described.

2.2 Background

Phenotypic variation is the raw material acted on by both natural and artificial selection; it provides the diversity required for species to adapt and respond to changing environments. Linking phenotypes to genotypes across evolutionarily distant lineages provides researchers with the ability to predict phenotypic outcomes of genotypic changes, and to compare genetic strategies that give rise to like phenotypes. This information in turn enables researchers to develop medical and agricultural innovations and to assess and manage the organismal and community-level variation critical to maintaining ecosystem processes and adaptive responses to climate change.

Phenotypic data are extremely diverse, ranging in scope from expression profiles, to quantitative information, to summarizing textual descriptions of development. Data can be associated to individuals, populations, or species and can be described in comparative terms (e.g., mutant versus wild type) or absolute measurements (e.g., days to flowering). The documentation of these data can be at a summary level (e.g., average height of plants studied) or measured (a particular plant is measured to be 62 cm tall). For these reasons and myriad others, the documentation, integration, representation, and accessibility of phenotype data is notoriously challenging (reviewed in Deans et al. (2015)). Adding to the complexity, new high-throughput measurements of phenotype can involve remote sensing, high-density imaging, and integration with geolocation data. Because phenotypic data are so diverse, and the rates, volumes, and complexities of data collection are only increasing, it is difficult to aggregate these complex datasets for downstream analyses (reviewed in Thessen et al. (2015)).

In an effort to leverage the wealth of phenotypic data available for making important biological inferences, McGary et al. (2010) developed a method of candidate gene discovery involving phenologs, or orthologous phenotypes. As defined by the authors, phenotype A from species A and phenotype B from species B are phenologs if their two sets of known causal genes have a significant overlap in the form of orthologous genes. Once phenotypes A and B have been identified as phenologs, the authors' methodology identifies candidate genes as those genes which are known to be causal in one species, but are not currently associated with that phenotype in the other species. If Gene 1 is causal to phenotype A, then its ortholog in species B is a candidate gene for phenotype B. The authors demonstrated the utility of this methodology by discovering non-obvious model systems for human disease phenotypes, predicting specific novel candidate genes associated with those phenotypes, and verifying selected predictions. For example, a significant overlap in orthologous genes revealed a phenolog relationship between mammalian abnormal angiogenesis and reduced rates of growth in lovastatin-treated yeast. A predicted candidate gene for angiogenesis, SOX13 (known to be causal to the yeast reduced growth rate phenotype), was experimentally confirmed through knockdown studies in both mouse and human cells.

The methodology of finding phenologs first proposed by McGary et al. (2010) relies on previously known genotype to phenotype associations and the use of orthology relationships between genes to reveal related phenotypes. Phenologs, however, may also be proposed based solely on the characteristics of the phenotypes themselves, represented in the form of textual or other data. Relying on textual descriptions rather than genotype to phenotype associations to identify phenologs is advantageous in the case of phenotypes for which associated genotypes (causal genes) are not known, or causal genes are involved in similar pathways between the species but are not necessarily orthologous. Example phenolog sets that generate similar phenotypes are kinesin motor proteins that, when mutated, cause trichome branching defects in Arabidopsis and neuronal branching defects in mice (Figure 2.1), and some genes involved in lesion formation in both humans and maize (Figure 2.2).

The lesion phenotypes shown in panels A and B of Figure 2.2 are caused by reduced activity of uroporphyrinogen decarboxylase that leads to the accumulation of uroporphyrin and related metabolites (Burns et al. (2008); Hu et al. (1998); Johal (2007)). The enzymes encoded by the UROD and Les22 genes in humans and maize, respectively are 35% identical and so would be readily discovered by the method of McGary et al. (2010) (GenBank: Accession No. NP_000365.3 and Kazic, unpublished). But approximately 54 other mutations producing lesion phenotypes in maize have been confirmed so far, and for most neither the gene nor the biochemical functions it encodes have been identified (Neuffer and Kazic, unpublished). All of these mutations produce discontiguous patches of chlorotic or necrotic tissue on leaves, often in response to light or developmental cues, and their morphology, behavior, spatial distribution, and time of onset are sensitive to genetic background and environmental perturbations. For example, Figure 2.2 panel C shows a classic oscillatory lesion phenotype produced by Les1 (Neuffer et al. (1975)). Mapping has placed the locus on the short arm of chromosome 2, and no biochemical function has been described (Neuffer and Pawar (1980)). Only searching over phenologs would discover Les1 and its phenotypes. Many other lesion mimic mutants of maize display these types of oscillations under appropriate conditions, and these provide important clues to the underlying causal mechanisms (Kazic, unpublished). Such searches would benefit considerably from also annotating other important dimensions of the phenotype, such as its spatiotemporal oscillation and sensitivity to ambient temperature and genetic background.

Regardless of the genotypic data associated with these phenotypes, they share characteristic morphologies across the species that are readily summarized by images. It is likely that morphological phenologs will eventually be directly discovered by sophisticated combinations of image analysis and pattern recognition techniques that can be used on distributed image databases, though these techniques are only just emerging and will require significant research. However, many dimensions of these phenotypes, and many other non-morphological phenotypes, are not neatly captured by an image. For example, the time of onset of a phenotype and the environmental perturbations that trigger or modify it are far more accurately and compactly expressed as text. Further, the genetic component that produces phenocopies is sharply reduced by definition, but these phenotypes can be particularly revealing: Type I porphyria cutanea tarda and lesion formation in response to pathogen infection both reveal important causal mechanisms. For these non-morphological phenotypes and phenotypic dimensions, computational analysis of phenotypic descriptions will be key to automating discovery of such associations, now and for the foreseeable future.

To enable computational discovery methods, free text descriptions of phenotypes need to be associated with standardized ontology terms which are placed in hierarchical directed acyclic graphs. The analysis of ontology terms and the relationship graph in which they are placed is called semantic reasoning; it is what allows machines to "understand" (reason over) domain knowledge. The graph nature of the terms and relationships in an ontology allows metrics utilizing graph theory to be applied to quantifying similarity between terms, and thus the phenotype descriptions linked to them. Following the identification of phenologs through semantic reasoning, the same method of candidate gene discovery suggested by McGary et al. (2010) can be applied. Multiple research groups are currently using semantic reasoning for novel candidate gene discovery. Hoehndorf et al. (2011) predicted the association between the genes Adam19 and Fgf15 with the Tetralogy of Fallot disease phenotype in humans, one result of the construction of a network of phenotype similarity scores among 86,203 phenotypes across five different species (yeast, fly, worm, mouse, zebrafish) called PhenomeNET (http://phenomebrowser.net/).

In seeking to construct a similar network to facilitate candidate gene discovery in plants, Oellrich et al. (2015) developed and tested a workflow to curate and standardize existing plant phenotype datasets. This approach employed curated data to demonstrate the feasibility of semantic comparison across plant species. Data for six plant species, encompassing both model species and crop plants with established genetic resources, were integrated and analyzed using a common set of ontologies, annotation standards, formats, and best practices. The study focused on mutant phenotypes associated with genes of known sequence in Arabidopsis thaliana (L.) Heynh. (Arabidopsis), Zea mays L. subsp. mays (maize), Medicago truncatula Gaertn. (barrel medic or Medicago), Oryza sativa L. (rice), Glycine max (L.) Merr. (soybean), and Solanum lycopersicum L. (tomato). Curated phenotypes from taxon-specific databases were converted into a common format using the Ontologies for Plant Biology (described in Cooper et al. (2018)) that include Plant Ontology (PO; Cooper et al. (2013)), Gene Ontology (GO; Ashburner et al. (2000)), Plant Experimental Conditions Ontology (PECO; Cooper et al. (2018)), Chemical Entities of Biological Interest (ChEBI; Hastings et al. (2012)) ontology, and Phenotype and Trait Ontology (PATO; Gkoutos et al. (2005)). The ontology annotations were used to construct a matrix of semantic similarity scores for all possible pairs of inter- and intra-specific genotypes. The constructed ontology annotations representing each phenotype are referred to as EQ (Entity-Quality) statements (Mungall et al. (2010)), as they are composed of ontology term(s) representing a biological structure or process (entity), and term(s) representing an aspect or modification of that entity (its quality (Gkoutos et al. (2018))).

From 2,866 genotypes yielding over 8 million possible combinations, 548,888 had non-zero semantic similarity scores. A similarity score of 0 indicates no semantic overlap with respect to the phenotype, while a similarity score of 1 indicates an identical semantic phenotype description (and therefore equivalent sets of EQs). Of these, 44% of the non-zero semantic similarity scores were below 0.1, indicating that many of the phenotypes show only a small overlap in their description, while 13% of the genotype pairs with non-zero scores fell into the 0.9-1 range. This indicates that for most of the genes the semantic similarity of their mutant phenotype descriptions with other genes is low. Some of the very high scores (scores near 1) are likely artifacts due to limited data curation. For example, if only some characteristics of genotypes have been annotated in the form of EQ statements, two genotypes may appear artificially much more similar or dissimilar than they would be had their phenotypes been annotated in full. Furthermore, not all phenotype changes may be reported in the literature for a given genotype in the first place. It is important to note that semantic similarity algorithms cannot compensate for such gaps in reporting or in annotation. Results of the semantic similarity analysis are provided through the Plant PhenomeNET (http://phenomebrowser.net/plant/) web interface, which was adapted from PhenomeNET. For each genotype, a detailed page provides information about similarity scores to any of the other genotypes as well as a link to an additional page providing the phenotype assigned by the curator and those translated to use terms represented in the ontologies.

The semantic similarity dataset was evaluated for its ability to enhance predictions of gene families, protein functions, and shared metabolic pathways that underlie informative plant phenotypes. In one example, Oellrich et al. were able to use Plant PhenomeNET to identify a set of maize gene models that participate in the initial reactions of flavonoid biosynthesis as part of the phenylpropanoid biosynthesis pathway. This result indicates that reasoning across curated phenotypes in plants is capable of recapitulating well-characterized biological phenomena and hints that, for plant species that are not genetically well-characterized, the ontological reasoning approach to predicting phenotypic associations can help with characterizing understudied species and assist in forward genetics approaches.

In a second example, Oellrich et al. (2015) were able to place 2,741 EQ-annotated genes from all six species into 1,895 gene families, of which 42 contain between 5 and 12 genes with EQ statements. These families were assessed for how often homologous genes have similar functions. There were also 147 families containing EQ statements from two or more species, which allowed the authors to assess how often functions are conserved between orthologs. For most families in this sample, gene function was conserved or similar, but there were some cases in which annotated phenotypes were quite different across orthologs.

2.3 Current Efforts

This method of comparing semantic similarity of mutant phenotypes has high potential for semantic prediction, but requires consistent, coherent, and complete phenotype annotations that computationally replicate the underlying biology of organisms, which in turn will require a much larger, more complete dataset. Within the plant kingdom, the Planteome group has begun working toward this goal.

The Planteome project (Cooper et al. (2018); http://www.planteome.org) provides a suite of interconnected reference and species-specific ontologies associated with a database of plant gene expression and function, traits, phenotypes, QTLs, and germplasm annotations spanning 95 plant taxa. The reference ontologies include the Plant Ontology, Plant Trait Ontology, and Plant Experimental Conditions Ontology, developed by the Planteome project, as well as those developed by collaborating groups, such as the GO, PATO, and ChEBI. An important feature of the Planteome database is an integration of species-specific Crop Ontologies describing traits and phenotype scoring standards being utilized by international plant breeding projects. In the Planteome 2.0 Release (February 2017), the Planteome database includes trait ontologies for eight crop species: maize (Zea mays), sweet potato (Ipomoea batatas), soybean (Glycine max), pigeon pea (Cajanus cajan), rice (Oryza sativa), cassava (Manihot esculenta), lentil (Lens culinaris) and wheat (Triticum aestivum). Planteome database users can access the ontologies and annotated data from the project website and ontology browser, perform faceted searches for ontology terms, annotations and bioentities, and download custom datasets for further analysis. Other tools offered by the Planteome include web services (http://planteome.org/web_services) for ontology terms and annotated data, and the Planteome Noctua platform (http://noctua.planteome.org/) for collaborative building of gene annotation models using ontology terms.

The current methods of annotating phenotypes are largely manual, which limits high volume data curation. Therefore, semiautomated methods dependent on data mining using reference ontologies and natural language processing methods have started to become available (Wei et al. (2013); Xu et al. (2016)). The Planteome project, among others, is working towards the goal of using ontologies and neural network-based methods to identify phenotypes and plant characters (phenotypes and traits) from both high-throughput phenotyping project data and plant taxonomic sample collections. Within the context of vertebrates, the Phenoscape comparative framework and Knowledgebase (Mabee et al. (2012); http://www.phenoscape.org) and see Chapter 11) is another platform that, if adapted to include plant data, may enable plant-centric analyses, as well as cross-domain analyses including ones such as those shown in Figure 2.1 and Figure 2.2. The Phenoscape Knowledgebase (http://kb.phenoscape.org) currently combines computable morphological phenotype descriptions from phylogenetic systematics publications on comparative fish morphology and the vertebrate fin-to-limb transition, on the one hand, and from mutant screens and other genetic perturbation experiments in vertebrate model organisms, on the other. By way of example, Figure 2.3 shows the results of querying the Phenoscape Knowledgebase with the search term “urod” (a gene associated with human porphyria cutanea tarda, with phenotype shown in Figure 2.2A). Not only are orthologs in zebrafish, Xenopus, mouse, and human returned by the search, in each species the phenotypes associated with UROD are listed.

Evolutionary phenotype profiles for taxa and clades are linked to the phenotypes of mutated genes in model organisms with the highest semantic similarity. This enables researchers to explore conservation of phenotype in distantly related organisms and leverage knowledge from model organisms to identify candidate genes for related phenotypes in non-model organisms.

An additional goal of the Phenoscape project is the development of natural language processing tools to increase the speed with which existing and new phenotypic data can be rendered computable. Such tools include CharaParser (Cui (2012)) and Phenex (Balhoff et al. (2010); Balhoff et al. (2014)) that enable semi-autonomous encoding of morphological descriptions and facilitate mapping to ontology terms, respectively. Scaling the computable phenotype dataset of the Plant PhenomeNET (6 taxa) to the size of the Phenoscape Knowledgebase (5,211 taxa) through incorporation of existing and future natural language processing tools would drive the prediction of novel candidate genes for crop plants, helping to extract as much information as possible from the wealth of phenotypic data.

2.4 Future Work

What would it take to create a PlantPhenoscape? To build such a resource would require phenotype data in the form of computable ontologies that are universal to all land plants, functional information in the form of Gene Ontology (GO) annotations and gene expression data, and evolutionary data in the form of phylogenetic relatedness among genes and species. Combining phenomic with genomic data in a single query-based database would enable researchers to advance understanding of the basic genomic mechanisms underlying plant development and evolution. Connection and integration of these resources with existing and emerging community data-centric projects (e.g., TAIR, MaizeGDB, Genomes to Fields, CyVerse, DivSeek, Planteome etc.) would ensure broad access and longevity for developed resources.

Once assembled, a PlantPhenoscape Knowledgebase could be used to combine existing phenotype data, including both natural variation and genetic mutant phenotypes, for agricultural and model plants. These data could be assembled based on the ontology-building workflow established by Oellrich et al. (2015). To be most useful, a PlantPhenoscape Knowledgebase should include annotations for genes and gene orthologs including molecular function, biological processes, and localization of gene products from GO and link these with the phenotype database using mutant EQ phenotypes extracted from PhenomeNet, Planteome, and PATO.

Once combined into a single database, the PlantPhenoscape Knowledgebase would allow researchers to ask questions about the evolutionary and phenotypic relatedness of particular structures, and to develop hypotheses concerning the genetic and developmental mechanisms underlying these particular changes. Such a resource could lead to the development of a fast semantic similarity engine for searching in real time across taxa or genotypes for shared phenotypic profiles. A phenotypic profile could include particular morphological or developmental descriptors or shared aspects of gene expression that result in phenotypic differences. While Phenoscape's existing architecture for semantic similarity tests and trait/character matrix capabilities could be used for comparative phylogenetic analyses, the breadth of Phenoscape's computable data types could also be expanded to include species interactions, developmental data, phenotype-genotype connections, gene network interactions, and genomic data.

By combining plant kingdom-wide ontologies for structural and anatomical characters with linked phenotypic and genotypic data from across plant genetic systems, one could describe plant diversity across many lineages and species and predict which genes and gene expression patterns may be responsible not only for ecologically driven and evolutionarily significant changes in plant form and function, but also for traits of interest in the world's major crops.

As noted earlier, the construction of the existing Plant PhenomeNET similarity matrix created by Oellrich et al. (2015) required manual creation of EQ statements from free-text phenotype descriptions sourced from phenotypic databases and literature papers. This conversion from human-readable phenotypic data to their computable representations demanded extensive time and effort from domain experts for each of the plant species included in the project. In a similar fashion, the entity-quality relations of the Phenoscape Knowledgebase are manually generated from phenotypic and morphological data reported in literature and character state matrices, albeit assisted through purpose-built auto-completion tools such as Phenex, mentioned previously. The automation of this process converting human-readable phenotype descriptions into computable EQ statements has the potential to reduce the number of human hours spent on this task, allowing curators to focus primarily on ensuring the quality of the EQ representations rather than generating them. In addition, automating this process would expand the total amount of phenotypic data that can be processed within a given time frame, proportionally expanding the scope of the analyses that can be performed.

Information extraction, one of the problems at the core of parsing out EQ statements from phenotype descriptions, is an established problem in the field of natural language processing (NLP). Within the scope of information extraction, one of the most notable challenges is adapting algorithms and techniques for specific biological domains. Whereas a general case NLP algorithm may identify people or places, biological applications require recognition of items such as gene and protein names (Settles (2005)), and more complex ideas such as disease-phenotype relations (Xu et al. (2013)) and descriptions of mutations within genes (Horn et al. (2004)). These specific information extraction algorithms have been applied and evaluated in the domain of biomedical texts, but differences in taxa notation and vocabulary and the wide variety of phenotypic information available (from anatomical phenotypes, to cellular concentrations, to biological processes, etc.), makes generalization difficult.

CharaParser, the computational tool reported in Cui (2012), addresses this problem of information extraction with respect to morphological phenotypes. From an input in a variety of natural language formats, CharaParser produces character-state formulated phenotype descriptions encoded in an XML file format. Words and phrases corresponding to characters and character-states are identified through an unsupervised learning algorithm (Cui et al. (2010)), preventing the need for the creation of domain-specific annotated training data. Instead, a small number of seed characters and character-states are fed to the algorithm and used to identify patterns leading to the identification of more character and character states in the input text itself, in an iterative process. Following the unsupervised learning algorithm, the proposed characters and character states extracted from the input text are verified, altered, or removed by a human reviewer. The widely used NLP parsing tool Stanford Parser (Klein and Manning (2003)) is then used in combination with sets of heuristic rules to identify the relationship between characters and character states in the input text and produce the final XML representation of the phenotypes. CharaParser has been shown to perform well on plant data sets, with 90% precision and recall at the sentence level on a North American Flora data set, with slightly lower performance on an invertebrate data set (Cui (2012)).

In the XML representation of phenotypes produced by tools such as CharaParser, characters and character-states are computationally identified and organized, but semantically they are represented with the exact vocabulary with which they were represented in the input text. This prevents comparison between multiple encoded descriptions, as is possible with EQ statements. The task of representing predicted characters and character states in terms of entities and qualities drawn from ontologies is referred to as concept coding, or concept mapping. More generally, concept coding refers to mapping between any word or set of words and a corresponding ontology term. As with other natural language processing problems, tools have been built to address the problem of concept coding in the biomedical text domain. Notably, Aronson et al. (Aronson (2001); Aronson and Lang (2010)) developed a concept coding algorithm called MetaMap for mapping text to concepts in the UMLS (Unified Medical Language System) Metathesaurus. The MetaMap algorithm accounts for possible variants of the input text (synonyms, abbreviations, etc.), and scores their similarity to available ontology terms through custom metrics, such as how many words were used to find the match, and how central those words are to the meaning of the input text. Combining the functionality of an information extraction tool like CharaParser, capable of identifying entity and quality-like terms in phenotypic data, with a concept coder capable of mapping those terms to ontological concepts, would allow for near-complete automation of EQ statement generation. Provided that the information extraction algorithms can be developed to perform on a diverse collection of datasets, the variety of EQ statements such a system could generate would only be limited by the availability of ontologies. Anatomical, chemical, and biological process ontologies already provide the basis for representing an extremely wide range of phenotypic information.

2.5 Outlook

Given recent innovations in gene editing and the availability of tools to design specific changes to gene regions of interest (reviewed in Brazelton Jr et al. (2015)), predictive phenomics can be used to target desired phenotypes and test correlations between phenomes and genomes in any species of interest, bringing functional genomics tools to bear on all phenotypes across many species and even domains of life. Together, ontology-based phenotypic prediction, coupled with simplified, broadly accessible gene editing capabilities, will not only advance our understanding of basic biological mechanisms and principles, but has the potential to improve disease models and agricultural innovation.

2.6 Acknowledgements

Thanks to Gerry Neuffer and John Gray for sharing images of the lesion mimic mutant phenotypes in maize. These authors were brought together largely due to the Phenotype Ontology Research Coordination Network supported by National Science Foundation award DBI-0956049. The Planteome Project is supported by National Science Foundation award IOS-1340112. The Phenoscape II project is funded by NSF collaborative grants DBI-1062404 and DBI-1062542 grant, and is supported by the National Evolutionary Synthesis Center (NESCent), NSF #EF-0905606. Its predecessor, the Phenoscape I project, was funded by NSF grant BDI-0641025, and supported by NESCent, NSF #EF-0423641. Contributions by IRB and CJLD were supported by the Predictive Plant Phenomics NSF Research Traineeship (NRT) DGE-1545453 and the Iowa State University Plant Sciences Institute Faculty Scholars program.

2.7 References

Aronson, A. R. (2001). Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association.

Aronson, A. R. and Lang, F.-M. (2010). An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. (2000). Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29.

Balhoff, J. P., Dahdul, W. M., Dececchi, T. A., Lapp, H., Mabee, P. M., and Vision, T. J. (2014). Annotation of phenotypic diversity: decoupling data curation and ontology curation using phenex. *Journal of biomedical semantics*, 5(1):1–5.

Balhoff, J. P., Dahdul, W. M., Kothari, C. R., Lapp, H., Lundberg, J. G., Mabee, P., Midford, P. E., Westerfield, M., and Vision, T. J. (2010). Phenex: ontological annotation of phenotypic diversity. *PLoS One*, 5(5):e10500.

Brazelton Jr, V. A., Zarecor, S., Wright, D. A., Wang, Y., Liu, J., Chen, K., Yang, B., and Lawrence-Dill, C. J. (2015). A quick guide to crispr sgrna design tools. *GM crops & food*, 6(4):266–276.

Burns, T., Breathnach, S., Cox, N., and Griffiths, C. (2008). *Rook's textbook of dermatology*. John Wiley & Sons.

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

Cooper, L., Walls, R. L., Elser, J., Gandolfo, M. A., Stevenson, D. W., Smith, B., Preece, J., Athreya, B., Mungall, C. J., Rensing, S., et al. (2013). The plant ontology as a tool for comparative plant anatomy and genomic analyses. *Plant and Cell Physiology*, 54(2):e1–e1.

Cui, H. (2012). Charaparser for fine-grained semantic annotation of organism morphological descriptions. *Journal of the American Society for Information Science and Technology*, 63(4):738–754.

Cui, H., Boufford, D., and Selden, P. (2010). Semantic annotation of biosystematics literature without training examples. *Journal of the American Society for Information Science and Technology*, 61(3):522–542.

Deans, A. R., Lewis, S. E., Huala, E., Anzaldo, S. S., Ashburner, M., Balhoff, J. P., Blackburn, D. C., Blake, J. A., Burleigh, J. G., Chanet, B., et al. (2015). Finding our way through phenotypes. *PLoS Biol*, 13(1):e1002033.

Gkoutos, G. V., Green, E. C., Mallon, A.-M., Hancock, J. M., and Davidson, D. (2005). Using ontologies to describe mouse phenotypes. *Genome biology*, 6(1):R8.

Gkoutos, G. V., Schofield, P. N., and Hoehndorf, R. (2018). The anatomy of phenotype ontologies: principles, properties and applications. *Briefings in Bioinformatics*, 19(5):1008–1021.

Hastings, J., de Matos, P., Dekker, A., Ennis, M., Harsha, B., Kale, N., Muthukrishnan, V., Owen, G., Turner, S., Williams, M., et al. (2012). The chebi reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic acids research*, 41(D1):D456–D463.

Hoehndorf, R., Schofield, P. N., and Gkoutos, G. V. (2011). Phenomenet: a whole-phenome approach to disease gene discovery. *Nucleic acids research*, 39(18):e119–e119.

Homma, N., Takei, Y., Tanaka, Y., Nakata, T., Terada, S., Kikkawa, M., Noda, Y., and Hirokawa, N. (2003). Kinesin superfamily protein 2a (kif2a) functions in suppression of collateral branch extension. *Cell*, 114(2):229–239.

Horn, F., Lau, A. L., and Cohen, F. E. (2004). Automated extraction of mutation data from the literature: application of mutext to g protein-coupled receptors and nuclear hormone receptors. *Bioinformatics*, 20(4):557–568.

Hu, G., Yalpani, N., Briggs, S. P., and Johal, G. S. (1998). A porphyrin pathway impairment is responsible for the phenotype of a dominant disease lesion mimic mutant of maize. *The Plant Cell*, 10(7):1095–1105.

Johal, G. S. (2007). Disease lesion mimic mutants of maize. *Online. APSnet Features. doi*, 10.

Klein, D. and Manning, C. D. (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st annual meeting of the association for computational linguistics*, pages 423–430.

Lu, L., Lee, Y.-R. J., Pan, R., Maloof, J. N., and Liu, B. (2005). An internal motor kinesin is associated with the golgi apparatus and plays a role in trichome morphogenesis in arabidopsis. *Molecular Biology of the Cell*, 16(2):811–823.

Mabee, P., Balhoff, J. P., Dahdul, W. M., Lapp, H., Midford, P. E., Vision, T. J., and Westerfield, M. (2012). 500,000 fish phenotypes: The new informatics landscape for evolutionary and developmental biology of the vertebrate skeleton. *Journal of Applied Ichthyology*, 28(3):300–305.

McGary, K. L., Park, T. J., Woods, J. O., Cha, H. J., Wallingford, J. B., and Marcotte, E. M. (2010). Systematic discovery of nonobvious human disease models through orthologous phenotypes. *Proceedings of the National Academy of Sciences*, 107(14):6544–6549.

Mungall, C. J., Gkoutos, G. V., Smith, C. L., Haendel, M. A., Lewis, S. E., and Ashburner, M. (2010). Integrating phenotype ontologies across multiple species. *Genome biology*, 11(1):R2.

Neuffer, M. and Pawar, S. (1980). Dominant disease lesion mutants. *Maize Genet. Coop. Newslett*, 54:34–36.

Neuffer, M. G., Calvert, O. H., et al. (1975). Dominant disease lesion mimics in maize. *Journal of Heredity*, 66(5):265–270.

Oellrich, A., Walls, R. L., Cannon, E. K., Cannon, S. B., Cooper, L., Gardiner, J., Gkoutos, G. V., Harper, L., He, M., Hoehndorf, R., et al. (2015). An ontology approach to comparative phenomics in plants. *Plant methods*, 11(1):1–15.

Settles, B. (2005). Abner: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics*, 21(14):3191–3192.

Thessen, A. E., Bunker, D. E., Buttigieg, P. L., Cooper, L. D., Dahdul, W. M., Domisch, S., Franz, N. M., Jaiswal, P., Lawrence-Dill, C. J., Midford, P. E., et al. (2015). Emerging semantics to link phenotype and environment. *PeerJ*, 3:e1470.

Wei, C.-H., Kao, H.-Y., and Lu, Z. (2013). Pubtator: a web-based text mining tool for assisting biocuration. *Nucleic acids research*, 41(W1):W518–W522.

Xu, R., Li, L., and Wang, Q. (2013). Towards building a disease-phenotype knowledge base: extracting disease-manifestation relationship from literature. *Bioinformatics*, 29(17):2186–2194.

Xu, W., Gupta, A., Jaiswal, P., Taylor, C., and Lockhart, P. (2016). Enhancing information accessibility of scientific publications with text mining and ontology. In *CEUR Workshop Proc*, volume 1747, pages 2–3.

2.8 Figures and Tables

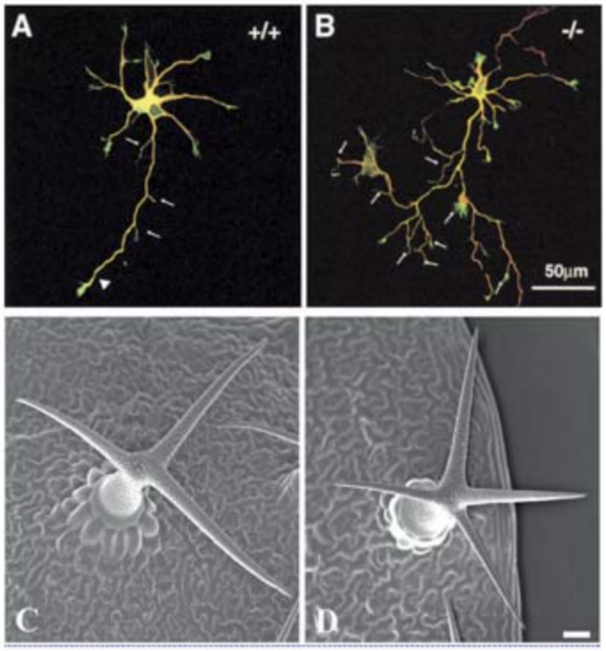


Figure 2.1 Branching defects shared between mouse and Arabidopsis cells. Homma et al. (2003) reported increased branching in cultured neurons of mouse Kinesin-13 mutant KIF2A (A) wild type and (B) mutant. A mutation in the Arabidopsis ortholog KIN-13A (At3g16630) also shows increased trichome branching (C) wild type and (D) mutant (Lu et al. (2005))

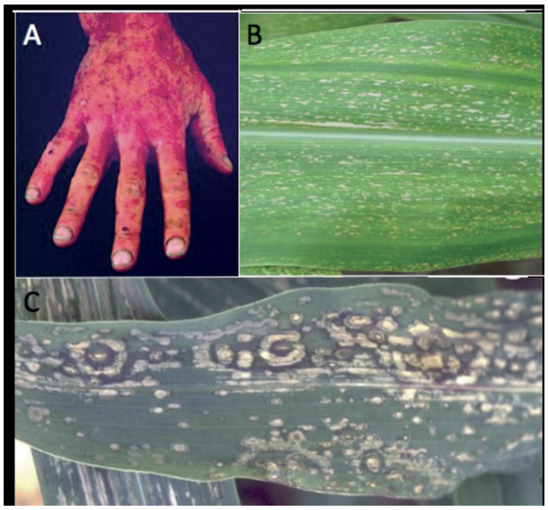
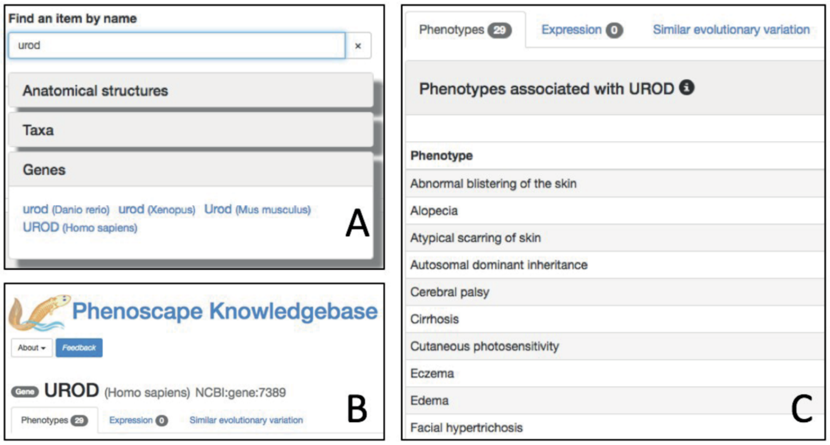


Figure 2.2 Lesion formation in humans and maize. A: Porphyria cutanea tarda in humans[12]. B: Lesion formation in a Les22 mutant plant due to a mutation in the UROD enzyme (image courtesy John Gray). C: The classic oscillatory lesions displayed by Les1 heterozygotes (Neuffer and Pawar (1980))



Find an item by name

urod ×

Anatomical structures

Taxa

Genes

urod (Danio rerio) urod (Xenopus) Urod (Mus musculus) UROD (Homo sapiens)

Phenoscape Knowledgebase

About ▾ Feedback

Gene UROD (Homo sapiens) NCBI:gene:7389

Phenotypes 29 Expression 0 Similar evolutionary variation

Phenotypes 29 Expression 0 Similar evolutionary variation

Phenotypes associated with UROD

Phenotype
Abnormal blistering of the skin
Alopecia
Atypical scarring of skin
Autosomal dominant inheritance
Cerebral palsy
Cirrhosis
Cutaneous photosensitivity
Eczema
Edema
Facial hypertrichosis

Figure 2.3 Results of a query on the Phenoscape Knowledgebase for UROD. A: The search term “urod” returns genes in zebrafish, Xenopus, mouse, and human. B: Clicking on the human UROD, the knowledgebase returns 29 phenotypes. C: A sample of the list of phenotypes associated with human UROD is shown.

CHAPTER 3. COMPUTATIONAL CLASSIFICATION OF PHENOLOGS ACROSS BIOLOGICAL DIVERSITY

Ian R Braun[1], Carolyn J Lawrence-Dill[1]

[1]Genetics, Development, and Cell Biology, Iowa State University, Ames, IA, United States

Modified from a manuscript published in *Proceedings of the 9th International Conference on Biological Ontology (ICBO 2018), Corvallis, Oregon, USA*

3.1 Abstract

Phenotypic diversity analyses are the basis for research discoveries ranging from basic biology to applied research. Phenotypic analyses often benefit from the availability of large quantities of high-quality data in a standardized format. Image and spectral analyses have been shown to enable high-throughput, computational classification of a variety of phenotypes and traits. However, equivalent phenotypes expressed across individuals or groups that are not anatomically similar can pose a problem for such classification methods. In these cases, high-throughput, computational classification is still possible if the phenotypes are documented using standardized, language-based descriptions. Conversion of language-based phenotypes to computer-readable "EQ" statements enables such large-scale analyses. EQ statements are composed of entities (e.g., leaf) and qualities (e.g., increased length) drawn from terms in ontologies. In this work, we present a method for automatically converting free-text descriptions of plant phenotypes to EQ statements using a machine learning approach. Random forest classifiers identify potential matches between phenotype descriptions and terms from a set of ontologies including GO (gene ontology), PO (plant ontology), and PATO (phenotype and trait ontology), among others. These candidate ontology terms are combined into candidate EQ statements, which are probabilistically evaluated with respect to a natural language parse of the phenotype description. Models and

parameters in this method are trained using a dataset of plant phenotypes and curator-converted EQ statements from the Plant PhenomeNET project (Oellrich et al. (2015)). Preliminary results comparing predicted and curated EQ statements are presented. Potential use across datasets to enable automated phenolog discovery are discussed.

3.2 Introduction

Identifying phenologs (comparable phenotypes with hypothesized shared genetic origin) within and between species enables candidate gene prediction for phenotypes of interest in agriculture and medicine alike (McGary et al. (2010); Hoehndorf et al. (2011); Oellrich et al. (2015)). For systems or species which are not anatomically similar, the use of image-based phenotype data makes phenolog identification difficult. In these cases however, semantic analysis of text-based representations of the phenotypes can provide enough information to identify phenologs and generate hypotheses about the underlying biology of interest (Braun et al. (2018)).

Plant PhenomeNET is a phenotype similarity network composed of phenotypes from six different model plant species that demonstrates the utility of this approach (Oellrich et al. (2015)). In the construction of Plant PhenomeNET, curators converted text-based representations of the phenotypes into sets of EQ statements, composed of entities (e.g., leaf) and qualities (e.g., increased length), both represented by ontology terms. The similarity for each pair of phenotypes was then calculated based on the overlap in the sets of ontology terms present in each phenotype's EQ statements. The goal of the work presented here is to automate the process of converting text-based phenotypes to EQ statements using machine learning and natural language processing techniques, so that such phenotype similarity networks can be generated and expanded more easily.

3.3 Methods

3.3.1 Plant PhenomeNET Dataset

The Plant PhenomeNET dataset of phenotype descriptions, corresponding atomized statements, and corresponding curator-generated EQ statements is used as the source of both training and testing data in this work. The atomized statements in this dataset are used as input to the described methods, with the aim of automatically generating logical EQ statements which are similar to those generated by the curators.

3.3.2 Mapping Text to Candidate Terms

The purpose of the first method employed is to map each input atomized statement to a subset of the available ontology terms, which contains only those terms that match the text (may be used to describe a portion of the text). To do this, random forest machine learning models specific to each ontology are trained to classify pairs of text and ontology terms as either matching or not, and are then used to produce probabilities with which the ontology terms may be ranked for a given atomized statement. Features used to represent pairs of text and ontology terms take into account semantic similarity, syntactic similarity, and contextual similarity with respect to the ontology structure. The top ranking ontology terms are taken as candidate terms.

3.3.3 Composing Candidate EQ Statements

For each atomized statement, the candidate ontology terms are used to construct a set of all possible candidate EQ statements. This is done by combining the terms from appropriate ontologies into appropriate roles within the EQ statement structure. Some rules specific to the ontologies used are enforced. For example, the inclusion of a relational PATO term as the quality necessitates a secondary entity term.

3.3.4 Evaluating Candidate EQ Statements

This process evaluates each candidate EQ statement that was composed in the previous step. The atomized statement that was used to generate the candidate EQ statements is processed with the Stanford CoreNLP pipeline, specifically to produce a dependency graph of the text. Each candidate ontology term identified in the previous step is assigned to a node in the dependency graph that is most similar to that ontology term (as measured by similarity metrics of high importance in the random forest models). With each candidate ontology term assigned to a node in the dependency graph, a given EQ statement can be represented by the shortest path in the graph from the Entity term to the Quality term. Distributions of the length of these paths and edge types along the paths are generated from the training data. The structural probability of a candidate EQ statement is defined as the frequency with which its E-to-Q path appears in the training data. The overall quality score q for an EQ is a weighted average of this structural probability and the average probability of the terms, as output by the random forest models.

3.4 Results and Discussion

Random forest classifiers specific to each ontology were evaluated using standard precision and recall curves (Figure 3.1). For the purposes of this evaluation, predicted probabilities for a term are considered correct if they exceed the threshold value and that term is present in the curated EQ statement for that atomized statement. In addition to binary precision and recall, hierarchical similarity metrics are used to evaluate the average similarity between predicted and curated terms with respect to the structure of the ontology (Figure 3.1). For each predicted EQ statement, its similarity to the corresponding curated EQ statement was measured (Figure 3.2). This preliminary work demonstrates the utility of using machine learning and natural language processing techniques for automating or assisting the work of translating text-based phenotypes into EQ statements. Our current and on-going work is focused on 1) adapting the methods to handle more complex phenotypes which map to multiple EQ statements, 2) using and adapting existing tools to extract phenotype descriptions from the literature in order to build an expanded dataset of text descriptions.

3.5 References

Braun, I., Balhoff, J. P., Berardini, T. Z., Cooper, L., Gkoutos, G. V., Harper, L. C., Huala, E., Jaiswal, P., Kazic, T., Lapp, H., et al. (2018). 'computable'phenotypes enable comparative and predictive phenomics among plant species and across domains of life.

Hoehndorf, R., Schofield, P. N., and Gkoutos, G. V. (2011). Phenomenet: a whole-phenome approach to disease gene discovery. *Nucleic acids research*, 39(18):e119–e119.

McGary, K. L., Park, T. J., Woods, J. O., Cha, H. J., Wallingford, J. B., and Marcotte, E. M. (2010). Systematic discovery of nonobvious human disease models through orthologous phenotypes. *Proceedings of the National Academy of Sciences*, 107(14):6544–6549.

Oellrich, A., Walls, R. L., Cannon, E. K., Cannon, S. B., Cooper, L., Gardiner, J., Gkoutos, G. V., Harper, L., He, M., Hoehndorf, R., et al. (2015). An ontology approach to comparative phenomics in plants. *Plant methods*, 11(1):1–15.

3.6 Figures and Tables

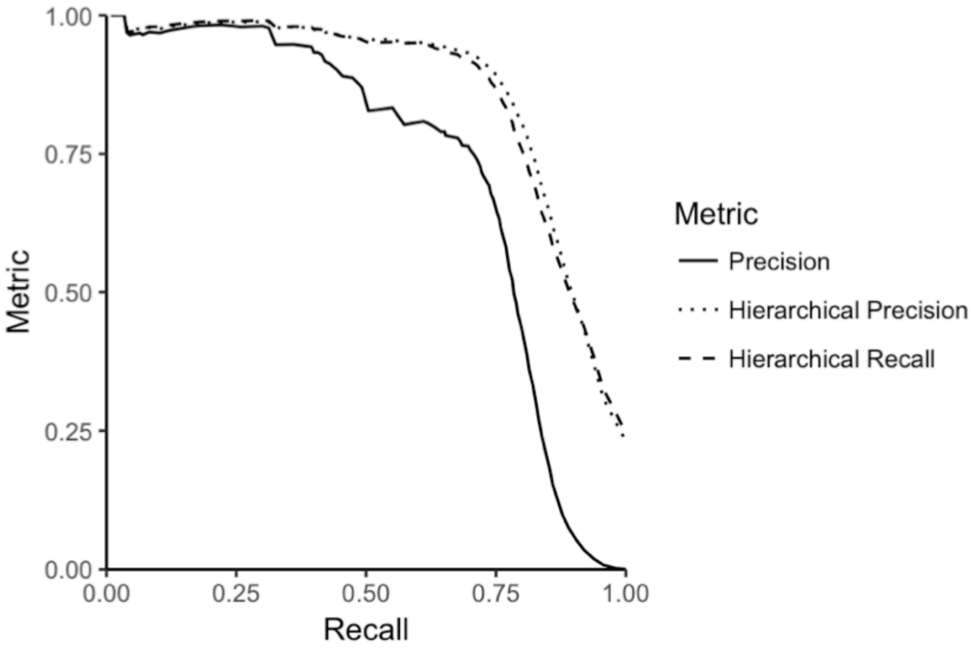


Figure 3.1 Precision recall curve for predicted PATO terms of holdout atomized-statements from Plant PhenomeNET. Average hierarchical precision and recall are shown between all positive predictions and the closest correct PATO terms.

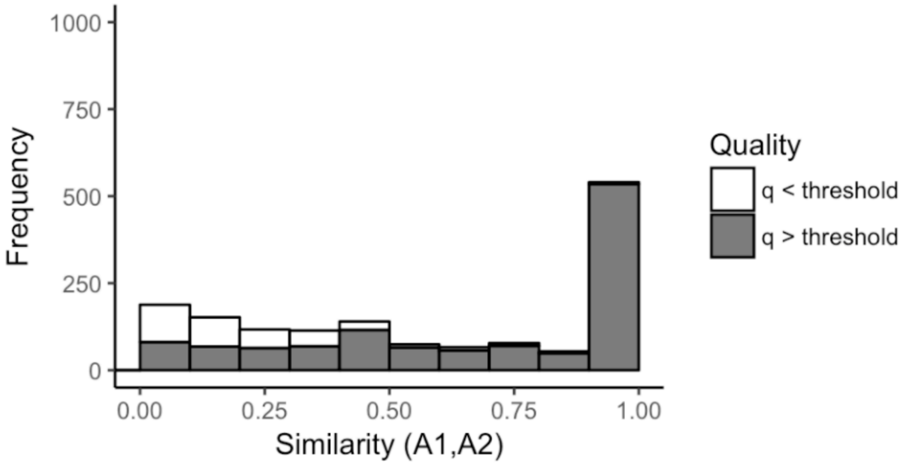


Figure 3.2 Histogram of similarities (weighted Jaccard) between predicted and curated EQ statements for holdout atomized statements from Plant PhenomeNET. Shaded predictions have quality scores exceeding the learned quality threshold value.

CHAPTER 4. AUTOMATED METHODS ENABLE DIRECT COMPUTATION ON PHENOTYPIC DESCRIPTIONS FOR NOVEL CANDIDATE GENE PREDICTION

Ian R. Braun[1,2] and Carolyn J. Lawrence-Dill1[2,3*]

[1]Department of Genetics, Development, and Cell Biology, Iowa State University, Ames, IA, United States

[2]Interdepartmental Bioinformatics and Computational Biology, Iowa State University, Ames, IA, United States

[3]Department of Agronomy, Iowa State University, Ames, IA, United States

Modified from a manuscript published in *Frontiers in Plant Science*

4.1 Abstract

Natural language descriptions of plant phenotypes are a rich source of information for genetics and genomics research. We computationally translated descriptions of plant phenotypes into structured representations that can be analyzed to identify biologically meaningful associations. These representations include the entity–quality (EQ) formalism, which uses terms from biological ontologies to represent phenotypes in a standardized, semantically rich format, as well as numerical vector representations generated using natural language processing (NLP) methods (such as the bag-of-words approach and document embedding). We compared resulting phenotype similarity measures to those derived from manually curated data to determine the performance of each method. Computationally derived EQ and vector representations were comparably successful in recapitulating biological truth to representations created through manual EQ statement curation. Moreover, NLP methods for generating vector representations of phenotypes are scalable to large quantities of text because they require no human input. These results

indicate that it is now possible to computationally and automatically produce and populate large-scale information resources that enable researchers to query phenotypic descriptions directly.

4.2 Background

Phenotypes encompass a wealth of important and useful information about plants, potentially including states related to fitness, disease, and agricultural value. They comprise the material on which natural and artificial selection act to increase fitness or to achieve desired traits, respectively. Determining which genes are associated with traits of interest and understanding the nature of these relationships is crucial for manipulating phenotypes. When causal alleles for phenotypes of interest are identified, they can be selected for in populations, targeted for deletion, or employed as transgenes to introduce desirable traits within and across species. The process of identifying candidate genes and specific alleles associated with a trait of interest is called candidate gene prediction.

Genes with similar sequences often share biological functions and therefore can create similar phenotypes. This is one reason sequence similarity search algorithms like BLAST (Altschul et al. (1990)) are so useful for candidate gene prediction. However, similar phenotypes can also be attributed to the function of genes that have no sequence similarity. This is how protein-coding genes that are involved in different steps of the same metabolic pathway or transcription factors involved in regulating gene expression contribute to shared phenotypes. For example, knocking out any one of the many genes involved in the maize anthocyanin pathway can result in pigment changes (reviewed in Sharma et al. (2011)). This concept is modelled in Figure 4.1, where, notably, the sequence-based search with Gene 1 as a query can only return genes with similar sequences, but querying for similar phenotypes to those associated with Gene 1 returns many additional candidate genes.

High-throughput and computational phenotyping methods are largely sensor and image-based (Fahlgren et al. (2015)). These methods can produce standardized datasets such that, for example, an image can be analyzed, data can be extracted, and those data can be interrogated

(Green et al. (2012); Gehan et al. (2017); Miller et al. (2017)). However, while such methods are adept at comparing phenotypic information between plants that are physically similar, they are limited in their ability to transfer this knowledge between physically dissimilar species. For example, traits such as leaf angle vary greatly among different species, and therefore cannot be compared directly. Moreover, where shared pathways and processes are conserved across broad evolutionary distances, it can be hard to identify equivalent phenotypes. McGary et al. (2010) call these non-obvious shared phenotypes phenologs. Between species, phenologs may present as equivalent properties in disparate biological structures (Braun et al. (2018)). For example, Arabidopsis KIN-13A mutants and mouse KIF2A mutants both show increased branching in single-celled structures, but with respect to neurons in mouse (Homma et al. (2003)) and with respect to trichomes in Arabidopsis (Lu et al. (2005)). Taken together, the ability to compute on phenotypic descriptions to identify phenologs within and across species has the potential to aid in the identification of novel candidate genes that cannot be identified by sequence-based methods alone and that cannot be identified via image analysis.

In order to identify phenologs, some methods rely on searching for shared orthologs between causal gene sets (McGary et al. (2010); Woods et al. (2013)). For example, McGary et al. (2010) identified a phenolog relationship between “abnormal heart development” in mouse and “defective response to red light” in Arabidopsis by identifying four orthologous genes between the sets of known causal genes in each species. However, these methods are not applicable when the known causal gene set for one phenotype or the other is small or non-existent. In these cases, using natural language descriptions to identify phenologs avoids this problem by relying only on characteristics of the phenotypes, per se. These phenotypic descriptions are a rich source of information that, if leveraged to identify phenolog pairs, can enable identification of novel candidate genes potentially involved in generating phenotypes beyond what has already been described.

Unfortunately, computing on phenotype descriptions is not straightforward. Text descriptions of phenotypes present in the literature and in online databases are irregular because natural language representations of even very similar phenotypes can be highly variable. This makes reliable quantification of phenotype similarity particularly challenging (Thessen et al. (2012); Braun et al. (2018)). To represent phenotypes in a computable manner, researchers have recently begun to translate and standardize phenotype descriptions into entity–quality (EQ) statements composed of ontology terms, where an entity (e.g., "leaf") is modified by a quality (e.g., "increased length"; Mungall et al. (2010)).1 Using this formalism, complex phenotypes are represented by multiple EQ statements. For example, multiple EQ statements are required to represent dwarfism, where the entity and quality pairs ("plant height," "reduced") and ("leaf width," "increased") may be used, among others. Each of these phenotypic components of the more general phenotype is termed a "phene." Because both entities and qualities are represented by terms from biological ontologies (fixed vocabularies arranged as hierarchical concepts in a directed acyclic graph), quantifying the similarity between two phenotypes that have been translated to EQ statements can be accomplished using graph-based similarity metrics (Hoehndorf et al. (2011); Slimani (2013)). Such techniques for estimating semantic similarity based on arranging concepts hierarchically in a graph have long been employed in the field of natural language processing (NLP; e.g., Resnik (1999)) and, as applied to biological ontologies, have been useful in applications from clustering gene function annotations for data visualization (Supek et al. (2011)) to assessing functional similarities between orthologous genes (Altenhoff et al. (2016)).

Oellrich et al. (2015) developed Plant PhenomeNET, an EQ statement-based resource primarily consisting of a phenotype similarity network containing phenotypes across six different model plant species, namely, Arabidopsis (Arabidopsis thaliana), maize (Zea mays ssp. mays), tomato (Solanum lycopersicum), rice (Oryza sativa), Medicago (Medicago truncatula), and soybean (Glycine max). Their analysis demonstrated that the method developed by Hoehndorf et al. (2011) could be used to recover known genotype to phenotype associations for plants. The authors found that highly similar phenotypes in the network (phenologs) were likely to share causal genes that were orthologous or involved in the same biological pathways. In constructing the network, text statements comprising each phenotype were converted by hand into EQ statements primarily composed of terms from the Phenotype and Trait Ontology (PATO; Gkoutos et al. (2005)), Plant Ontology (PO; Cooper et al. (2013)), Gene Ontology (GO; Ashburner et al. (2000)), and Chemical Entities of Biological Interest (ChEBI; Hastings et al. (2012)) ontology.

The success of this plant phenotype pilot project was encouraging, but to scale up to computing on all available phenotypic data for each of the six species was not a reasonable goal given that curating data for this pilot project took approximately 2 years and covered only phenotypes of dominant alleles for 2,747 genes across the six species. More specifically, human translation of text statements into EQ statements is the most time-consuming aspect of generating phenotype similarity networks using this method. Automation of this translation promises to increase the rate at which such networks can be generated and expanded. Notable efforts to automate this process include Semantic Charaparser (Cui (2012); Cui et al. (2015)), which extracts characters (entities) and their corresponding states (qualities) after a curation step that involves assigning terms to categories and then mapping these characters and states to EQ statements constructed from input ontologies. Other existing annotation tools such as NCBO Annotator (Musen et al. (2012)) and NOBLE Coder (Tseytlin et al. (2016)) are fully automated, relying only on input ontologies. Both map words in the input text to ontology terms without imposing an EQ statement structure. State-of-the-art machine learning approaches to annotating text with ontology terms also have been developed (Hailu et al. (2019)). These can be trained using a dataset such as the Colorado Richly Annotated Full-Text corpus (CRAFT; Bada et al. (2012)), but are not readily transferable to ontologies that are not represented in the training set.

In addition to using ontology-based methods, similarity between text descriptions of phenotypes can also be quantified using NLP techniques such as treating each description as a bag-of-words and comparing the presence or absence of those words between descriptions, or using neural network-based tools such as Doc2Vec to embed descriptions into abstract high-dimensional numerical vectors between which similarity metrics can then be easily applied (Mikolov et al. (2013); Le and Mikolov (2014)). Conceptually, this process involves converting natural language descriptions into locations in space, such that descriptions that are near each other are interpreted as having high similarity and those that are distant have low similarity.

In this work, we demonstrate that automated techniques for generating computable representations of natural language can be applied to a dataset of phenotypic descriptions in order to generate biologically meaningful phenotype similarity networks. See Figure 4.2 for an overview of how phenotype similarity networks are computationally generated as an output when text descriptions are provided as the input. We first show that these computational techniques are limited in their capability to exactly reproduce the annotations and corresponding phenotype similarity networks generated with hand-curation. However, we subsequently show that the hand-curated network does not outperform networks built with purely computational approaches on dataset-wide tasks of biological relevance, such as organizing genes by function and predicting membership in biochemical pathways. Most importantly, we discuss how we can now use these computational approaches to automatically generate new datasets necessary to identify phenotypic similarities and predict gene function within and across species without requiring the use of time-consuming and costly hand-curation.

4.3 Methods

4.3.1 Dataset of Phenotypic Descriptions and Curated EQ Statements

The pairwise phenotype similarity network described in Oellrich et al. (2015) was built based on a dataset of phenotype descriptions across six different model plant species (A. thaliana, Z. mays ssp. mays, S. lycopersicum, O. sativa, M. truncatula, and G. max). In that work, each phenotype description was split into one or more atomized statements describing individual phenes, each of which mapped to exactly one curated EQ statement (Table 4.1). The EQ statements in this dataset were primarily built from terms present in PATO, PO, GO, and ChEBI. For this work, we used this existing dataset as the source of genes and associated phenotypic descriptions on which to test automated methods for assessing similarity networks between phenotypes and using the resulting phenotype similarity networks to perform comparative analyses across the whole dataset to predict gene function.

4.3.2 Computationally Generating EQ Statements From Phenotypic Descriptions

For each phenotype and phene description in the dataset, we computationally generated corresponding EQ statements without human interaction. To accomplish this, terms were first annotated to each text description and then combined to form complete EQ statements. Two different existing computational tools and a simple machine learning technique were used to map ontology terms to text descriptions. Specifically, these were NCBO Annotator and NOBLE Coder, which are tools for matching ontology terms to specific words in text, and a Naïve Bayes bag-of-words classifier, which assigns terms to descriptions based on the observed frequencies of term–word co-occurrence in a training dataset. The Oellrich et al. (2015) dataset of descriptions and curated EQ statements was split into four groups such that any three groups of the dataset were used to train a Naïve Bayes model that was then applied to the remaining group. The result of applying these three annotation methods was a set of ontology terms from PATO, PO, GO, and ChEBI assigned to each text description. Terms were then combined to form full EQ statements by assigning default root terms where none were matched, such as the entity term whole plant (PO:0000003), and organizing the matched terms into the different roles of the EQ statement by removing overlapping terms and automatically applying compositional rules used by curators in Oellrich et al. (2015). As an example, these rules include the fact that ChEBI terms cannot be the primary entity. The EQ statements were scored based on how well the terms aligned with the text description they were annotated to, so that the closest matching EQ statements for each text description were output and used downstream to generate phenotype similarity networks. See the Supplemental Methods section for a more detailed description of this process.

4.3.3 Computationally Generating Numerical Vectors From Phenotypic Descriptions

In addition to generating EQ statements for each phenotype and phene description in the dataset, Doc2Vec was used for generating numerical vectors for each description. A model pre-trained on Wikipedia was used (Lau and Baldwin (2016)). In these document embeddings, positions within the vector do not refer to the presence of specific words but rather abstract features learned by the model. A size of 300 was used for each vector representation, which is the fixed vector size of the pre-trained model. In addition, vectors were generated for each description using bag-of-words and set-of-words representations of the text. For these methods, each position within the vector refers to a particular word in the vocabulary. Each vector element with bag-of-words refers to the count of that word in the description, and each vector element with set-of-words is a binary value indicating presence or absence of the word. In cases where phene descriptions were used instead of phenotype descriptions, the descriptions were concatenated prior to embedding to obtain a single vector.

4.3.4 Creating Gene and Phenotype Networks

Oellrich et al. (2015) developed a network with phenotypes as nodes and similarity between them as edges for all the phenotypes in the dataset. For each type of text representations that we generated with computational methods, comparable networks were constructed. For EQ statement representations, Jaccard similarity either taking the structure and order of terms in the EQ statement into account (referred to as metric S1) or ignoring the structure and treating the ontology terms in the EQ statement as an unordered set (referred to as metric S2) were used to determine edge values. See the Supplemental Methods section for a more detailed description of these similarity metrics. For vector representations generated using Doc2Vec and bag-of-words, cosine similarity was used. For the vector representations generated using set-of-words, Jaccard similarity was used. These networks are considered to be simultaneously gene and phenotype similarity networks because each phenotype in the dataset corresponds to a specific causal gene and a node in the network represents both that causal gene and its cognate phenotype. However, two phenotype descriptions corresponding to the same gene are retained as two separate nodes in the network, so while each node represents a unique gene/phenotype pair, a single gene may be represented within more than one node.

4.4 Results

4.4.1 Performance of Computational Methods in Reproducing Hand-Curated Annotations

We tested the ability of computational semantic annotation methods to assign ontology terms similar to those selected by curators to phenotype and phene descriptions in the Oellrich et al. (2015) dataset. Specifically, the ontology terms mapped by each method to a particular description were compared against the terms present in the EQ statement(s) that were created by hand-curation for that same description. Metrics of partial precision (PP) and partial recall (PR), as well as the harmonic mean of these values (PF_1) as a summary statistic, were used to evaluate performance (Table 4.2). Metrics PP and PR were applied as in Dahdul et al. (2018); see the Supplemental Methods section for a detailed description of these metrics.

NOBLE Coder and NCBO Annotator generally produced semantic annotations more similar to the hand-curated dataset using phenotype descriptions as inputs than using the set of phene descriptions as inputs, a result consistent across ontologies. We considered this to be counter-intuitive because the phene descriptions are more directly related to the individual EQ statements in terms of semantic content. However, the set of target ontology terms considered correct is larger in the case of the phenotype descriptions because this set of terms includes all terms in any EQ statements derived from that phenotype rather than a single EQ statement, which could contribute to this measured increase in both partial recall and partial precision. Accounting for synonyms and related words generated through Word2Vec models increased PR in the case of specific annotation methods as the threshold for word similarity was decreased (from

1.0 to 0.5), but did not increase PF_1 in any instance due to the corresponding losses in PP (Supplemental Figure 1).

NOBLE Coder and NCBO Annotator performed comparably in the case of each type of text description and ontology, with NOBLE Coder using the precise matching parameter slightly outperforming the other annotation method with respect to these particular metrics for these particular descriptions. Both outperformed the Naïve Bayes classifier, for which performance dropped significantly for the ontologies with smaller relative representation in the dataset (GO and ChEBI), as might be expected. When the results were aggregated, the increase in partial recall for PATO, PO, and GO terms relative to the maximum recall achieved by any individual method indicates that the curated terms that were recalled by each method were not entirely overlapping. This is as expected given that different methods used for semantic annotation recalled target (curated) ontology terms to different degrees, as measured by Jaccard similarity of a given target term to the closest predicted term annotated by that particular method. These sets of obtained similarities to target terms were comparable between NCBO Annotator and NOBLE Coder ($\rho = 0.84$ with phene descriptions and $\rho = 0.86$ with phenotype descriptions) and dissimilar between either of those methods and the Naïve Bayes classifier ($\rho < 0.10$ in both cases for either type of description) using Spearman rank correlation adjusted for ties.

These results indicate that automated annotation methods (NCBO Annotator, NOBLE Coder, and Naïve Bayes classifier) do not reproduce the exact same ontology term annotations selected by hand-curation for each phenotypic description, as expected. Given this result, we next assessed how these differences between the hand-curated annotations and computationally generated annotations translated into differences between the phenotype similarity networks based on these annotations.

4.4.2 Comparing Computational Networks to the Hand-Curated Network

Oellrich et al. (2015) developed a network with phenotype/gene pairs as nodes and similarity between them as edges for all phenotypes in the dataset. In this work, comparable networks were constructed for the same dataset using a number of computational approaches for representing phenotype and phene descriptions and for predicting similarity. For the purposes of this assessment, the network built from hand-curated EQ statements and described in Oellrich et al. (2015) is considered the gold standard against which each network we produced is compared. The computational and gold standard networks were compared using the F1 metric to assess similarity in predicted phenolog pairs at a range of k values, where k is the allowed number of phenolog pairs predicted by the networks (the k most highly valued edges). Results are reported through k = 583,971, which is the number of non-zero similarities between phenotypes in the gold standard network, and were repeated using phenotype descriptions and phene descriptions as inputs to the computational methods (Figure 4.3). The simplest NLP methods for assessing similarity (set-of-words and bag-of-words) consistently recapitulated the gold standard network the best using phenotype descriptions, whereas the document embedding method using Doc2Vec outperformed these methods for values of k 200,000 based on phene descriptions. The differences in the performance of each method are robust to 80% subsampling of the phenotypes present in the dataset.

These results illustrate that computational methods do not exactly reproduce the phenotype similarity network built from the hand-curated EQ statements. However, this does not necessarily mean that the hand-curated network is inherently more biologically meaningful. To assess how useful each network is in a biological context, we next compared how the hand-curated network and each computational network performed on the task of sorting genes into functional groups.

4.4.3 Computational Methods Outperform Hand-Curation for Gene Functional Categorization in Arabidopsis

Lloyd and Meinke (2012) previously organized a set of Arabidopsis genes with accompanying phenotype descriptions into a functional hierarchy of groups (e.g., “morphological”), classes (e.g., “reproductive”), and finally subsets (e.g., “floral”), in order from most general to most specific. See Supplemental Table 1 in Lloyd and Meinke (2012) for a full specification of this hierarchy to which the genes were assigned, and Supplemental Table 2 in Lloyd and Meinke (2012) for a mapping between genes and this hierarchical vocabulary. Oellrich et al. (2015) later used this set of genes and phenotypes to validate the quality of their dataset of hand-curated EQ statements by reporting the average similarity of phenotypes (translated into EQ statements) that belonged to the same functional subset. We used this same functional hierarchy categorization and a similar approach to assess the utility of computationally generated representations of phenotypes towards correctly categorizing the functions of the corresponding genes and to compare this utility against that of the dataset of hand-curated EQ statements. For each class and subset in the hierarchy, the mean similarity between any two phenotypes related to genes within that class or subset ("within" mean) was quantified using each computable representation of interest and compared to the mean similarity between a phenotype related to a gene within that class or subset and one outside of it ("between" mean), quantified in terms of standard deviation of the distribution of all similarity scores generated for each given method. The difference between the "within" mean and "between" mean (referred to here as the Consistency Index) for each functional category for each method indicates the ability of that method to generate strong similarity signal for phenotypes in this dataset that share that function (Figure 4.4). In the case of these data, most computational methods using either phene or phenotype descriptions as the input text were able to recapitulate the signal present in the network Oellrich et al. (2015) generated from hand-curated EQ statements, and the simplest NLP methods (bag-of-words and set-of-words) produced the most consistent signal.

In order to more directly compare each method on a general classification task, networks constructed from curated EQ statements and those generated using each computational method were used to iteratively classify each Arabidopsis phenotype into classes and subsets. This was accomplished by removing one phenotype at a time and withholding the remaining phenotypes as training data, learning a threshold value from the training data, and then classifying the held-out phenotype by calculating its average similarity to each training data phenotype in each class or subset and classifying it as belonging to any category for which the average similarity to other phenotypes in that category exceeded the learned threshold. Performance on this classification task using each network was assessed using the F1 metric, where the functional category assignments for each gene reported by Lloyd and Meinke (2012) were considered to be the correct classifications (Table 4.3). The simplest NLP methods (bag-of-words and set-of-words) outperformed the Oellrich et al. (2015) hand-curated EQ statement network on this classification task in all cases, while using the computationally generated EQ statements or document embeddings generated with Doc2Vec only outperformed the curated EQ statement network in some cases.

Taken together, these results indicate that even though the computationally generated networks are significantly different than the hand-curated network (Figure 4.3), they generally perform equally well or better on tasks related to organizing Arabidopsis genes into functional groups. We next examined how these networks compare on the task of predicting biochemical pathway membership for specific genes, both within a single species and across multiple species.

4.4.4 Computational Methods Outperform Hand-Curation for Recovering Genes Involved in Anthocyanin Biosynthesis Both Within and Between Species

Oellrich et al. (2015) illustrated the utility of using EQ statement representations of phenotypes to provide semantic information necessary to recover shared membership of causal genes in regulatory and metabolic pathways. Specifically, they showed that by querying a six-species phenotype similarity network with the c2 (colorless2) gene in maize, which is involved in anthocyanin biosynthesis, genes c1, r1, and b1 (colorless1, red1, and booster1), which are also involved in anthocyanin biosynthesis in maize, are recovered. Querying in this instance is defined as returning other genes in the similarity network, ranked using the maximal value of the edges connecting a phenotype corresponding to the query gene and a phenotype corresponding to each other gene in the network. There are 2,747 genes in the dataset, so querying with one gene returns a ranked list of 2,746 genes. This result was included by Oellrich et al. (2015) as a specific example of the general utility of the phenotype similarity network to return other members of a pathway or gene regulatory network when querying with a single gene. See Figure 4.1 for a general illustration of this concept.

To evaluate this same utility in the phenotype similarity networks we generated using computational methods and to compare their utility to that of the network from Oellrich et al. (2015) generated using hand-curated EQ statements, we first expanded the set of maize anthocyanin pathway genes to include those present in the description of the pathway given by Li et al. (2019), and listed in Supplementary Table 1 of that publication. Of those genes, 10 are present in the Oellrich et al. (2015) dataset (Table 4.4). Additionally, we likewise identified the set of Arabidopsis genes known to be involved in anthocyanin biosynthesis (listed in Table 1 of Appelhagen et al. (2014)) that were present in the Oellrich et al. (2015) dataset. This yielded a total of 16 Arabidopsis genes (Table 4.5).

4.4.5 Recovering Anthocyanin Biosynthesis Genes Within a Single Species

Using each phenotype similarity network, each anthocyanin biosynthesis gene from one species was iteratively used as a query against the network. The rank of each other gene in the set of anthocyanin biosynthesis genes corresponding to the same species as the query was quantified. We grouped the ranks into bins of width 10 for ranks less than or equal to 50 and combined all ranks greater than 50 into a single bin. For each phenotype similarity network, the mean and standard deviation of the number of anthocyanin biosynthesis genes in each bin were calculated (Figure 4.5). The average number of pathway genes ranked within the top 10 across all queries was greater for all computationally generated networks than for the network built from hand-curated EQ statements, although variance across the queries was high. In general, computational networks built from predicted EQ statements performed best for this task, whereas the network built using the hand-curated EQs performed the worst. The networks constructed using the numerical vector representations (set-of-words, bag-of-words, and Doc2Vec) were intermediate in performance as a group (Figure 4.5).

4.4.6 Recovering Anthocyanin Biosynthesis Genes Between Two Species

To determine whether the methods performed similarly both within and across species, we repeated the analysis described in the previous section (Recovering Anthocyanin Biosynthesis Genes Within a Single Species), but instead of quantifying the ranks of all anthocyanin biosynthesis genes from the same species as the query gene, we quantified the ranks of all anthocyanin genes that derived from the other species. In other words, Arabidopsis genes were used to query for maize genes, and maize genes were used to query for Arabidopsis genes. As shown in Figure 4.6, the phenotype similarity network constructed from hand-curated EQ statements did not recover (provide ranks of less than or equal to 50) any of the anthocyanin biosynthesis genes when queried with genes from the other species. Networks generated using the set-of-words and bag-of-words approaches, or with Doc2Vec, performed similarly, recovering on average less than one anthocyanin biosynthesis gene per query. Only networks built from computationally generated EQ statements recovered an appreciable number of anthocyanin biosynthesis genes on average across the queries between species (Figure 4.6).

4.5 Discussion

4.5.1 Computationally Generated Phenotype Representations Are Useful

A primary purpose for generating representations of phenotypes that are easy to compute on (EQ statements, vector embeddings, etc.) is to construct similarity networks that enable the use of one phenotype as a query to retrieve similar phenotypes. This process serves as a means of discovering relatedness between phenotypes (potential phenologs) within and across species, thus generating hypotheses about underlying genetic relatedness (reviewed in Oellrich et al. (2015)).

The computational methods discussed in this work were demonstrated to only partially recapitulate the phenotype similarity network constructed by Oellrich et al. (2015) using hand-curated EQ statements (Comparing Computational Networks to the Hand-Curated Network). Despite the limited similarity between the network built from hand-curated annotations and the computationally generated networks, the computationally generated networks performed as well or better than the hand-curated network (based on curated EQ statements) in terms of correctly organizing phenotypes and their causal genes into functional categories at multiple hierarchical levels (Computational Methods Outperform Hand-Curation for Gene Functional Categorization in Arabidopsis). In addition, each computationally generated network performed better than the hand-curated network for querying with either maize or Arabidopsis anthocyanin biosynthesis genes to return other anthocyanin biosynthesis genes from the same species (Recovering Anthocyanin Biosynthesis Genes Within a Single Species), a task originally used to demonstrate the utility of the phenotype similarity network constructed in Oellrich et al. (2015).

Moreover, the networks built from computationally generated EQ statements were useful for recapturing anthocyanin biosynthesis genes from a species different than the species of origin for the queried gene/phenotype pair. None of the other networks, including the network built from curated EQ statements, exhibited this utility for this task (Recovering Anthocyanin Biosynthesis Genes Between Two Species). This particular result indicates that high accuracy of constructed EQ statements is not specifically necessary for tasks such as querying for related genes across species because potentially inaccurate (computationally predicted) EQ statements generated a more successful network for the task. Replicating these analyses with phenotype descriptions in a different biological domain, such as vertebrates, would determine whether these results generalize to additional species groups and datasets.

Taken together, these results over this particular dataset of phenotype descriptions suggest that while the EQ statements generated through manual curation are likely the most accurate and informative computable representation of a given phenotype in specific cases, other representations generated entirely computationally with no human intervention are capable of meeting or exceeding the performance of the hand-curated annotations on dataset-wide tasks such as sorting phenotypes and genes into functional categories, as well as in the case of specific tasks such as querying with particular genes to recover other genes involved in the same pathway.

Therefore, in cases where the volume of data is large, the results are understood to be predictive, and manual curation is impractical, using automated annotation methods to generate large-scale phenotype similarity networks is a worthwhile goal and can provide biologically relevant information that can be used for hypothesis generation, including novel candidate gene prediction.

4.5.2 Multiple Approaches to Representing Natural Language Are Useful

EQ statement annotations comprising ontology terms allow for interoperability with compatible annotations from varied data sources. They are also a human-readable annotation format, meaning that a knowledgeable human could fix an incorrect annotation by selecting a more appropriate ontology term (a process that is not possible using abstract vector embeddings). Their uniform structure also provides a means of explicitly querying for phenotypes involving a biological entity that is similar to some structure or process (e.g., trichomes) or matches some quality (e.g., an increase in physical size). Ontology-based annotations have the potential to increase the information attached to a phenotype (through inferring ancestral terms which are not specifically referred to in the phenotype description), but do not necessarily fully capture the detail and semantics of the natural language description.

For this reason, future representations of phenotypes in relational databases for the purpose of generating phenotype similarity networks across a large volume of phenotypes described in literature and in databases likely should include both ontology-based annotations describing the phenotypes, as well as the original natural language descriptions. Although the number of phenotypes in the dataset used here and described in Oellrich et al. (2015) is relatively small, the results of this work suggest utility of original text representations as a powerful means of calculating similarity between phenotypes, especially within a single species. Computationally generated EQ statements, which in the context of this study do not often meet the criteria for a fully logical curated EQ statement, were demonstrated to be more useful in any other approach for recovering biologically related genes across species.

Ensemble methods are often applied in the field of machine learning, where multiple methods are used to solve a problem, with a higher-level model determining which method will be most useful in solving each new instance of the problem. It is possible that such an approach could be applied to measuring similarity between phenotypes to generate a single large-scale network, where similarity values are based on the best possible method to assess the text representations of each pair of particular phenotypes.

4.5.3 Additional Challenges With EQ Statement Representation

Although ontology terms and EQ statements composed of ontology terms are an information-rich representation of phenes and phenotypes, flexibility in which terms and statements can represent a particular phenotype can limit the ability to computationally recognize true biological similarity. The graph structures of the ontologies themselves, the metrics used to assess semantic similarity, and the ambiguity inherent in both natural language and EQ statement representations of phenes and phenotypes can all potentially contribute to this problem.

As one example in the Oellrich et al. (2015) dataset used here, the phene description "complete loss of flower formation" was annotated with an EQ statement whose entity is flower development, whereas the computationally identified entity using the methods described in this work was flower formation. In this instance, the Jaccard similarity between these two ontology terms was 0.286, which by comparison is less than the Jaccard similarity between flower formation and leaf formation in the context of the ontology graph. This selected example illustrates the possible discrepancies between true biological similarity and semantic similarity as measured using graph-based metrics. Although each semantic similarity metric calculates this value differently, those that use the hierarchical nature of the ontology are all constrained by the structure of the graph itself.

Variation in how humans and computational methods interpret how a phenotype as a whole should be conceptualized also has the potential to produce representations that obscure true similarity, as measured by graph-based metrics. In another example from the Oellrich et al.

(2015) dataset, the phene description “stamens transformed to pistils” was annotated with two different EQ statements. The first EQ statement uses the relational quality has fewer parts of type to indicate the absence of stamen in this phenotype, and the second uses the relational quality has extra parts of type to indicate the presence of pistils in this phenotype. This representation of the phenotype makes logical sense, but is not easy to generate computationally because it abstractly describes the outcome of the transformation that is explicitly present in the natural language description and is dissimilar from computationally generated representations that focus on the explicit content (i.e., those which use the relational quality transformed to).

Finally, this study looked at a dataset consisting entirely of phenotypic descriptions in English, and the generalizability of these methods to other languages is not discussed. It is certainly likely that structural differences between languages would result in differences in how certain methods of computing over descriptions in those languages perform, but such analysis is outside the scope of this work.

4.5.4 Extending This Work to the Wealth of Text Data Available in Databases and the Literature

We plan to apply the methods of semantic annotation, ontology-based semantic similarity calculation, and natural language-based semantic similarity calculation to the wealth of text data available in existing plant model organism databases and biological literature. For the latter, doing so will involve the additional challenge of extracting phenotype descriptions as well as the genes causative to those phenotypes as a separate identification and processing step. We plan to leverage existing work in the areas of named entity recognition specific to genes (Wei et al. (2015)) and relation extraction, as well as existing methods for extracting information related to phenotypes such as those developed using vector-based representations of phenotype descriptions (Xing et al. (2018)) and grammar-tree representations of phenotype descriptions (Collier et al. (2015)). As the size of the applicable dataset is increased by these means, we will continue to analyze the performance of methods from the domains of machine learning and NLP towards constructing biologically meaningful networks from this phenotypic data, including additional techniques that were not included in the results presented here. For example, Sent2Vec (Pagliardini et al. (2017)) is another technique for assessing text similarity that takes a different approach from Doc2Vec for embedding text as numerical vectors and has been shown to perform well when trained on life science corpora (Chen et al. (2019)). These next steps are anticipated to enable researchers to begin to compute on phenotype descriptions directly and will drive a promising future for forward genetics research approaches where phenotypes can be used for novel candidate gene prediction as easily as sequence similarity searches can be used to identify putative homologs from sequence data.

4.6 Data Availability Statement

The dataset of phenotype and phene descriptions and the corresponding hand-curated EQ statements used in this work are available as supplemental data of Oellrich et al. (2015). The hierarchical functional categorization of the set of Arabidopsis genes used in this work is available as supplemental data of Lloyd and Meinke (2012). The code used to produce the results of this work is available at `github.com/irbraun/phenologs`. Files necessary to reproduce the discussed results, datasets used to generate figures presented in this work, and other supplemental files are available at `doi.org/10.5281/zenodo.3255020`. This data repository also includes versions of the previously described datasets available as supplemental data of Oellrich et al. (2015) and Lloyd and Meinke (2012), for the purpose of making this study reproducible without any additional external files.

4.7 Funding

The authors were supported to carry out this work by an Iowa State University Presidential Interdisciplinary Research Seed Grant, the Iowa State University Plant Sciences Institute Faculty Scholars Program, and the Predictive Plant Phenomics NSF Research Traineeship (#DGE-1545453).

4.8 Authors Contributions

IB and CL-D together contributed to the conception and design of the study. IB organized the data, performed the analyses, and wrote the manuscript. IB and CL-D contributed to manuscript revision and read and approved the final version.

4.9 Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

4.10 Acknowledgments

This manuscript has been released as a preprint at doi.org/10.1101/689976. We thank Lisa Harper, Sowmya Vajjala, and Ramona Walls for helpful discussions and suggestions. We are grateful to the NSF Phenotype Ontology RCN (#DBI-0956049) for creating foundations for this work by bringing plant and computational biologists together to develop a common vocabulary and for their support to the Plant Phenotype Pilot Project participants who developed the Oellrich et al. (2015) datasets that our analyses relied upon. We thank the reviewers for their valuable guidance. Based upon their suggestions, the manuscript was improved significantly.

4.11 Supplementary Material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2019.01629/full#supplementary-material

4.12 References

Altenhoff, A. M., Boeckmann, B., Capella-Gutierrez, S., Dalquen, D. A., DeLuca, T., Forslund, K., Huerta-Cepas, J., Linard, B., Pereira, C., Pryszcz, L. P., et al. (2016). Standardized benchmarking in the quest for orthologs. *Nature methods*, 13(5):425–430.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410.

Appelhagen, I., Thiedig, K., Nordholt, N., Schmidt, N., Huep, G., Sagasser, M., and Weisshaar, B. (2014). Update on transparent testa mutants from arabidopsis thaliana: characterisation of new alleles from an isogenic collection. *Planta*, 240(5):955–970.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. (2000). Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29.

Bada, M., Eckert, M., Evans, D., Garcia, K., Shipley, K., Sitnikov, D., Baumgartner, W. A., Cohen, K. B., Verspoor, K., Blake, J. A., et al. (2012). Concept annotation in the craft corpus. *BMC bioinformatics*, 13(1):161.

Braun, I., Balhoff, J. P., Berardini, T. Z., Cooper, L., Gkoutos, G. V., Harper, L. C., Huala, E., Jaiswal, P., Kazic, T., Lapp, H., et al. (2018). 'computable'phenotypes enable comparative and predictive phenomics among plant species and across domains of life.

Chen, Q., Peng, Y., and Lu, Z. (2019). Biosentvec: creating sentence embeddings for biomedical texts. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 1–5. IEEE.

Collier, N., Groza, T., Smedley, D., Robinson, P. N., Oellrich, A., and Rebholz-Schuhmann, D. (2015). Phenominer: from text to a database of phenotypes associated with omim diseases. *Database*, 2015.

Cooper, L., Walls, R. L., Elser, J., Gandolfo, M. A., Stevenson, D. W., Smith, B., Preece, J., Athreya, B., Mungall, C. J., Rensing, S., et al. (2013). The plant ontology as a tool for comparative plant anatomy and genomic analyses. *Plant and Cell Physiology*, 54(2):e1–e1.

Cui, H. (2012). Charaparser for fine-grained semantic annotation of organism morphological descriptions. *Journal of the American Society for Information Science and Technology*, 63(4):738–754.

Cui, H., Dahdul, W., Dececchi, A. T., Ibrahim, N., Mabee, P., Balhoff, J. P., and Gopalakrishnan, H. (2015). Charaparser+ eq: Performance evaluation without gold standard. *Proceedings of the Association for Information Science and Technology*, 52(1):1–10.

Dahdul, W., Manda, P., Cui, H., Balhoff, J. P., Dececchi, T. A., Ibrahim, N., Lapp, H., Vision, T., and Mabee, P. M. (2018). Annotation of phenotypes using ontologies: a gold standard for the training and evaluation of natural language processing systems. *Database*, 2018.

Fahlgren, N., Gehan, M. A., and Baxter, I. (2015). Lights, camera, action: high-throughput plant phenotyping is ready for a close-up. *Current opinion in plant biology*, 24:93–99.

Gehan, M. A., Fahlgren, N., Abbasi, A., Berry, J. C., Callen, S. T., Chavez, L., Doust, A. N., Feldman, M. J., Gilbert, K. B., Hodge, J. G., et al. (2017). Plantcv v2: Image analysis software for high-throughput plant phenotyping. *PeerJ*, 5:e4088.

Gkoutos, G. V., Green, E. C., Mallon, A.-M., Hancock, J. M., and Davidson, D. (2005). Using ontologies to describe mouse phenotypes. *Genome biology*, 6(1):R8.

Green, J. M., Appel, H., Rehrig, E. M., Harnsomburana, J., Chang, J.-F., Balint-Kurti, P., and Shyu, C.-R. (2012). Phenophyte: a flexible affordable method to quantify 2d phenotypes from imagery. *Plant methods*, 8(1):1–12.

Hailu, N. D., Bada, M., Hadgu, A. T., and Hunter, L. E. (2019). Biomedical concept recognition using deep neural sequence models. *bioRxiv*, page 530337.

Hastings, J., de Matos, P., Dekker, A., Ennis, M., Harsha, B., Kale, N., Muthukrishnan, V., Owen, G., Turner, S., Williams, M., et al. (2012). The chebi reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic acids research*, 41(D1):D456–D463.

Hoehndorf, R., Schofield, P. N., and Gkoutos, G. V. (2011). Phenomenet: a whole-phenome approach to disease gene discovery. *Nucleic acids research*, 39(18):e119–e119.

Homma, N., Takei, Y., Tanaka, Y., Nakata, T., Terada, S., Kikkawa, M., Noda, Y., and Hirokawa, N. (2003). Kinesin superfamily protein 2a (kif2a) functions in suppression of collateral branch extension. *Cell*, 114(2):229–239.

Lau, J. H. and Baldwin, T. (2016). An empirical evaluation of doc2vec with practical insights into document embedding generation. *arXiv preprint arXiv:1607.05368*.

Le, Q. V. and Mikolov, T. (2014). Distributed representations of sentences and documents. corr abs/1405.4053 (2014). *arXiv preprint arXiv:1405.4053*.

Li, T., Zhang, W., Yang, H., Dong, Q., Ren, J., Fan, H., Zhang, X., and Zhou, Y. (2019). Comparative transcriptome analysis reveals differentially expressed genes related to the tissue-specific accumulation of anthocyanins in pericarp and aleurone layer for maize. *Scientific reports*, 9(1):1–12.

Lloyd, J. and Meinke, D. (2012). A comprehensive dataset of genes with a loss-of-function mutant phenotype in arabidopsis. *Plant physiology*, 158(3):1115–1129.

Lu, L., Lee, Y.-R. J., Pan, R., Maloof, J. N., and Liu, B. (2005). An internal motor kinesin is associated with the golgi apparatus and plays a role in trichome morphogenesis in arabidopsis. *Molecular Biology of the Cell*, 16(2):811–823.

McGary, K. L., Park, T. J., Woods, J. O., Cha, H. J., Wallingford, J. B., and Marcotte, E. M. (2010). Systematic discovery of nonobvious human disease models through orthologous phenotypes. *Proceedings of the National Academy of Sciences*, 107(14):6544–6549.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Miller, N. D., Haase, N. J., Lee, J., Kaeppler, S. M., de Leon, N., and Spalding, E. P. (2017). A robust, high-throughput method for computing maize ear, cob, and kernel attributes automatically from images. *The Plant Journal*, 89(1):169–178.

Mungall, C. J., Gkoutos, G. V., Smith, C. L., Haendel, M. A., Lewis, S. E., and Ashburner, M. (2010). Integrating phenotype ontologies across multiple species. *Genome biology*, 11(1):R2.

Musen, M. A., Noy, N. F., Shah, N. H., Whetzel, P. L., Chute, C. G., Story, M.-A., Smith, B., and team, N. (2012). The national center for biomedical ontology. *Journal of the American Medical Informatics Association*, 19(2):190–195.

Oellrich, A., Walls, R. L., Cannon, E. K., Cannon, S. B., Cooper, L., Gardiner, J., Gkoutos, G. V., Harper, L., He, M., Hoehndorf, R., et al. (2015). An ontology approach to comparative phenomics in plants. *Plant methods*, 11(1):1–15.

Pagliardini, M., Gupta, P., and Jaggi, M. (2017). Unsupervised learning of sentence embeddings using compositional n-gram features. *arXiv preprint arXiv:1703.02507*.

Resnik, P. (1999). Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of artificial intelligence research*, 11:95–130.

Sharma, M., Cortes-Cruz, M., Ahern, K. R., McMullen, M., Brutnell, T. P., and Chopra, S. (2011). Identification of the pr1 gene product completes the anthocyanin biosynthesis pathway of maize. *Genetics*, 188(1):69–79.

Slimani, T. (2013). Description and evaluation of semantic similarity measures approaches. *arXiv preprint arXiv:1310.8059*.

Supek, F., Bošnjak, M., Škunca, N., and Šmuc, T. (2011). Revigo summarizes and visualizes long lists of gene ontology terms. *PloS one*, 6(7):e21800.

Thessen, A. E., Cui, H., and Mozzherin, D. (2012). Applications of natural language processing in biodiversity science. *Advances in bioinformatics*, 2012.

Tseytlin, E., Mitchell, K., Legowski, E., Corrigan, J., Chavan, G., and Jacobson, R. S. (2016). Noble–flexible concept recognition for large-scale biomedical natural language processing. *BMC bioinformatics*, 17(1):32.

Wei, C.-H., Kao, H.-Y., and Lu, Z. (2015). Gnormplus: an integrative approach for tagging genes, gene families, and protein domains. *BioMed research international*, 2015.

Woods, J. O., Singh-Blom, U. M., Laurent, J. M., McGary, K. L., and Marcotte, E. M. (2013). Prediction of gene-phenotype associations in humans, mice, and plants using phenologs. *BMC bioinformatics*, 14(1):1–17.

Xing, W., Qi, J., Yuan, X., Li, L., Zhang, X., Fu, Y., Xiong, S., Hu, L., and Peng, J. (2018). A gene–phenotype relationship extraction pipeline from the biomedical literature using a representation learning approach. *Bioinformatics*, 34(13):i386–i394.

4.13 Figures and Tables

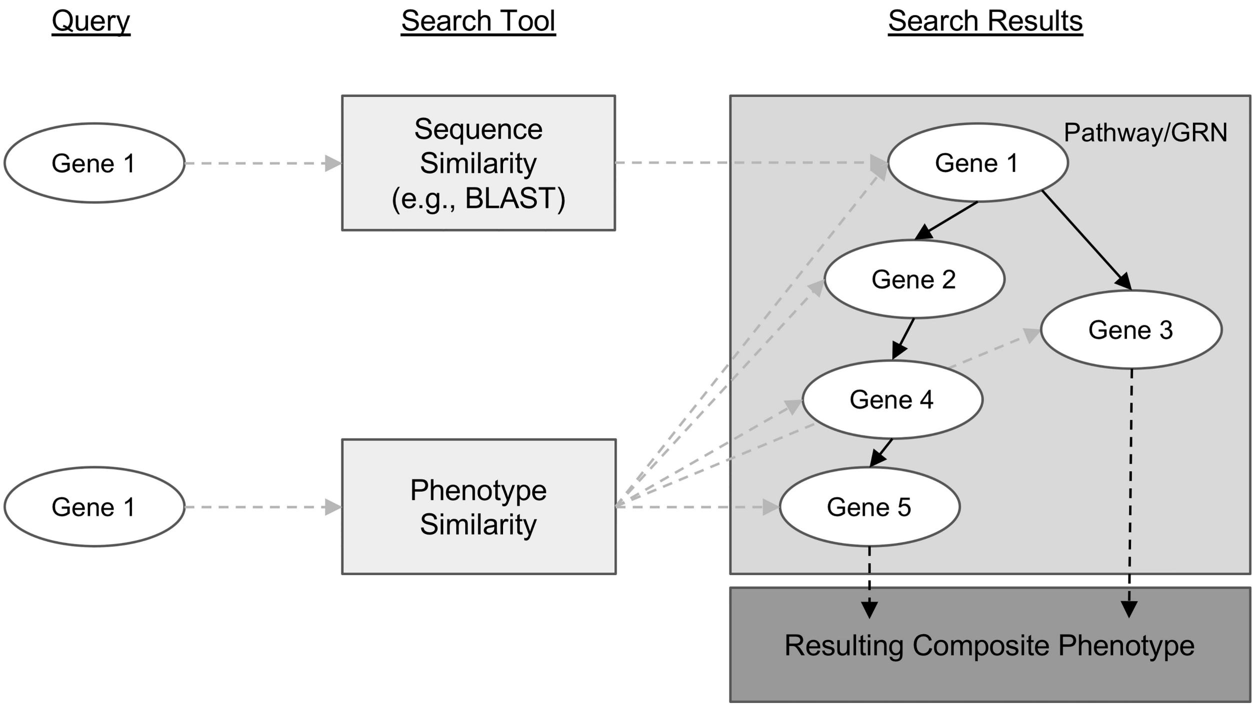


Figure 4.1 Conceptual comparison of querying with a gene sequence or its associated phenotypic description. Genes are shown as white ovals. Methods of searching for related genes are shown as light gray boxes. Gray dashed arrows indicate the path from the query gene to the set of genes that are returned from the search. Solid black arrows indicate relationships between genes in a biological pathway or gene regulatory network. Dashed black arrows indicate relationships between the pathway or network and the resulting phenotype.

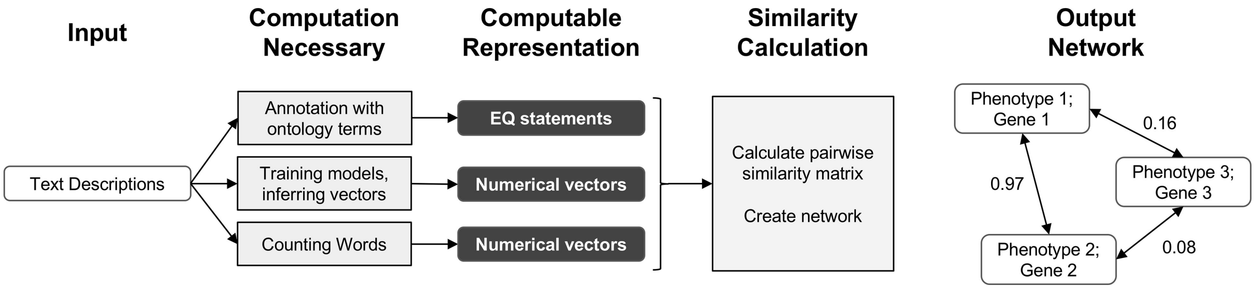


Figure 4.2 Overview of computational pipelines used here to generate phenotype similarity networks from text descriptions of phenotypes. Rounded white rectangles represent data in the form of text descriptions as input or network nodes as output. Rounded black rectangles represent the intermediate data forms that are computable representations of text descriptions. These allow for quantitative similarity metrics to be applied. Gray rectangles represent computational methods carried out at each step. Single-headed arrows represent flow of data through each pipeline. Double-headed arrows represent edges between nodes in resulting similarity networks. Values next to double-headed arrows indicate magnitude of phenotype similarity. One output network is created for each computable representation, but only one example is shown here.

Table 4.1 Description of the Oellrich et al. (2015) dataset in terms of number of phenotype descriptions, phene descriptions, and EQ statements.

Species	Phenotypes	Phenes	EQ Statements
Arabidopsis	1385	5172	5172
Maize	117	373	373
Tomato	90	269	269
Rice	86	340	340
Medicago	40	149	149
Soybean	24	61	61
Example gene: Arabidopsis PKS2 (AT1G14280.1)	Phenotype: short hypoctyl and expanded cotelydon under hourly far red pulses	Phene 1: short hypocotyl	PO:0020100 (hypocotyl) + PATO:0000574 (decreased length)
		Phene 2: expanded cotyledon	PO:0020030 (cotyledon) + PATO:0000586 (increased size)

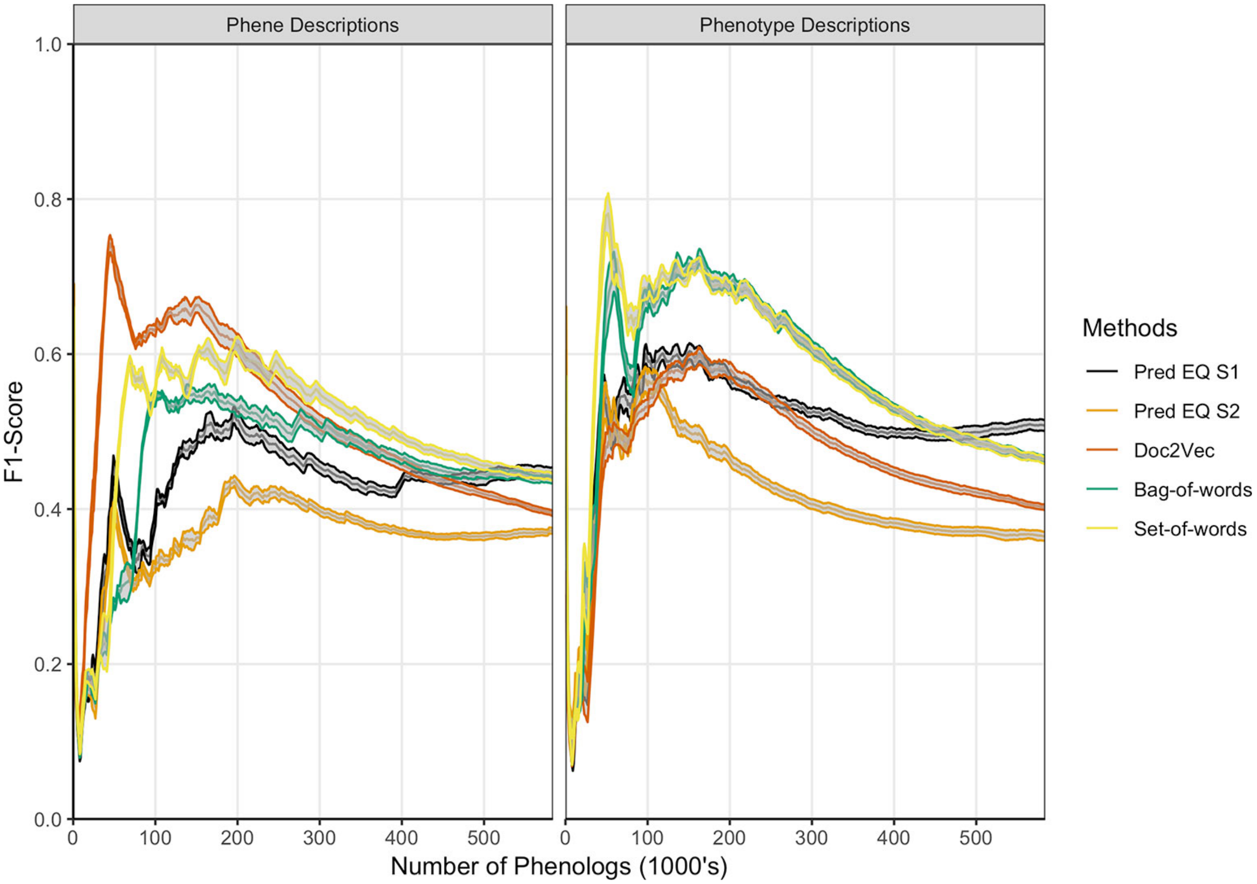


Figure 4.3 Comparison of phenolog pairs identified by predictive methods in comparison to the Oellrich et al. (2015) dataset. The x-axis indicates the number of phenologs pairs (highest valued edges in the phenotype similarity network) at each point. The standard deviation of resampling with 80% of the phenotypes in the dataset (network nodes) are indicated by ribbons for each method. Phene descriptions (left) or phenotype descriptions (right) were used as the text input for each particular method.

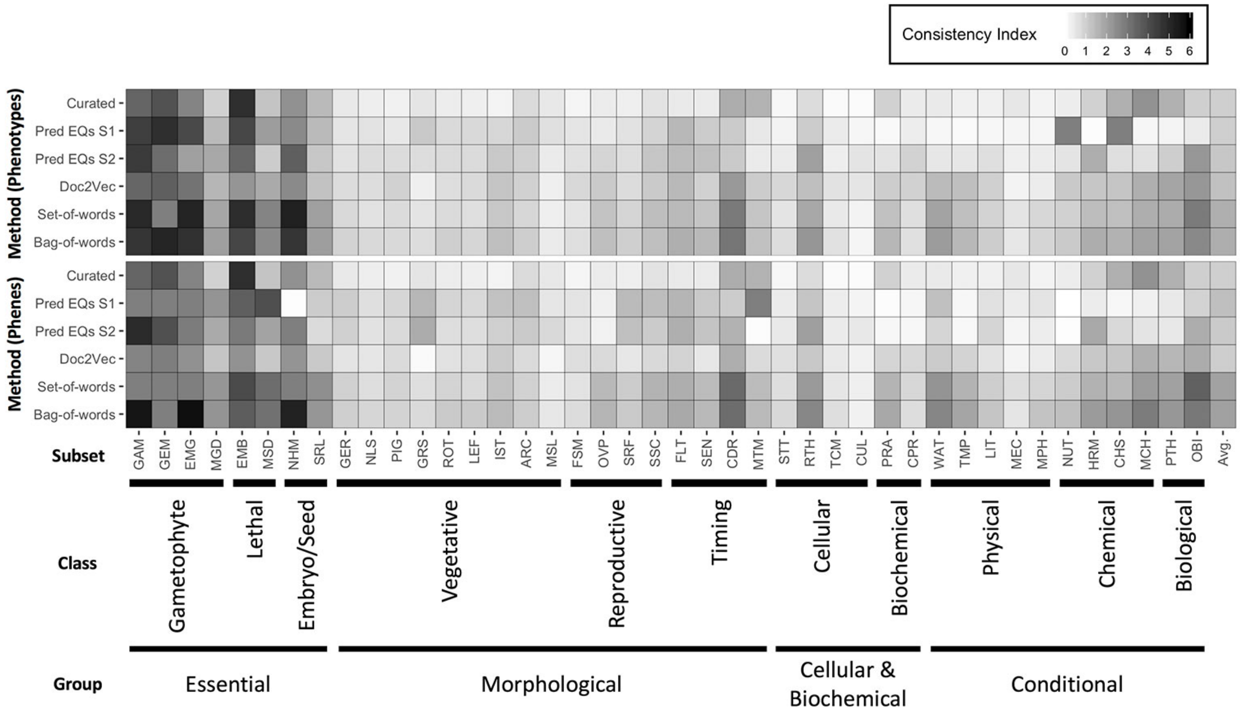


Figure 4.4 Heatmap of Consistency Index. The difference between average similarity for two phenotypes within a subset and one phenotype within and one outside, for each functional subset defined in the dataset of Arabidopsis phenotypes, and for each method of quantifying similarity between phenotypes is shown, with darker cells indicating higher consistency within a subset. Differences are measured in standard deviations of the distributions of similarities obtained for each method. The meaning of subset abbreviations are specified in Supplemental Table 1 of Lloyd and Meinke (2012). Methods are listed at left. Input text for calculating similarities between the phenotypes were either derived from phenotype descriptions (top) or phene descriptions (bottom). The far right column in the heatmap refers to an average Consistency Index for a given method across all subsets.

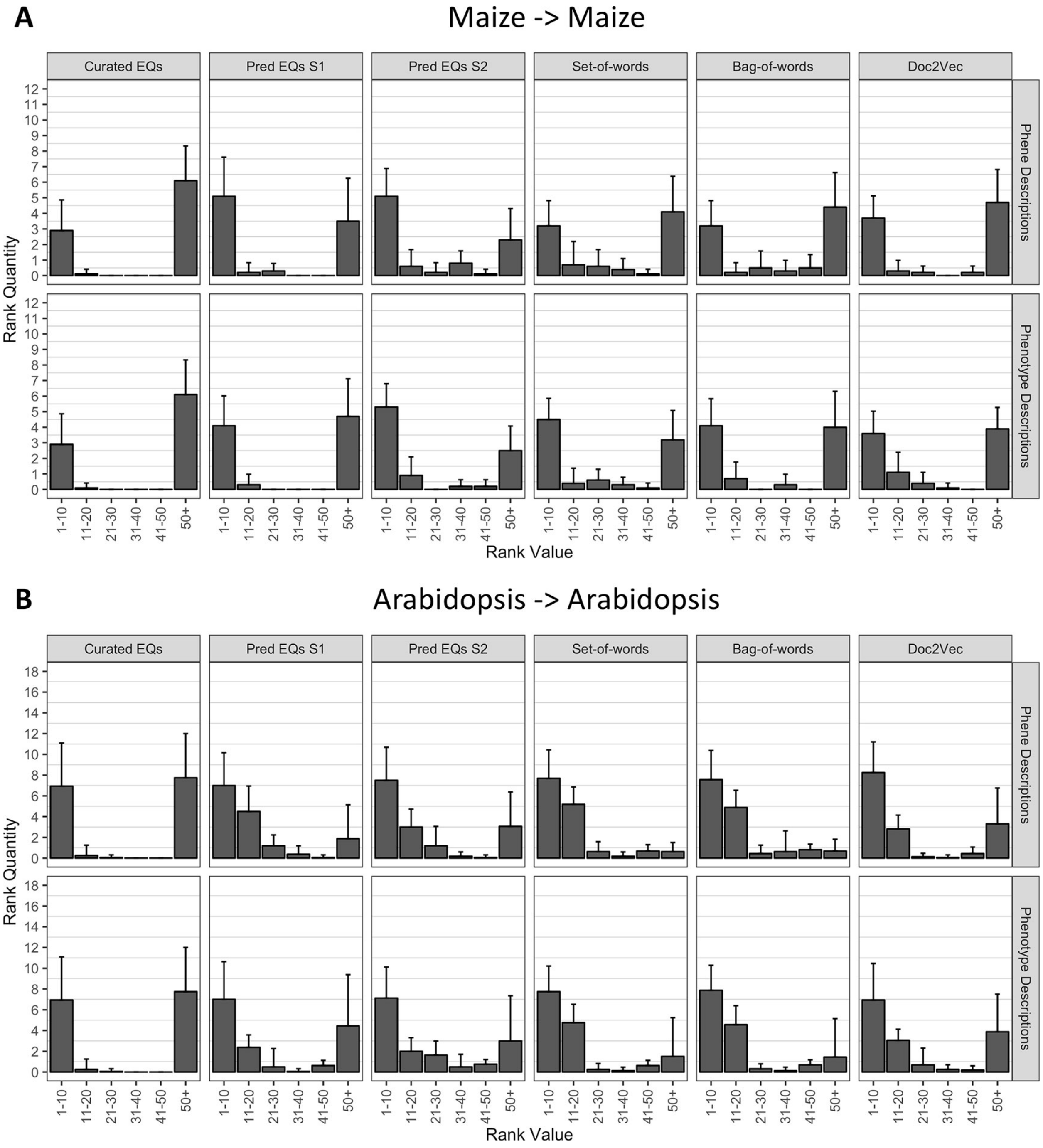


Figure 4.5 Rankings of anthocyanin biosynthesis genes in either maize (A) or Arabidopsis (B) upon querying phenotype similarity networks generated with genes from the same species. Phenotype networks are organized by the method used to generate them (columns) and by whether those methods were applied to phenotype or phene descriptions (rows). Rank value specifies a range of rankings for each bar in the plots (1–10, 11–20, etc.) and rank quantity indicates the average number of anthocyanin biosynthesis genes that were ranked in a given range over all queries. Error bars indicate one standard deviation of the rank quantities in each range over all queries.

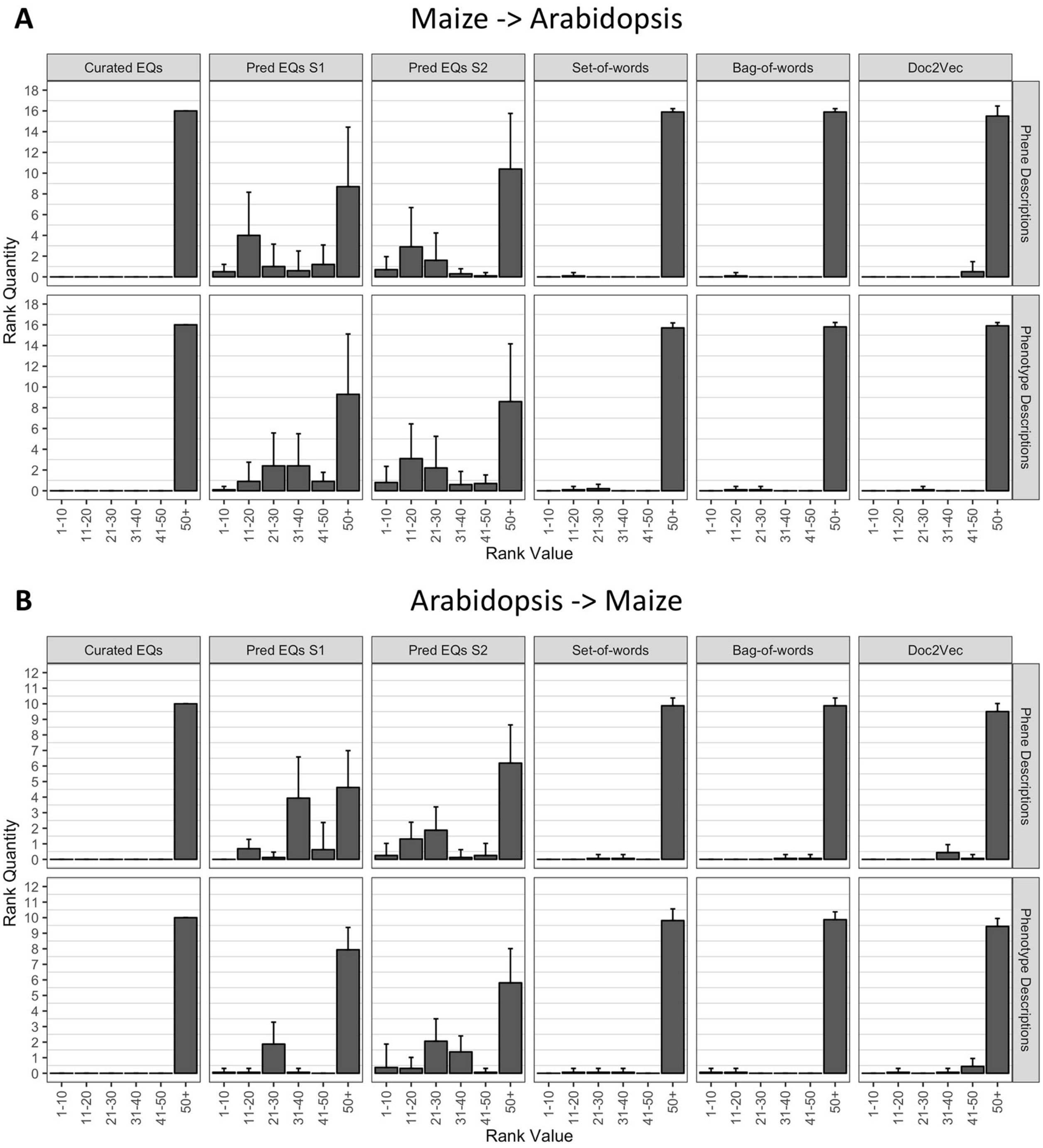


Figure 4.6 Rankings of anthocyanin biosynthesis genes in either maize (A) or Arabidopsis (B) upon querying phenotype similarity networks generated with genes from the other species. Phenotype networks are organized by the method used to generate them (columns) and by whether those methods were applied to phenotype or phene descriptions (rows). Rank value specifies a range of rankings for each bar in the plots (1–10, 11–20, etc.) and rank quantity indicates the average number of anthocyanin biosynthesis genes that were ranked in a given range over all queries. Error bars indicate one standard deviation of the rank quantities in each range over all queries.

Table 4.2 Performance metrics for semantic annotation methods.

Annotator	Ontology	n	Phenotype Description			Phene Descriptions		
			PP	PR	PF_1	PP	PR	PF_1
NOBLE Coder (precise)	PATO	7882	0.641	0.627	0.634	0.601	0.572	0.586
NOBLE Coder (precise)	PO	5634	0.622	0.380	0.472	0.546	0.294	0.382
NOBLE Coder (precise)	GO	1505	0.514	0.521	0.517	0.510	0.514	0.512
NOBLE Coder (partial)	PATO	7882	0.412	0.748	0.532	0.375	0.689	0.486
NOBLE Coder (partial)	PO	5634	0.309	0.758	0.439	0.269	0.659	0.382
NOBLE Coder (partial)	GO	1505	0.102	0.846	0.182	0.091	0.839	0.165
NCBO Annotator	PATO	7882	0.640	0.619	0.629	0.598	0.563	0.580
NCBO Annotator	PO	5634	0.550	0.259	0.352	0.458	0.170	0.248
NCBO Annotator	GO	1505	0.478	0.433	0.454	0.480	0.424	0.450
NCBO Annotator	ChEBI	775	0.429	0.888	0.579	0.431	0.913	0.586
Naïve Bayes Classifier	PATO	7882	0.517	0.394	0.447	0.642	0.484	0.552
Naïve Bayes Classifier	PO	5634	0.474	0.258	0.334	0.636	0.429	0.512
Naïve Bayes Classifier	GO	1505	0.091	0.073	0.081	0.155	0.157	0.156
Naïve Bayes Classifier	ChEBI	775	0.035	0.031	0.033	0.001	0.001	0.001
Aggregate Annotations	PATO	7882	0.412	0.798	0.543	0.383	0.815	0.522
Aggregate Annotations	PO	5634	0.351	0.809	0.489	0.304	0.831	0.445
Aggregate Annotations	GO	1505	0.107	0.839	0.190	0.090	0.839	0.163
Aggregate Annotations	ChEBI	775	0.366	0.890	0.519	0.305	0.913	0.457

Table 4.3 Evaluation (F1 scores) for each method used to categorize Arabidopsis genes by function.

Method	Phenes		Phenotypes	
	Class	Subset	Class	Subset
Curated EQs	0.470	0.359	0.470	0.359
Pred EQs S1	0.472	0.472	0.369	0.320
Pred EQs S2	0.504	0.413	0.437	0.368
Set-of-words	0.613	0.447	0.587	0.426
Bag-of-words	0.595	0.423	0.549	0.409
Doc2Vec	0.455	0.331	0.486	0.377

Table 4.4 Maize genes involved in anthocyanin biosynthesis.

Gene Name (Symbol)	B73 RefGen v3 ID	Functional Category	Encoded Protein
colorless2 (*c2*)	**GRMZM2G422750**	Enzyme	naringenin-chalcone synthase
chalcone flavonone isomeras1 (*chi1*)	GRMZM2G155329	Enzyme	chalcone isomerase
red aleurone1 (*pr1*)	GRMZM2G025832	Enzyme	flavonoid 3'-hydroxylase (flavonoid 3'-monooxygenase)
flavanone 3-hydroxylase1 (fht1; F3H)	GRMZM2G062396	Enzyme	flavonone 3'-hydroxylase (flavonol synthase)
anthocyaninless1 (*a1*)	**GRMZM2G026930**	Enzyme	dihydroflavonol 4-reductase (flavonone 4-reductase)
anthocyaninless2 a2	**GRMZM2G345717**	Enzyme	anthocyanidin synthase (leucoanthocyanidin dioxygenase)
bronze1 (*bz1*)	**GRMZM2G165390**	Enzyme	flavonol 3-O-glucosyltransferase
bronze2 bz2	**GRMZM2G016241**	Enzyme	glutathione transferase (maleylacetoacetate isomerase)
multidrug resistance associated protein3 (*mrpa3; ZmMrp4*)	GRMZM2G111903	Transporter	multidrug-resistance-like-transporter
scutellar node color1 (*sn1*)	GRMZM5G822829	TF	bHLH
colorless1 (*c1*)	**GRMZM2G005066**	TF	R2 R3-MYB
pericarp color1 p1	**GRMZM2G084799**	TF	R2 R3-MYB
purple plant1 (*pl1*)	**GRMZM2G701063**	TF	R2 R3-MYB
colored1 (*r1*)	**GRMZM5G822829**	TF	bHLH
colored plant1 (*b1*)	**GRMZM2G172795**	TF	bHLH
pale aleurone color1 (*pac1*)	GRMZM2G058432	TF	WD40

Table 4.5 Arabidopsis genes involved in anthocyanin biosynthesis.

Gene Name (Symbol)	Locus Name	Functional Category	Encoded Protein
TRANSPARENT TESTA 4 (*TT4*)	**At5g13930**	Enzyme	naringenin-chalcone synthase
TRANSPARENT TESTA 5 (*TT5*)	**At3g55120**	Enzyme	chalcone isomerse
TRANSPARENT TESTA 6 (*TT6*)	**At3g51240**	Enzyme	flavanone 3'-hydroxylase (flavonol synthase
TRANSPARENT TESTA 7 (*TT7*)	**At5g07990**	Enzyme	flavonoid 3'-hydroxylase (flavonoid 3'-monooxygenase)
TRANSPARENT TESTA 3 (*TT3*)	**At5g42800**	Enzyme	dihydroflavonol 4-reductase (flavonone 4-reductase)
TRANSPARENT TESTA 11 (*TT11*) *TRANSPARENT TESTA 17* (*TT17*) *TRANSPARENT TESTA 18* (*TT18*) *TANNIN-DEFICIENT SEED 4* (*TDS4*)	At4g22880	Enzyme	anthocyanidin synthase (leucoanthocyanidin dioxygenase)
ARABIDOPSIS SIP1 CLADE *TRIHELIX 1* (*AST1*) *BANYULS* (*BAN1*)	**At1g61720**	Enzyme	anthocyanidin reductase
TRANSPARENT TESTA 14 (*TT14*) *TRANSPARENT TESTA 19* (*TT19*)	**At5g17220**	Enzyme	glutathione transerase (maleylacetoacetate isomerase)
AUTOINHIBITED H^+-ATPASE (*AHA10*)	At1g17260	Enzyme	ATP-ase
TRANSPARENT TESTA 10 (*TT10*)	**At5g48100**	Enzyme	laccase
TRANSPARENT TESTA 5 (*TT15*)	**At1g43620**	Enzyme	3β-hydroxy sterol glucosyltransferase
TRANSPARENT TESTA 12 (*TT12*)	**At3g59030**	Transporter	MATEefflux proton antiporter
TRANSPARENT TESTA 16 (*TT16*)	**At5g23260**	TF	K-box, MADS-box
TRANSPARENT TESTA 1 (*TT1*)	**At1g34790**	TF	C2H2
TRANSPARENT TESTA 2 (*TT2*)	**At5g35550**	TF	bHLH
TRANSPARENT TESTA 8 (*TT8*)	**At4g09820**	TF	bHLH
TRANSPARENT TESTA GLABRA 1 (*TTG1*)	**At5g24520**	TF	WD40
TRANSPARENT TESTA GLABRA 2 (*TTG2*)	**At2g37260**	TF	WRKY

CHAPTER 5. COMPUTING ON PHENOTYPIC DESCRIPTIONS FOR CANDIDATE GENE DISCOVERY AND CROP IMPROVEMENT

Ian R. Braun[1,2], Colleen F. Yanarella[1,2], and Carolyn J. Lawrence-Dill[1,2,3]

[1]Interdepartmental Bioinformatics and Computational Biology, Iowa State University, Ames, IA 50011, USA

[2]Department of Genetics, Development and Cell Biology, Iowa State University, Ames, IA 50011, USA

[3]Department of Agronomy, Iowa State University, Ames, IA 50011, USA

Modified from a manuscript published in *Plant Phenomics*

5.1 Abstract

Many newly observed phenotypes are first described, then experimentally manipulated. These language-based descriptions appear in both the literature and in community datastores. To standardize phenotypic descriptions and enable simple data aggregation and analysis, controlled vocabularies and specific data architectures have been developed. Such simplified descriptions have several advantages over natural language: they can be rigorously defined for a particular context or problem, they can be assigned and interpreted programmatically, and they can be organized in a way that allows for semantic reasoning (inference of implicit facts). Because researchers generally report phenotypes in the literature using natural language, curators have been translating phenotypic descriptions into controlled vocabularies for decades to make the information computable. Unfortunately, this methodology is highly dependent on human curation, which does not scale to the scope of all publications available across all of plant biology. Simultaneously, researchers in other domains have been working to enable computation on natural language. This has resulted in new, automated methods for computing on language that

are now available, with early analyses showing great promise. Natural language processing (NLP) coupled with machine learning (ML) allows for the use of unstructured language for direct analysis of phenotypic descriptions. Indeed, we have found that these automated methods can be used to create data structures that perform as well or better than those generated by human curators on tasks such as predicting gene function and biochemical pathway membership. Here, we describe current and ongoing efforts to provide tools for the plant phenomics community to explore novel predictions that can be generated using these techniques. We also describe how these methods could be used along with mobile speech-to-text tools to collect and analyze in-field spoken phenotypic descriptions for association genetics and breeding applications.

5.2 Background

The volume of data related to phenotyping of plants is enormous and growing consistently. While sensor-based high-throughput technologies (described elsewhere in this issue) are responsible for much of this growth in phenotype data, text-based phenotype descriptions also contribute significantly. The scientific literature serves as the primary source of phenotype descriptions, where an example might look something like “maize line with specific mutation exhibits delayed flowering under stress condition .” Some phenotype descriptions find their way into model organism databases (e.g., TAIR, MaizeGDB, and SGN) through dedicated curation efforts (Berardini et al. (2015); Portwood et al. (2019); Fernandez-Pozo et al. (2015)).

Given the volume of phenotype descriptions available and the relevance of these descriptions to biological problems generally, interest in finding ways to compute on phenotypic descriptions is quite high. The most common method for making phenotypic descriptions computable involves representing the data using terms from large but finite and highly structured vocabularies such as the gene ontology (GO; Ashburner et al. (2000)), the plant ontology (PO; Cooper et al. (2013)), or the plant trait ontology (TO;Cooper et al. (2018)), among others (reviewed in Braun and Lawrence-Dill (2019)). The utility of using such vocabularies has been immense across the life sciences generally, with over 27,000 citations to the first GO publication alone (see Ashburner et al. (2000)). Use of these controlled vocabularies allows for increased consistency in how phenotypes are described, and the architecture of these data structures makes querying over a large volume of phenotypes realistic. Their hierarchical nature also enhances the meaning of each phenotype collected as a data point by inheriting implicit knowledge. For example, the GO hierarchy (Figure 5.1(a)) specifies that fruit ripening is a type of aging, so the association of a phenotype related to fruit ripening with this term allows that phenotype to be recovered by a query for aging, without that association being explicitly stated.

Despite the computational and inferential advantages that this type of annotation confers, detailed manual curation comes at the cost of the time and effort required to construct high-quality annotations for the large number of phenotypes observed, and the simplification of phenotypic descriptions to match the architecture of a particular knowledge representation necessarily reduces the specificity of a phenotypic description, thus losing some shades of meaning that are conveyed using natural language directly. How can these shortcomings be addressed? There are several applications for which unannotated natural language is becoming directly computable, a fact which has been largely underexploited in the biological disciplines.

The field of natural language processing (NLP) has made great advancement in recent years. NLP methods are used to compute on language directly to gain insights from semantic (meaning-based) and syntactic (structural) patterns. In the field of human health, applications of NLP with machine learning (ML) have been used to discover hidden patterns which can aid in informing patient care decisions. Such applications include text mining of medical records to predict probabilities of disease, machine translation of physician notes, and automated identification of articles relevant to disease phenotypes, to name just a few (reviewed in Ohno-Machado (2011)). These types of text analyses typically involve representing natural language using numerical vectors, which can then be used as inputs for ML models or to derive similarity scores (Figure 5.1(b)).

In a recent publication, we used NLP and ML to encode descriptions of plant phenotypes and measured pairwise similarity to construct similarity networks (Braun and Lawrence-Dill (2019)).

These computationally generated networks were shown to recover underlying gene functions and to predict membership in biochemical pathways, even on datasets distributed across multiple species. Most importantly, these computationally generated networks outperformed networks constructed using high-quality, ontology-based manual annotations in many cases, demonstrating that for these types of predictive tasks involving large datasets, applying computational methods over natural language descriptions yields comparable results to what can be achieved using a slower, labor-intensive, manual curation-based approach. Although high-quality curation plays an invaluable role in organizing phenotypic data, our findings suggest that there is much to be gained by applying purely computational approaches to phenotypic descriptions in plants.

5.3 What Do Phenotype Networks Look Like and How Can They Be Used?

Figures 5.1(c) and 5.1(d) illustrate what two types of similarity networks inferred from natural language descriptions of phenotypes look like. The first is useful for novel candidate gene prediction, and the second could become useful for genome-wide association studies (GWAS) through specification of a concept we call “synthetic traits” where clustered phenotypes are treated as a single trait.

For the novel candidate gene prediction application (Figure 5.1(c)), each node in the network refers to a particular gene and its corresponding phenotype. The similarity between two nodes implies an increased probability that the pair of genes is involved in a common regulatory network, biochemical pathway, or similar shared process. For example, two genes associated with phenotype descriptions that mention leaf size and shape are predicted to be involved in the same pathway or process. This sort of data structure enables researchers to generate new hypotheses about which genes may be involved in processes that generate a given phenotype.

For gene discovery, computationally generated phenotype similarity networks would be generated with no associations to genes asserted within the network (Figure 5.1(d)). In such a network, highly related phenotypes would create clusters, which we are defining as “synthetic traits.” Sequence data from plants with and without each synthetic trait could then be analyzed with well-understood GWAS approaches (Visscher et al. (2017)) to correlate specific genetic loci with the synthetic traits. This methodology could lead to the discovery of genes related to some phenotype properties that a researcher was not specifically looking to discover but that may be well represented in a specific growing environment by the germplasm under observation. For example, the graph may contain a cluster with words or phrases related to aerial root mucilage (Figure 5.1(d)) enabling this property to be used as a trait in downstream analyses like GWAS, even if this phenotype was not previously well understood (Van Deynze et al. (2018)). For collecting these data in a field environment, we envision phenotypic descriptions of plants being spoken and recorded, translated to text, then parsed computationally into specific statements. As such, this methodology is applicable to qualitative descriptions, rather than continuous numerical measurements. From there, the networks are created, highly interconnected clusters are identified as synthetic traits, and those traits are associated with genomic variants.

5.4 What Seems Unexpected (to Us) about the Use of Automated Methods for Computing on Phenotypic Descriptions?

The diversity of phenotype descriptions is beneficial to (rather than a hindrance to) this method of computing on the data. It is not necessary to standardize the words used to describe phenotypes for computational analysis, and the diversity of descriptions actually improves the quality of the result if enough phenotypic observations are recorded. By using data-driven approaches to specify synthetic traits, the concept of a trait becomes objective. This objectivity in grouping observations means that scientists may discover phenotype and trait groups that have not yet been conceived of and described previously. We are at the beginning of a new era for computing on phenotypic descriptions. In the past, researchers had to create simplified and structured descriptions to make phenotypes computable. Put another way, researchers were asked to think and behave like computers. Now, computational methods can accommodate the rich language that experts use to describe phenotypes. With NLP and ML, computers are able to reason like humans.

5.5 Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the contents of this manuscript or its publication.

5.6 Authors' Contributions

IRB, CFY, and CJLD contributed to the conception of the ideas presented here and to the writing and revision of the manuscript. Ian R. Braun and Colleen F. Yanarella contributed equally to this work.

5.7 Acknowledgments

We thank the Iowa State University Plant Science Institute Faculty Scholars and the Iowa State University Predictive Plant Phenomics graduate students for helpful discussions on the topics discussed in this manuscript. The authors were supported by an Iowa State University Presidential Interdisciplinary Research Seed Grant (CJLD), an Iowa State University Plant Sciences Institute Faculty Scholar Award (CJLD), the Predictive Plant Phenomics NSF Research Traineeship (DGE-1545453; CJLD is a coprincipal investigator; IRB and CFY are trainees), and a seed grant from National Science Foundation (OAC-1636865) that partially supported some work by IRB.

5.8 References

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. (2000). Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29.

Berardini, T. Z., Reiser, L., Li, D., Mezheritsky, Y., Muller, R., Strait, E., and Huala, E. (2015). The arabidopsis information resource: making and mining the "gold standard" annotated reference plant genome. *genesis*, 53(8):474–485.

Braun, I. R. and Lawrence-Dill, C. J. (2019). Automated methods enable direct computation on phenotypic descriptions for novel candidate gene prediction. *Frontiers in Plant Science*, 10:1629.

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

Cooper, L., Walls, R. L., Elser, J., Gandolfo, M. A., Stevenson, D. W., Smith, B., Preece, J., Athreya, B., Mungall, C. J., Rensing, S., et al. (2013). The plant ontology as a tool for comparative plant anatomy and genomic analyses. *Plant and Cell Physiology*, 54(2):e1–e1.

Fernandez-Pozo, N., Menda, N., Edwards, J. D., Saha, S., Tecle, I. Y., Strickler, S. R., Bombarely, A., Fisher-York, T., Pujar, A., Foerster, H., et al. (2015). The sol genomics network (sgn)—from genotype to phenotype to breeding. *Nucleic acids research*, 43(D1):D1036–D1041.

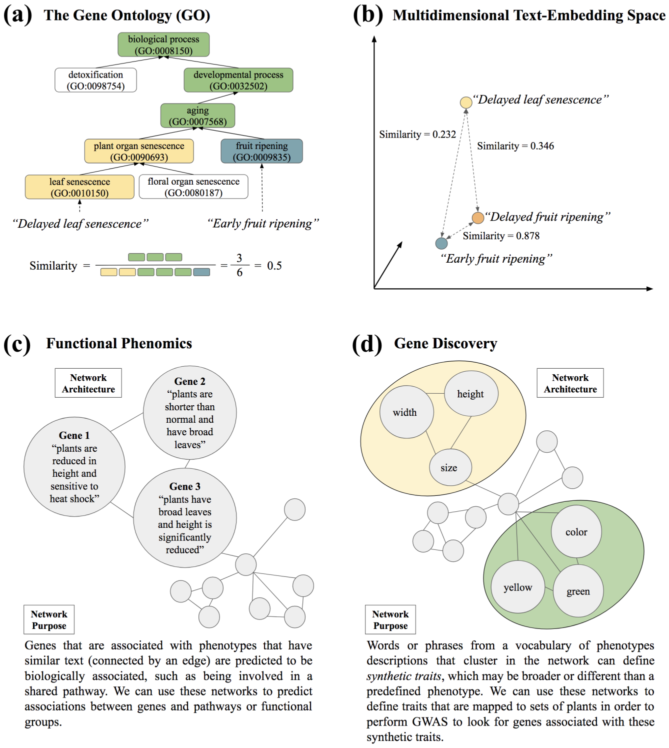
Ohno-Machado, L. (2011). Realizing the full potential of electronic health records: the role of natural language processing. *Journal of the American Medical Informatics Association*, 18(5):539–539.

Portwood, J. L., Woodhouse, M. R., Cannon, E. K., Gardiner, J. M., Harper, L. C., Schaeffer, M. L., Walsh, J. R., Sen, T. Z., Cho, K. T., Schott, D. A., et al. (2019). Maizegdb 2018: the maize multi-genome genetics and genomics database. *Nucleic acids research*, 47(D1):D1146–D1154.

Van Deynze, A., Zamora, P., Delaux, P.-M., Heitmann, C., Jayaraman, D., Rajasekar, S., Graham, D., Maeda, J., Gibson, D., Schwartz, K. D., et al. (2018). Nitrogen fixation in a landrace of maize is supported by a mucilage-associated diazotrophic microbiota. *PLoS biology*, 16(8):e2006352.

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., and Yang, J. (2017). 10 years of gwas discovery: biology, function, and translation. *The American Journal of Human Genetics*, 101(1):5–22.

5.9 Figures and Tables


(a) The Gene Ontology (GO)
biological process (GO:0008150)
detoxification (GO:0098754)
developmental process (GO:0032502)
aging (GO:0007568)
plant organ senescence (GO:0090693)
fruit ripening (GO:0009835)
leaf senescence (GO:0010150)
floral organ senescence (GO:0080187)
"Delayed leaf senescence"
"Early fruit ripening"
Similarity = 3/6 = 0.5
(b) Multidimensional Text-Embedding Space
"Delayed leaf senescence"
Similarity = 0.232
Similarity = 0.346
"Delayed fruit ripening"
Similarity = 0.878
"Early fruit ripening"
(c) Functional Phenomics
Network Architecture
Gene 2 "plants are shorter than normal and have broad leaves"
Gene 1 "plants are reduced in height and sensitive to heat shock"
Gene 3 "plants have broad leaves and height is significantly reduced"
Network Purpose
(d) Gene Discovery
Network Architecture
width
height
size
color
yellow
green
Network Purpose

Genes that are associated with phenotypes that have similar text (connected by an edge) are predicted to be biologically associated, such as being involved in a shared pathway. We can use these networks to predict associations between genes and pathways or functional groups.

Words or phrases from a vocabulary of phenotypes descriptions that cluster in the network can define *synthetic traits*, which may be broader or different than a predefined phenotype. We can use these networks to define traits that are mapped to sets of plants in order to perform GWAS to look for genes associated with these synthetic traits.

Figure 5.1 Phenotypic similarity. (a) For the GO, the similarity between two concepts can be evaluated based on the relationship between the sets of terms from the ontology that represent those concepts. This relationship can be quantified using metrics such as Jaccard similarity (shown). (b) Natural language processing technique such as sentence embedding using machine learning models or presence and absence of individual words can be used to produce high-dimensional vector representations of concepts, where their position within the vector space allows for quantification of similarity. The example shown plots concepts within three dimensions. (c) Example phenotypic similarity network where nodes represent genes and any associated phenotypic text descriptions. (d) Example phenotypic similarity networks where nodes represent words or phrases drawn from a set of descriptions about some population of plants.

CHAPTER 6. THE CASE FOR RETAINING NATURAL LANGUAGE DESCRIPTIONS OF PHENOTYPES IN PLANT DATABASES AND A WEB APPLICATION AS PROOF OF CONCEPT

Ian R. Braun[1,2], Diane C. Bassham[3], Carolyn J. Lawrence-Dill[1,2,3,*]

[1]Interdepartmental Bioinformatics and Computational Biology, Iowa State University

[2]Department of Agronomy, Iowa State University

[3]Department of Genetics, Development and Cell Biology, Iowa State University

[*]To whom correspondence should be addressed

Modified from a manuscript submitted to *bioRxiv*

6.1 Abstract

Finding similarity across phenotypic descriptions is not straightforward, with previous successes in computation requiring significant expert data curation. Natural language processing of free text phenotype descriptions is often easier to apply than intensive curation. It is therefore critical to understand the extent to which these techniques can be used to organize and analyze biological datasets and enable biological discoveries. A wide variety of approaches from the natural language processing domain perform as well as similarity metrics over curated annotations for predicting shared phenotypes. These approaches also show promise both for helping curators organize and work through large datasets as well as for enabling researchers to explore relationships among available phenotype descriptions. Here we generate networks of phenotype similarity and share a web application for querying a dataset of associated plant genes using these text mining approaches. Example situations and species for which application of these techniques is most useful are discussed.

6.2 Introduction

Phenotypes, defined as measurable characteristics or properties of an organism that result from interactions between genetics and the environment, comprise an enormous portion of the biological data that is considered important across a wealth of domains in the life sciences and beyond. Phenotypes are everything we see or measure in biology. On a more practical note, phenotypes encompass critical information related to human health and medicine, and important agronomic traits such as plant height and biomass of crop species. The scope of phenotypic information also ranges widely, from cellular phenotypes such as membrane composition or chemical concentrations, to community-level phenotypes like total leaf surface area in a field of crops. The extreme diversity in how phenotypes can be observed and represented makes handling this information on a computational level fundamentally different than genomic data, which lends itself to computational means of representation and analysis based on the existing natural codes of bases and amino acids (reviewed in Braun et al. 2018). This is especially true for phenotypes that are qualitative in nature, such as abnormal morphology, rather than phenotypes that are easily translated into a quantitative value, such as height (reviewed in Yanarella et al. 2020).

Despite these challenges, bio-ontologies have greatly helped to enable computation on phenotypic information by providing standardized, hierarchical sets of descriptors (terms) that can be used to annotate phenotypic information. Doing so enables comparison between data points, including comparisons across multiple species, studies, and sources in a meaningful way, which has contributed to the use of these data structures in recent years. Using terms from the Gene Ontology (GO; Ashburner et al. 2000) to describe cellular components, functions, and processes allows researchers to quickly find genes related to a biological concept of interest, and to understand which biological processes are potentially carried out or influenced by a group of genes of interest (Huang et al. 2009). Using this same ontology as the format for predictions about gene functions allows datasets of predicted gene functions to be seamlessly incorporated with and compared to known information (Zhou et al. 2019). Biomedical vocabulary graphs such as the Human Phenotype Ontology (HPO; Robinson et al. 2008) and Disease Ontology (DO; Schriml et al. 2012) allow for organization and interoperability of the vast and growing body of knowledge surrounding human medicine. Efforts such as Phenoscape (Edmunds et al. 2015), the Monarch Initiative (Mungall et al. 2017), and Planteome (Cooper et al. 2018), use ontologies to provide common data representations and allow for comparisons across diverse species or across evolutionary history.

At the same time, both the performance and availability of natural language processing (NLP) and machine learning (ML) methods for working with natural language and text data have continually improved. This is largely due to recent and continued innovations in how neural networks are designed to handle this type of information (Mikolov et al. 2013; Le and Mikolov 2014; Vaswani et al. 2017), and how they can be trained on massive volumes of unlabeled data (such as Wikipedia or PubMed) to provide systems for accurately modeling text in computable formats, and allowing for transferring to other domains and fine-tuning for more specific problems (Devlin et al. 2018, Wolf et al. 2020). One result of this progress is that such techniques now represent a complementary approach for computationally handling the diversity of phenotypic information, at least for cases where phenotypes are represented as text descriptions. Given that phenotypes have been described in academic articles for more than a century, sources for phenotypic descriptions abound. Although the vast majority of phenotypes described in the literature have not been extracted and represented in computationally accessible community databases, some databases do exist that contain phenotype descriptions in free text fields.

Previously, we demonstrated that for some organizational tasks (like grouping functionally similar genes together) using computational approaches that process text descriptions of phenotypes can work as well or better than using curated ontology term annotations to create similarity measurements (Braun and Lawrence-Dill 2019). Here, we demonstrate that this finding holds true for a larger dataset of the available phenotype text descriptions from across six different plant species. This means that where available, text descriptions of phenotypes have the potential to provide useful biological insight when combined with a variety of methods from the field of NLP. We therefore make a case for expanded inclusion of free text descriptions as a valuable component of biological databases going forward, whether as a supplemental data type to more standardized ontology term annotations, or as a potential short-term alternative for species currently lacking the curatorial resources to produce large scale datasets of high-confidence, curated annotations.

In demonstrating the utility of analyzing text descriptions of phenotypes with NLP approaches, we focus on what can be learned from evaluating similarity between descriptions as a measure of gene pair similarity. This is closely comparable to the ongoing problem in NLP of measuring sentence similarity, which has applications for text querying, text classification, and other tasks (De Boom et al. 2016). An enormous variety of solutions have been put forward for this problem, including both general solutions as well as more narrowly focused solutions for working in particular domains, such as biomedical literature (Soğancıoğlu et al. 2017, Chen et al. 2019). The number of solutions to this task is related to the fact that virtually all approaches for dealing with text computationally involve representing words or sentences as numerical vectors, on top of which similarity or distance metrics can then be applied to quantify relatedness between the two texts. In other words, all approaches for vectorizing text, which is typically the first step in handling any problem with NLP, can subsequently be used to find similarity between two texts by applying a similarity metrics to their vector representations. This enables the generation of networks for organizing data across large datasets. In this work, we assess the performance of a variety of both simple and state-of-the-art methods for translating plant phenotype descriptions into numerical vectors and build networks that can be used to make inferences from pairwise similarities.

We will also discuss and demonstrate how these same techniques can be applied for organizing and analyzing large datasets of text descriptions of phenotypes, accounting for phenotypic characteristics that have not yet been explicitly defined by the input data. Finally, we also provide a web application that enables others to explore and make use of phenotypic similarities identified. The application, called QuOATS (Querying with Ontology Annotations and Text

Similarity), can be used to search for plant genes with similar phenotypic descriptions using gene identifiers, ontology terms, keywords, or similarity to searched phenotype descriptions as input.

6.3 Methods

6.3.1 Datasets

We collected a dataset of available phenotype descriptions that have been mapped to specific plant genes, primarily through mutation studies, from the model species databases TAIR (Berardini et al. 2015), MaizeGDB (Portwood et al. 2019), and SGN (Fernandez-Pozo et al. 2015), as well as from a dataset of phenotype descriptions created by Oellrich et al. 2015. After merging data from multiple sources and preprocessing the texts, the combined dataset consisted of 7,907 genes from 6 plant species, with the quantity of genes and the text describing their associated phenotypes varying across species (Table 1). The distributions of sentences and words quantities present per gene also vary broadly across species (Figure 6.1). Portions of the vocabulary used to describe phenotypes in each of the species are unique to that particular species, but in all cases more than 80% of the vocabulary was shared with at least one other species (Figure 6.2).

For the genes in this dataset, we also collected three types of ontology term annotations: Gene Ontology (GO; Ashburner et al. 2000) annotations, Plant Ontology (PO; Jaiswal et al. 2005; Cooper et al. 2013) annotations, and entity-quality (EQ) statements composed of multiple ontology terms. For in-depth discussion on how EQ statements are composed and compared to one another, see (Hoehndorf et al. 2011; Oellrich et al. 2015; Braun and Lawrence-Dill 2019). GO and PO annotations were additionally sourced from the model species databases (Berardini et al. 2015; Portwood et al. 2019; Wimalanathan et al. 2018) and Planteome (Cooper et al. 2018; http://www.planteome.org), and were limited to those with evidence codes indicating they were either experimentally determined or created through author or curator statements (Consortium 2012; Giglio et al. 2019). The EQ statements were sourced from the dataset of curator-defined EQ statements created by Oellrich et al. 2015. Not all genes in the dataset had at least one annotation of each type, and these quantities are given in Table 6.1. The preprocessed, merged, and cleaned dataset described here is available and further described through a dedicated repository (see Code and Data Availability).

We also mapped the genes in this dataset to objects from additional bioinformatics resources, namely biochemical pathways in KEGG (Kanehisa et al. 2002) and PlantCyc (Schläpfer et al. 2017), protein-protein associations in STRING (Szklarczyk et al. 2016), ortholog relationships in PANTHER (Thomas et al. 2003), and a hierarchical Arabidopsis gene classification based on phenotypes (Lloyd and Meinke 2012). A subset of the genes in the complete dataset are found in each of these resources (Table 6.1).

6.3.2 Measure of gene pair similarity

We used a set of approaches for generating n by n pairwise similarity matrices, where n is the number of genes in the dataset, and the values in the matrix are some measure of the similarity between a given pair of genes. Each approach yields one matrix. The approaches belong to two main groups: text-based approaches that translate the text descriptions of phenotype(s) associated with each gene into numerical vectors, so that gene pair similarity can then be found using cosine similarity, and curator-based approaches, that rely on similarities between existing annotations for each gene (GO terms, PO terms, or EQ statements) to quantify gene pair similarity. Each of the text-based approaches used is described in overview here, as well as how the curator-based approaches determine gene pair similarity from annotations.

6.3.2.1 Tokenizing sentences

For each of the text-based approaches, we determined the effects of treating the entirety of the phenotype descriptions associated with a gene as one concatenated text, and comparing between those texts for pairs of genes to measure gene pair similarity, or by first tokenizing (separating) the phenotype descriptions into individual sentences, and treating those sentences as individual text instances. Then the maximum similarity scores obtained by any pair of sentences between those for those genes was taken as the gene pair relatedness score. This measure is intended to alleviate the effects of genes with longer phenotype descriptions seeming to appear unrelated to ones with shorter ones, and is analogous to looking for local alignments in the text, rather than global ones. In the subsequent Methods sections, we use the word ‘text’, to mean either the concatenation of all phenotype descriptions associated with a gene, or a single sentence from those descriptions, depending on which of these two methods is being described. Sentence tokenization was done with the NLTK package (Bird et al. 2009).

6.3.2.2 Baseline approach

Some genes in the collected dataset have identical phenotype descriptions. As a baseline approach against which to compare the subsequently described approaches, we include an approach that simply yields a similarity value of 1 for gene pairs that have identical texts, and 0 for gene pairs with texts that differ in any way, after preprocessing.

6.3.2.3 TF-IDF

Constructing tf-idf (term frequency-inverse document frequency) vectors is one of simplest ways of representing text in a computable format. With this approach, phenotype descriptions are treated as a bag-of-words, and translated to a vector which is the same length as the total number of unique words in the dataset vocabulary, where each position in the vector corresponds to a particular word. The value at the position in the vector for a particular word is the number of times that word appears in the phenotype description (term frequency) weighted by the inverse of the fraction of phenotype descriptions in which that word appears (inverse document frequency). Weighting by the inverse document frequency emphasizes the importance of rarer words (e.g., ‘gametophyte’) and de-emphasizes the importance of more common words (e.g., ‘plant’) in the vector encoding. In addition to this straightforward implementation of the tf-idf approach, we also used as a bigram approach where positions in the vector represent a sequence of two consecutive words (as opposed to the unigram approach, where positions are a single word, as described above). We also used a tf-idf monogram approach where the phenotype descriptions in the datasets are first subset to only include words that are over-represented in journal articles abstracts related to plant phenotypes. The criteria for inclusion was that a word appeared at least twice as frequently in the dataset of plant phenotype related abstracts compared to a general domain corpus. In all cases, cosine similarity was used to calculate gene pair similarity after phenotype descriptions were translated into vectors.

6.3.2.4 Computational annotation (NOBLE Coder)

NOBLE Coder (Tseytlin et al. 2016) is a computational tool for annotating text with ontology terms. We used NOBLE Coder to annotate phenotype descriptions with terms from a set of bio-ontologies (GO, PO, and PATO), inheriting additional terms using the hierarchical structure of the ontologies. We used NOBLE Coder with both the exact and partial match parameters, which alters how strictly an ontology term must match a text string for an annotation to be assigned. After assigning terms to phenotype descriptions for genes by this method, each gene is represented by a set of terms rather than a set words, and the process of translating these representations into numerical vectors and calculating gene pair relatedness using cosine similarity is the same as with the tf-idf approach, with positions in the resulting vectors referencing terms instead of words. Again, cosine similarity was applied to yield similarity matrices from these resulting vectors.

6.3.2.5 Topic modeling (LDA and NMF)

We used Latent Dirichlet-Allocation (LDA; Blei et al. 2003) and Non-negative Matrix Factorization (NMF; Lee and Seung 1999) to perform topic modelling on the dataset of phenotype descriptions. These are decomposition algorithms that are widely used in NLP applications (reviewed in Jelodar et al. 2019), and result in translating a document-term matrix into a document-topic matrix (in our case, documents are phenotype descriptions). If the algorithm is run to learn 10 topics, then the outcome is that each phenotype is represented by a vector of length 10 where each position indicates the probability that the phenotype is derived from that particular topic. Determining the appropriate number of topics to use for a particular dataset is often a matter of trying a range of values, and looking at which value produces the most coherent or logical topics given the subject matter. Based on the word probability distributions created using a range of topic quantities, we used our best judgement to elect to use 50 topics and 100 topics for our embedding approaches using each of these algorithms.

6.3.2.6 Neural network-based embeddings (Word2Vec, Doc2Vec, BERT, BioBERT)

We also used machine learning approaches designed to find vector embeddings that represent the semantics of input text in a compressed space, with positions in the embedding representing abstract semantic features. Word2Vec (Mikolov et al. 2013) is an approach for generating word embeddings based on the contexts in which words appear in a corpus. We used a skip-gram model, where a shallow network is trained to take one word at a time from our corpus as input and predict surrounding context words. The result of this self-supervised training step is a vector embedding for each word that occurs in the dataset of descriptions that reflects the context those words appear in, in a compressed feature space (200 dimensions). To supplement our dataset of phenotype descriptions to build a larger corpus for self-supervised training, we shuffled in sentences accessed from PubMed that were present in abstracts retrieved with queries for the word 'phenotype' and any of the names of the species present in our dataset. Hyperparameters for model construction were selected through a validation task of predicting whether ontology term names and synonyms from PATO and PO were parent-child or sibling pairs, or more distantly related. This validation task led to the selection of a skip-gram model using a window size of 8, and a hidden layer size of 200 (see genism package (Rehurek and Sojka 2010) for parameter details). In addition, as a point of comparison, we also used pre-trained published models trained on PubMed (Moen and Ananiadou 2013) and Wikipedia (Lau and Baldwin 2016).

Doc2Vec is an extension of Word2Vec that either exclusively learns embeddings for documents (texts with multiple words) or learns embeddings for documents simultaneously with word embeddings. We used a distributed bag of words architecture where the arbitrary document tags are used as an input in a self-supervising to predict randomly selected words form the input documents, resulting in network architecture that can be used to infer document-specific embeddings (Le and Mikolov 2014). We utilized the same training approach as for word embeddings, using only concept pairs with multiple words as validation data. In addition, we used a pre-trained Doc2Vec model trained on Wikipedia (Lau and Baldwin 2016).

BERT (Bidirectional Encoder Representations from Transformers) is a large-scale neural network architecture trained on large unlabeled text datasets to predict masked words in sentences and predict whether one sentence follows another in a corpus (Devlin et al. 2018). This results in a network where the encoder can be used to generate context-specific vector embeddings for words in an input sentence. We used both the BERT base model (Devlin et al. 2018) and BioBERT models fine-tuned on abstracts from PubMed and articles from PubMed Central (Lee et al. 2020).

The Doc2Vec models were used to directly infer vector embeddings for phenotype descriptions. The Word2Vec and BERT models generate vector embeddings for each word in phenotype descriptions, so these individual word-embeddings were combined to produce a single vector embedding for each phenotype description. Whether the vectors are summed or averaged is a hyperparameter choice, along with how many encoder layers are used to build the BERT word vectors, and whether those layers should be summed or concatenated. These hyperparameter choices were made using performance on the validation task described previously for the networks trained on phenotype descriptions, and for the pre-trained models we selected hyperparameters based on their performance on a related biomedical sentence similarity problem with the BIOSSES dataset (Soğancıoğlu et al. 2017), and went forward with the hyperparameters that provided the best results on that separate dataset. As with the other approaches, cosine similarity was applied to the resulting vectors to yield similarity matrices.

6.3.2.7 Using embeddings to generate meaningful vectors with word replacement

Producing the most informative vector representations of phenotype descriptions requires combining the tf-idf approach of explicitly representing the quantity of each particular word from the vocabulary that is present in each phenotype description, and also accounting for semantics through learning vector embeddings of particular words relative to their own meanings in this vocabulary or their meaning relative to the words around them in these phenotypes. We used an approach where pairwise word-similarity matrices for each word in the vocabulary as represented by our Word2Vec models were used to replace each word in all descriptions with the most common word in the vocabulary out of the word itself and the three other most similar words predicted by that model (algorithm detailed in Pontes et al. 2016). This results in substitutions such as 'susceptible' to 'resistance' that may allow comparisons to be made between phenotypes that simpler bag-of-words approaches would consider as distinct. The resulting vector representations are tf-idf vectors, but the semantic relationships between words as informed by the neural network models is already account for prior to encoding.

6.3.2.8 Curated annotations (GO, PO, EQ statements)

For a point of comparison to the text-based approaches described above, we also used the curator-based annotations to quantify gene pair relatedness. For GO and PO annotations, we calculated similarities as the maximum information content of any single term shared between the annotation sets for a given pair of genes. The more similar two sets of annotations are, the more specific (with higher information content) the terms shared between the two sets will be with respect to the ontology graph structure, leading to greater similarity. In this case, information content is transformed to be in the range of 0 to 1, so that it can be used as a similarity metric compatible with the other approaches used. To quantify similarity between genes using EQ statements, we used the pairwise similarities provided in Oellrich et al. 2015.

6.3.3 Formulating Biologically Relevant Questions

We used additional bioinformatic resources (KEGG, PlantCyc, STRING, PANTHER, etc.) to assess representation of biologically relevant relationships between gene pairs in the dataset, that each approach described above can attempt to recover by quantifying the similarity for that pair of genes, allowing for direct comparison among the approaches (Table 6.2). Because not all genes in the dataset map to each resource (Table 6.1), the number of gene pairs that are applicable to each question are not consistent (Table 6.3). Although these questions are likely related to one another in terms of true biology (e.g., if a pair of genes are related to the same observable phenotype, they are probably more likely to act in a shared pathway), these questions are neither identical nor redundant in the context of this work, because different questions apply to different portions of the dataset, and even within the overlaps of gene pairs that apply to multiple questions, the set of positives (gene pairs for which the correct answer is 'true') are not the same (Table 6.4). For example, the two most similar tasks are 'Associations' and 'Pathways', where 1,271,297 of the same gene pairs are considered in both tasks, and the Jaccard similarity between the two sets of target values ('true', 'false') between those gene pairs is only 0.172 (Table 6.4). For this reason, we looked at the results of each of these questions individually rather than combining them.

6.4 Results

6.4.1 Text-based approaches recover biological relationships

Using each of the text-based approaches as well as using similarity metrics over the existing curated annotations, we calculated gene pair similarity values for all pairs of genes in our dataset. We measured the success of each approach for (1) predicting whether two genes were orthologs (as specified in PANTHER), (2) predicting known protein associations specified in STRING, (3) predicting whether two genes functioned in at least one of the same biochemical pathways (as specified PlantCyc and KEGG), and (4) at predicting whether two Arabidopsis genes belonged to one of the phenotype categories specified by Lloyd and Meinke 2012. For each of these biological questions, a given approach for measuring gene similarity is considered useful if the distribution of values for gene pairs for which the answer to the question is true is distinct from the distribution of values for gene pairs for which the answer to question is false. The success of each approach for each biological question was calculated in terms of the maximum F_1 statistic. We also recalculated the maximum F_1 statistic for just the genes for which we have GO annotations, PO annotations, and EQ statements, to directly compare performance of each approach on each question with approaches that are based on curation (Table 6.5, Supplemental Table 6.6).

6.4.1.1 Text-based approach performance is dependent on biological query type

Of the four biological questions assessed for this analysis, predicting whether two genes were orthologous, whether two genes shared an association, or whether two genes belonged to a shared biochemical pathway were infeasible for any of the text-based or curation-based approaches, in terms of broad performance measured with maximum F_1 statistics (Table 6.5, Supplemental Table 6.6). The largest F_1 statistic obtained across all three of these tasks for any approach was 0.140 using the curated GO annotations, with all other approaches yielding F_1 values less than 0.12 (Table 6.5, Supplemental Table 6.6). However, F_1 statistics were much higher for the task of predicting whether two genes belonged to the same phenotypic category, an expected result given that this prediction follows directly from the explicit contents of the phenotypic descriptions (Table 6.5). This was true for both the text-based and curation-based approaches, but the best performance was achieved using text-based approaches (Table 6.5). Performance on this task of predicting whether two genes share a phenotypic category can be broken down by general classes of approach (Figure 6.3).

As previously stated, all approaches were unsuccessful in predicting ortholog relationships (Supplemental Table 6.6). In addition, all approaches were completely unsuccessful in predicting whether two genes from different species were involved in a common biochemical pathway (Supplemental Table 6.7). Even though the maximum F_1 statistics for predicting whether two genes share a pathway were already low, these values were even lower when filtering the dataset to only look at interspecies gene pairs, and marginally greater when filtering the dataset to only look at intraspecies gene pairs (Supplemental Table 6.7). Therefore, even the very small amount of biological information recovered only applies to looking at genes from within the same species. This indicates that comparing the text of phenotype descriptions across different species is not biologically informative in this case. This might not be true for all species or all phenotypes, but it does not generalize across the current dataset of available plant phenotype descriptions.

6.4.1.2 Significant description similarity within individual phenotype and pathway gene groups

Although predicting whether two genes shared a biochemical pathway was generally unsuccessful (as evaluated by low maximum F_1 values), this is in part a consequence of the fact that pathways vary greatly in how related the phenotype descriptions for their component genes are. We evaluated this property by plotting the average gene-to-gene similarity for all possible gene pairs in each individual pathway, as a percentile of the similarities between all genes pairs (Figure 6.4. As a point of reference, we repeated this analysis for phenotypic categories, where larger F_1 values were obtained. We randomly sampled groups of genes at each value of n to calculate p-values for each phenotype category and pathway, calculating the probability of each approach generating a mean similarity value between genes in that group that is that large or larger, controlling for false discovery rate for each approach with the Benjamini–Hochberg procedure (Table 6.5). For text-based approaches using sentence tokenization, 81% to 100% of the phenotypic categories had a significantly large average similarity value (with respect to the Benjamini–Hochberg procedure), while between 6% and 39% of the pathways obtained significant average similarity values, for these same approaches, with an average of 23% (Table 6.5). Taken together, these results indicate that while text-based similarity values are not broadly indicative of whether or not two genes share a pathway, there is a significant subset of known pathways for which this is the case. In the case of groups of genes belonging to the same pathway that do have similar phenotype descriptions, these are generally either due to mentions of downstream phenotypic effects of pathway disruption, or more direct mentions of the pathway function or role. For example, the descriptions associated with genes in the chlorophyll degradation pathway include mentions of necrotic lesions, and the descriptions associated with genes in the phospholipid desaturation pathway include mentions of fatty acid levels or composition.

6.4.1.3 Combining syntactic and semantic approaches improves recovery of phenotypic categories

The purely syntactic text-based approaches (tf-idf) were among the most successful in terms of maximum F_1 statistic for predicting whether gene pairs belonged to the same phenotypic category (Table 6.5, Figure 6.3). In general, semantic approaches that use machine learning techniques to drastically reduce the dimensionality of the vector encoding for each text instance were comparably successful (Table 6.5, Figure 6.3). However, the combined approaches where semantic techniques were used to augment the information in the tf-idf vectors by replacing words with similar words prior to encoding provided a boost in performance over other approaches (Table 6.5, Figure 6.3). Taken together, this indicates that this dataset contains phenotype descriptions for genes in the same phenotypic category that are similar both in terms of explicitly shared words (where syntactic approaches are most helpful), as well as genes that are similar only in terms of shared meaning but not specific words (where semantic approaches provide an advantage). Using word embedding models trained on plant phenotype specific data provided marginal improvement over models trained on PubMed generally or the Wikipedia corpus, but all three models provided the same boost over other approaches when applied to word replacement, indicating that useful associations between words for recovering common phenotypic categories from descriptions are not limited to relationships only represented in a narrow corpus of text related to plant phenotypes. Given that using bio-ontologies for this same task did not perform as well as text-based approaches, and one of the main functions of such ontologies in this case is to inject domain-specific inferences into the similarity metrics, this result is not surprising.

6.4.1.4 Sentence tokenization is important for comparing phenotypes

For all the text-based approaches on all the biological questions posed, the preprocessing step of tokenizing phenotype descriptions into sentences and evaluating gene pair relatedness as the maximum pairwise sentence similarity resulted in greater F_1 statistics (Table 6.5, Table 6.6). Unexpectedly, this held true even for approaches that are generally intended for use with larger input texts, such as Doc2Vec, and topic modeling algorithms LDA and NMF. This indicates that when predicting whether two genes share a common role, it is important to account for 'local alignments' in their associated phenotype descriptions, as the similarity might exist between single sentences associated with those genes while other sentences act as noise obscuring this relationship.

6.4.2 Enabling biologists to use these methods and dataset

6.4.2.1 Web application (QuOATS)

We have developed a web application called QuOATS (Querying with Ontology Annotations and Text Similarity) for querying the dataset described here through leveraging the computational methods described here (Figure 6.5A). The underlying dataset of plant genes is the same as is described previously (Table 6.1), and can be filtered to include particular species (Figure 6.5B). The application supports four different query types (Figure 6.5D), with the primary purpose being to obtain lists of genes that are related to phenotypes described similarly to some phenotypic characteristic(s) or if interest. Firstly, a free text query can be used to search the dataset for any genes related to phenotypes that are described similarly to text strings separated by periods in the query (Figure 6.5E). Secondly, a keyword query can be used to input any number of strings of any length, and genes whose phenotype descriptions contain those strings (after preprocessing including stemming and case-normalization) are returned (Figure 6.5F). Thirdly, an ontology term query can be used to search for any genes annotated by curators with one or more ontology terms, either directly or inherited through the ontology hierarchy (Figure 6.5G). Lastly, a gene identifier query can be carried out to search for any gene name, protein name, gene model, or any other gene identifier potentially represented in the dataset. Selecting a gene from the returned list of candidates that match the query will auto-complete a second query that returns genes related to phenotypes that are described similarly to the selected genes (Figure 6.5H).The similarity scores used to rank genes in the returned list are calculated using approaches described here, selected from a drop-down menu in the web application (Figure 6.5C).

6.4.2.2 Proof of concept applications of the web tool

In our previous findings illustrated in Braun and Lawrence-Dill 2019, we discussed how a set of genes related to anthocyanin biosynthesis could be used to demonstrate recovering gene groups by querying specifically with phenotype descriptions or computationally generated annotations from those descriptions. Specifically, we looked at a dataset of 16 maize genes (Li et al. 2019) and 21 genes from Arabidopsis (Appelhagen et al. 2014) but only 10 of the maize genes and 16 of the Arabidopsis genes were present in the dataset. Our expanded dataset in this work includes 14 of those maize genes and 18 of the Arabidopsis genes. We now evaluate the results of querying with each of these genes in the web application QuOATS, to recover both genes in the same species from these sets and genes in the alternate species. Over the 64 total queries (32 within the same species and 32 between species), we quantified the average and standard deviation of the number of target genes contained in bins of ranks in the query results, in bin sizes of 10 up to 50, and a final bin for genes that obtain ranks higher than 50 (Figure 6.6). Additionally, we also repeated this analysis for a set of 9 core autophagy genes in Arabidopsis (Figure 6.6). These queries illustrate a proof-of-concept for using this web application to use phenotypic descriptions associated with one gene to recover other related genes. This application demonstrates the utility of applying text-based algorithms in cases where ontology annotations are either not present, are insufficient, or could simply be augmented for allowing additional, less rigidly-defined phenotype descriptions to be searchable as well.

6.5 Discussion

The difficulty in computing on phenotypic data is largely a consequence of extreme variability with which these data are represented, and the diversity of ways that phenotypes are measured, quantified, and described. This is in contrast with sequence data; biology as a field has enormously benefited from the ways in which sequence data are intuitively computed on, given the naturally occurring nucleic acid and amino acid coding systems. Sequencing technology provided the datasets to compute on, and algorithms and applications like BLAST provided the means to make use of these data. Ontologies have begun to provide a similar means for making phenotypic data computable, and processing of natural language provides an additional avenue by which we can make biological inferences if we have the datasets on which to apply them. The combination of biological ontologies, machine learning approaches, and NLP provide strategies for handling phenotypic descriptions and learning from it where it exists.

Plant phenotypes are frequently described as text within academic papers or research notes. However, these text descriptions are rarely incorporated into relevant research community databases, associated with a specific gene or genotype, and made readily available as part of the growing data resources for that species. This could be the case for a variety of reasons, including the difficulties involved with extracting phenotype descriptions from larger texts, the curatorial effort necessary to produce high quality datasets of phenotypes descriptions associated with genes, or because these text representations of phenotypes are considered a non-valuable data type, and are instead represented by annotations using structured vocabularies of hierarchical terms such as biological ontologies. Notable exceptions to this situation exist, including The Arabidopsis Information Resource (TAIR), which contains thousands of text descriptions of phenotypes mapped to specific Arabidopsis genes (Berardini et al. 2015).

In this work, we have shown that a variety of NLP approaches for vectorizing phenotype descriptions in order to generate gene pair similarity matrices are equally or more predictive in general of known phenotype categorizations compared to using existing curated annotations for this task. Based on these results, we argue that it is worthwhile for databases that contain gene-to-phenotype information to include natural language descriptions of phenotypes.

The natural language descriptions of phenotypes are useful, and when combined with NLP approaches for computationally representing text can be leveraged to provide a way for researchers to quickly identify genes associated with phenotypes similar to the ones that they are observing or studying. Natural language descriptions can also be used to organize genes computationally on a large scale and discover which categorizations of phenotypes are present in a dataset, with techniques like clustering and topic modeling. In some cases, this natural language data may be easier to generate than ontology annotations. In situations where curators are not available (or have limited time) to generate the high-confidence ontology term annotation datasets, it may be faster or still possible for authors or someone else to at least identify the free-text portions of the manuscript that include phenotype descriptions, and the genes associated with them. In the near future, NLP techniques for parsing full-texts may also progress to the point where this phenotype identification could be done automatically as well. In these instances, we argue it is worthwhile to generate and make accessible this free text phenotype information. In other cases, these text data might already be generated, but are potentially discarded. In situations where curators are actively involved in generating ontology annotations from papers, this process often involves the tasks of highlighting text from the paper, or possibly writing down the phenotype descriptions first then producing the ontology term representation of those associations. Given that the free text itself is useful, we argue it should be retained in the final mapping in the resulting database or dataset rather than being discarded as an intermediate data form. It is possible that for some applications the ontology annotations will be more useful than the natural language descriptions, for example when making comparisons between species, but we have shown that this is not always the case, and if it is being generated regardless, it makes sense to retain the natural language and make it available.

We envision that the area where the application of these methods would make the most difference is in the case of species where phenotype descriptions have not already been thoroughly standardized and special vocabularies have already been created and assigned to phenotype data in a pervasive, large-scale way (as is the case with human phenotypes and diseases), but where phenotypes are still largely described in general biological terms. In addition, these approaches would make the most sense to use when high quality, curated datasets of ontologized phenotypes are either not available or not financially feasible. In these cases, if at a minimum phenotype descriptions are extracted from literature and associated with specific genes in an accessible community database, these methods can be applied to organize these data, group by genes into sets that impact similar phenotypes, and allow researchers to search based on linguistic similarity.

In 2011, Mike Freeling made an impression by saying, “Ontologies are for people who don’t understand their phenotype,” to CJLD at the Annual Maize Meeting Genetics Conference in response to a request to review the completeness of the MaizeGDB Phenotypic Controlled Vocabulary (Michael Freeling personal communication). While ontologies have proved invaluable for managing and analyzing the massive quantity of data that biologists deal with, we think that this quote emphasizes the key finding for the efforts here: that we should not undervalue the utility of free text as a datatype, and that it should be made available through bioinformatic resources that provide phenotypic data to the research community, given that we have the computational tools to leverage it in useful ways. Not only do plant scientists understand their phenotypes and use rich language to describe them, there is a diversity of algorithms available to enable computation on phenotypic descriptions so that the scope of data any single researcher can access becomes quite expansive.

Data and Code Availability

The dataset of plant genes collected from other sources for this work is available at https://git.io/JTutQ, along with all the code for preprocessing, reshaping, and merging this data. The code for carrying out the analysis shown here has its own repository at https://git.io/JTutN. The results given here can be reproduced using code and datasets at those locations. In addition, a Python package called OATS (Ontology Annotation and Text

Similarity) for working with gene-phenotype datasets, ontology annotations, and free-text was developed in parallel with this work. This package was used extensively for this analysis, and can be found at https://git.io/JTuqv, with documentation available at https://irbraun-oats.readthedocs.io. We have combined the dataset and some of the techniques for identifying similar texts into a streamlit web application named QuOATS available at https://quoats.dill-picl.org/. Use this tool for looking up genes by phenotype keywords or phrases, or finding genes with similar descriptions to a searched phenotype description. The code for this web application is available at https://git.io/Jtv9J.

Acknowledgements

The authors thank Carson Andorf, John Portwood, and Naama Menda for their assistance in obtaining the dataset of plant phenotype descriptions and annotations. The authors thank Darwin Campbell and Scott Zarecor for their assistance with dataset creation and hosting and maintenance of the web application discussed here. The authors thank Marna Yandeau-Nelson, Iddo Friedberg, Baskar Ganapathysubramanian, Annette O'Connor, and Qi Li for useful discussions on the topics discussed here, and particularly thank Marna Yandeau-Nelson for assistance in editing and Iddo Friedberg for assistance in reviewing the design of the web application.

Funding

This work has been supported by an Iowa State University Presidential Interdisciplinary Research Seed Grant (PI Diane Bassham; CJLD is a coPI), the Iowa State University Plant Sciences Institute Faculty Scholars Program to CJLD, the Predictive Plant Phenomics NSF Research Traineeship (#DGE-1545453) to CJLD (IRB is a trainee), and a USDA NIFA Agriculture and Food Research Initiative Predoctoral Research Grant (#2020-67034-31745) to IRB.

6.6 References

Appelhagen, I., Thiedig, K., Nordholt, N., Schmidt, N., Huep, G., Sagasser, M., and Weisshaar, B. (2014). Update on transparent testa mutants from arabidopsis thaliana: characterisation of new alleles from an isogenic collection. *Planta*, 240(5):955–970.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. (2000). Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29.

Berardini, T. Z., Reiser, L., Li, D., Mezheritsky, Y., Muller, R., Strait, E., and Huala, E. (2015). The arabidopsis information resource: making and mining the “gold standard” annotated reference plant genome. *genesis*, 53(8):474–485.

Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit.* ” O’Reilly Media, Inc.”.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Braun, I., Balhoff, J. P., Berardini, T. Z., Cooper, L., Gkoutos, G. V., Harper, L. C., Huala, E., Jaiswal, P., Kazic, T., Lapp, H., et al. (2018). ’computable’phenotypes enable comparative and predictive phenomics among plant species and across domains of life.

Braun, I. R. and Lawrence-Dill, C. J. (2019). Automated methods enable direct computation on phenotypic descriptions for novel candidate gene prediction. *Frontiers in Plant Science*, 10:1629.

Chen, Q., Peng, Y., and Lu, Z. (2019). Biosentvec: creating sentence embeddings for biomedical texts. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 1–5. IEEE.

Consortium, G. O. (2012). Gene ontology annotations and resources. *Nucleic acids research*, 41(D1):D530–D535.

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

Cooper, L., Walls, R. L., Elser, J., Gandolfo, M. A., Stevenson, D. W., Smith, B., Preece, J., Athreya, B., Mungall, C. J., Rensing, S., et al. (2013). The plant ontology as a tool for comparative plant anatomy and genomic analyses. *Plant and Cell Physiology*, 54(2):e1–e1.

De Boom, C., Van Canneyt, S., Demeester, T., and Dhoedt, B. (2016). Representation learning for very short texts using weighted word embedding aggregation. *Pattern Recognition Letters*, 80:150–156.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Edmunds, R. C., Su, B., Balhoff, J. P., Eames, B. F., Dahdul, W. M., Lapp, H., Lundberg, J. G., Vision, T. J., Dunham, R. A., Mabee, P. M., et al. (2015). Phenoscape: identifying candidate genes for evolutionary phenotypes. *Molecular biology and evolution*, 33(1):13–24.

Fernandez-Pozo, N., Menda, N., Edwards, J. D., Saha, S., Tecle, I. Y., Strickler, S. R., Bombarely, A., Fisher-York, T., Pujar, A., Foerster, H., et al. (2015). The sol genomics network (sgn)—from genotype to phenotype to breeding. *Nucleic acids research*, 43(D1):D1036–D1041.

Giglio, M., Tauber, R., Nadendla, S., Munro, J., Olley, D., Ball, S., Mitraka, E., Schriml, L. M., Gaudet, P., Hobbs, E. T., et al. (2019). Eco, the evidence & conclusion ontology: community standard for evidence information. *Nucleic acids research*, 47(D1):D1186–D1194.

Hoehndorf, R., Schofield, P. N., and Gkoutos, G. V. (2011). Phenomenet: a whole-phenome approach to disease gene discovery. *Nucleic acids research*, 39(18):e119–e119.

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research*, 37(1):1–13.

Jaiswal, P., Avraham, S., Ilic, K., Kellogg, E. A., McCouch, S., Pujar, A., Reiser, L., Rhee, S. Y., Sachs, M. M., Schaeffer, M., et al. (2005). Plant ontology (po): a controlled vocabulary of plant structures and growth stages. *Comparative and functional genomics*, 6(7-8):388–397.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., and Zhao, L. (2019). Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, 78(11):15169–15211.

Kanehisa, M. et al. (2002). The kegg database. In *Novartis Foundation Symposium*, pages 91–100. Wiley Online Library.

Lau, J. H. and Baldwin, T. (2016). An empirical evaluation of doc2vec with practical insights into document embedding generation. *arXiv preprint arXiv:1607.05368*.

Le, Q. V. and Mikolov, T. (2014). Distributed representations of sentences and documents. corr abs/1405.4053 (2014). *arXiv preprint arXiv:1405.4053*.

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2020). Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Li, T., Zhang, W., Yang, H., Dong, Q., Ren, J., Fan, H., Zhang, X., and Zhou, Y. (2019). Comparative transcriptome analysis reveals differentially expressed genes related to the tissue-specific accumulation of anthocyanins in pericarp and aleurone layer for maize. *Scientific reports*, 9(1):1–12.

Lloyd, J. and Meinke, D. (2012). A comprehensive dataset of genes with a loss-of-function mutant phenotype in arabidopsis. *Plant physiology*, 158(3):1115–1129.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Moen, S. and Ananiadou, T. S. S. (2013). Distributional semantics resources for biomedical text processing. *Proceedings of LBM*, pages 39–44.

Mungall, C. J., McMurry, J. A., Köhler, S., Balhoff, J. P., Borromeo, C., Brush, M., Carbon, S., Conlin, T., Dunn, N., Engelstad, M., 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.

Oellrich, A., Walls, R. L., Cannon, E. K., Cannon, S. B., Cooper, L., Gardiner, J., Gkoutos, G. V., Harper, L., He, M., Hoehndorf, R., et al. (2015). An ontology approach to comparative phenomics in plants. *Plant methods*, 11(1):1–15.

Pontes, E. L., Huet, S., Torres-Moreno, J.-M., and Linhares, A. C. (2016). Automatic text summarization with a reduced vocabulary using continuous space vectors. In *International Conference on Applications of Natural Language to Information Systems*, pages 440–446. Springer.

Portwood, J. L., Woodhouse, M. R., Cannon, E. K., Gardiner, J. M., Harper, L. C., Schaeffer, M. L., Walsh, J. R., Sen, T. Z., Cho, K. T., Schott, D. A., et al. (2019). Maizegdb 2018: the maize multi-genome genetics and genomics database. *Nucleic acids research*, 47(D1):D1146–D1154.

Rehurek, R. and Sojka, P. (2010). Software framework for topic modelling with large corpora. In *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Citeseer.

Robinson, P. N., Köhler, S., Bauer, S., Seelow, D., Horn, D., and Mundlos, S. (2008). The human phenotype ontology: a tool for annotating and analyzing human hereditary disease. *The American Journal of Human Genetics*, 83(5):610–615.

Schläpfer, P., Zhang, P., Wang, C., Kim, T., Banf, M., Chae, L., Dreher, K., Chavali, A. K., Nilo-Poyanco, R., Bernard, T., et al. (2017). Genome-wide prediction of metabolic enzymes, pathways, and gene clusters in plants. *Plant physiology*, 173(4):2041–2059.

Schriml, L. M., Arze, C., Nadendla, S., Chang, Y.-W. W., Mazaitis, M., Felix, V., Feng, G., and Kibbe, W. A. (2012). Disease ontology: a backbone for disease semantic integration. *Nucleic acids research*, 40(D1):D940–D946.

Soğancıoğlu, G., Öztürk, H., and Özgür, A. (2017). Biosses: a semantic sentence similarity estimation system for the biomedical domain. *Bioinformatics*, 33(14):i49–i58.

Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N. T., Roth, A., Bork, P., et al. (2016). The string database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic acids research*, page gkw937.

Thomas, P. D., Kejariwal, A., Campbell, M. J., Mi, H., Diemer, K., Guo, N., Ladunga, I., Ulitsky-Lazareva, B., Muruganujan, A., Rabkin, S., et al. (2003). Panther: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification. *Nucleic acids research*, 31(1):334–341.

Tseytlin, E., Mitchell, K., Legowski, E., Corrigan, J., Chavan, G., and Jacobson, R. S. (2016). Noble–flexible concept recognition for large-scale biomedical natural language processing. *BMC bioinformatics*, 17(1):32.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Wimalanathan, K., Friedberg, I., Andorf, C. M., and Lawrence-Dill, C. J. (2018). Maize go annotation—methods, evaluation, and review (maize-gamer). *Plant Direct*, 2(4):e00052.

Wolf, T., Chaumond, J., Debut, L., Sanh, V., Delangue, C., Moi, A., Cistac, P., Funtowicz, M., Davison, J., Shleifer, S., et al. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Yanarella, C. F., Braun, I. R., Lawrence-Dill, C. J., et al. (2020). Computing on phenotypic descriptions for candidate gene discovery and crop improvement. *Plant Phenomics*, 2020:1963251.

Zhou, N., Jiang, Y., Bergquist, T. R., Lee, A. J., Kacsoh, B. Z., Crocker, A. W., Lewis, K. A., Georghiou, G., Nguyen, H. N., Hamid, M. N., et al. (2019). The cafa challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome biology*, 20(1):1–23.

6.7 Figures and Tables

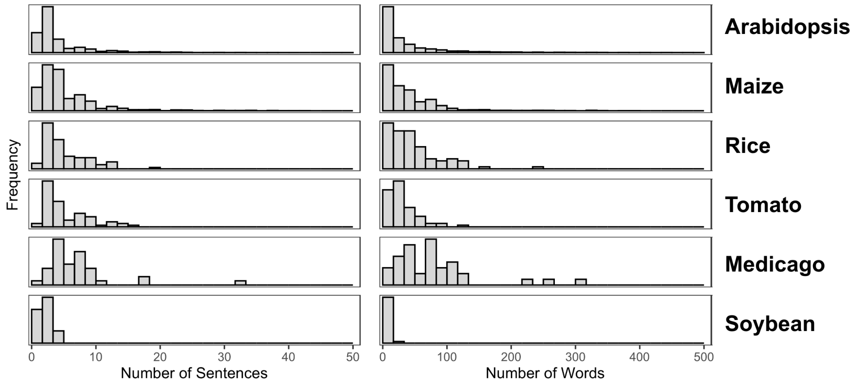


Figure 6.1 Phenotype description text length distributions across six plant species. The distributions for quantities of text in terms of both sentences (Left) and words (Right) describing phenotypes for genes in each of the plant species. Outliers with very long descriptions are not shown, which includes <1% of the genes belonging to Arabidopsis and <0.1% of the genes belonging to maize. The y-axis is scaled to be proportional to the quantity of genes for each individual species.

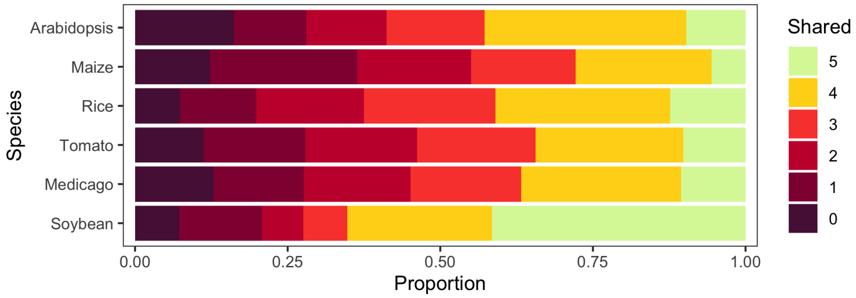


Figure 6.2 Overlap among vocabularies used to describe phenotypes in each species. For each of the six species in the dataset, listed on the left, the proportion of words in the total vocabulary used in all phenotype descriptions of that species that are shared with the vocabularies of a given additional number of species are shown, with colors indicated on the right. For example, plum/purple indicates the proportion of words used only in that species, and light green indicates the proportion of that species vocabulary that is shared with the vocabulary of all five other species.

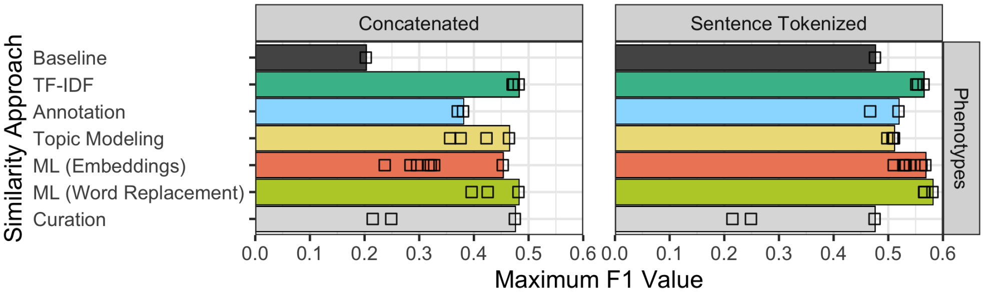


Figure 6.3 Comparing the groups of gene pair similarity approaches. The maximum F_1 statistics for each approach in each broad category for measuring gene similarity is shown, with the bar indicating the best F_1 statistics among all the approaches in that general group. Bars on the left indicate performance when phenotype descriptions are treated as one concatenated piece of text, and bars on the right indicate performance when the descriptions are sentence tokenized first.

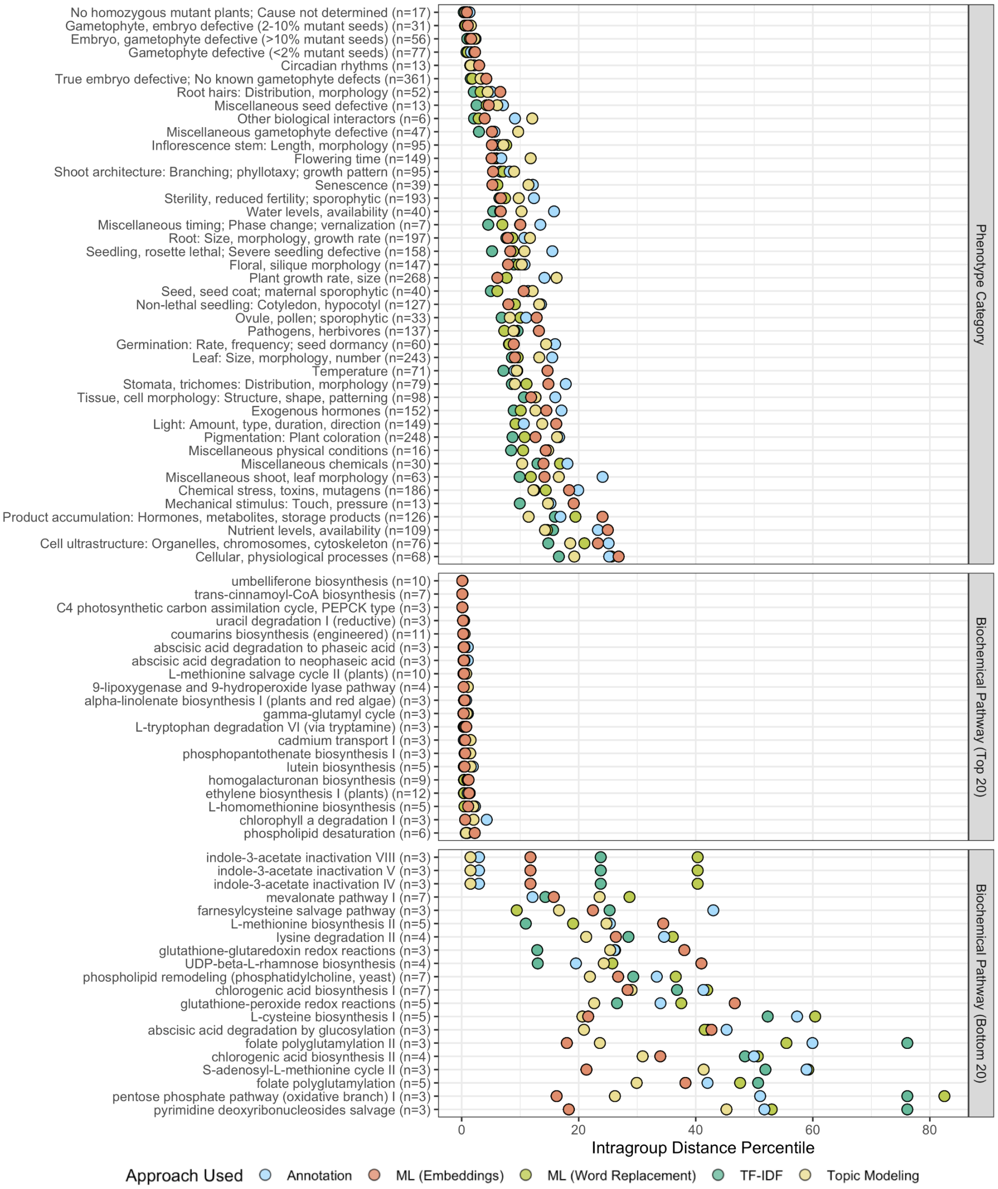
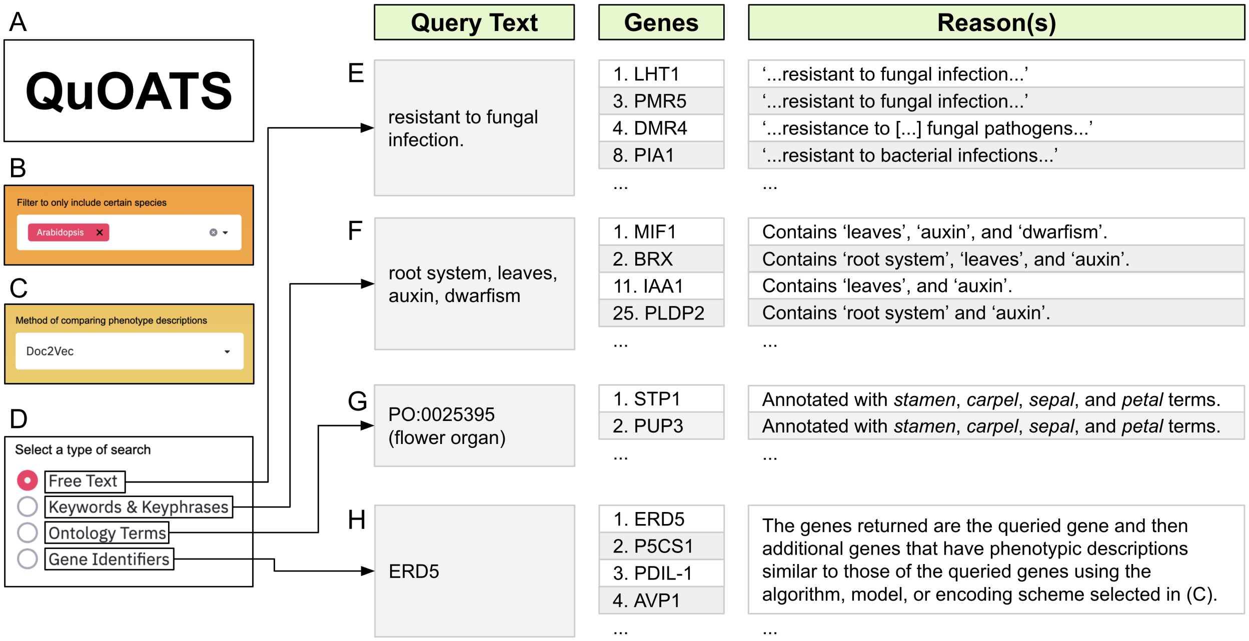


Figure 6.4 Cohesiveness of phenotype and pathway gene groups. Phenotype categories (Top) and Biochemical Pathways (Middle and Bottom) are listed, with the number of genes in this dataset belonging to each group listed to the right of the group's name. The x-axis indicates group cohesiveness, given as the percentile against all pairwise gene distances that the average distance between any two genes in that group falls in. The minimum value of this metric achieved by any approach that is in the listed category is shown. For example, the location of the yellow dot in a particular row indicates the smallest intragroup distance percentile obtained by any approach in the topic modeling category of text-based approaches for that particular group of genes.



Query Text	Genes	Reason(s)
resistant to fungal infection.	1. LHT1	'...resistant to fungal infection...'
	3. PMR5	'...resistant to fungal infection...'
	4. DMR4	'...resistance to [...] fungal pathogens...'
	8. PIA1	'...resistant to bacterial infections...'
	...	...
root system, leaves, auxin, dwarfism	1. MIF1	Contains 'leaves', 'auxin', and 'dwarfism'.
	2. BRX	Contains 'root system', 'leaves', and 'auxin'.
	11. IAA1	Contains 'leaves', and 'auxin'.
	25. PLDP2	Contains 'root system' and 'auxin'.
	...	...
PO:0025395 (flower organ)	1. STP1	Annotated with *stamen*, *carpel*, *sepal*, and *petal* terms.
	2. PUP3	Annotated with *stamen*, *carpel*, *sepal*, and *petal* terms.
	...	...
ERD5	1. ERD5	The genes returned are the queried gene and then additional genes that have phenotypic descriptions similar to those of the queried genes using the algorithm, model, or encoding scheme selected in (C).
	2. P5CS1	
	3. PDIL-1	
	4. AVP1	
	...	...

Figure 6.5 Querying plant genes, annotations, and phenotype descriptions. A. The name of the web application we have developed. B. Option to subset the available dataset to only include certain species. C. Option to select the algorithm or method used to compare phenotype descriptions. D. Four different types of querying are supported. E, F, G, H. The information given here for each query type is presented when using the webtool, but has been re-organized and truncated for the sake of illustration. The queries listed are the text strings that are entered into the search bar to generate the results shown. The returned genes appear in the results in the row indicated by the number to the left of the gene names. The reasons that these genes appear in this order given these particular queries are described to the right of the gene names.

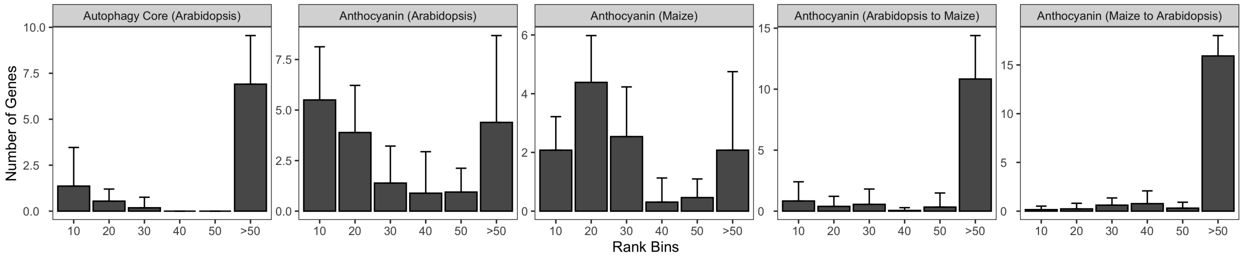


Figure 6.6 Querying with autophagy core genes and anthocyanin biosynthesis genes in QuOATS. The labels above each plot indicate the set of genes, the species of the genes used as the queries, and then the species for which the resulting ranked genes were filtered (in the case of the left three plots the species is the same for queries and targets). Bars represent bins of rank values for returned genes. Their height indicates the average number of genes with those ranks returned in each query. The error bar indicates the standard deviation in each case. Bars in each plot are labeled with the rank that falls in the right-most edge of that bin. For example, the bar labelled 20 represents genes that were ranked between 11 and 20 in the query results.

Table 6.1 Scope and scale of the complete dataset.

		Phenotype Descriptions				Annotations			Other Databases			
Species	Number of Genes	Total Sentences	Total Words	Unique Words	Unique to Species	Mapped to EQ Stmt(s)	Mapped to GO Term(s)	Mapped to PO Term(s)	Mapped in PlantCyc	Mapped in KEGG	Mapped in STRING	Mapped in PANTHER
Arabidopsis	6,274	30,123	261,422	6,792	5,084	2,393	5,984	4,395	843	1,435	4,317	377
Maize	1,405	7,512	47,139	1,722	498	114	190	761	133	157	229	443
Rice	92	478	3,689	760	97	92	46	92	3	0	45	86
Tomato	69	359	1,678	552	99	72	25	64	2	18	11	10
Medicago	37	263	2,447	671	123	40	30	32	2	0	13	0
Soybean	30	62	222	78	12	30	28	27	0	1	5	0
Total	7,907	38,797	316,597	7,663	5,913	2,741	6,303	5,371	983	1,611	4,620	916

Table 6.2 Biological relationships tested in each task.

Task	Description (Are genes A and B. . .)	Knowledge Source
Phenotypes	...impacting the same phenotype?	Lloyd and Meinke, 2012
Pathways	...functioning in the same pathway?	PlantCyc, KEGG
Associations	...known to share a function or process?	STRING
Orthologs	...orthologous to one another?	PANTHER

Table 6.3 Number of genes and gene pairs used for each task.

	All Text Data				With Annotations			
Question	Genes	Pairs	Positive Pairs		Genes	Pairs	Positive Pairs	
Phenotypes	2,356	2,774,190	303,009	10.92%	2,284	2,607,186	293,221	11.25%
Pathways	1,838	1,688,203	45,847	2.72%	1,045	545,490	14,853	2.72%
Associations	4,620	9,343,325	147,271	1.58%	2,377	2,530,556	52,541	2.08%
Orthologs	921	248,913	65	0.03%	368	43,187	23	0.05%

Table 6.4 Similarities among datasets across biological tasks.

Task 1	Task 2	Overlap Size	Jaccard (Pairs)	Jaccard (Values)
Associations	Pathways	1,271,297	0.130	0.172
Phenotypes	Pathways	511,566	0.129	0.050
Phenotypes	Associations	2,687,721	0.285	0.032
Pathways	Orthologs	29,654	0.016	0.012
Phenotypes	Orthologs	0	0.000	
Associations	Orthologs	0	0.000	

Table 6.5 Comparing F_1scores and group significance rates for phenotype and pathway relationships.

			Phenotypes (F_1, % Significant)				Pathways (F_1, % Significant)			
Approach	Category	Concat	All Genes		Curated		All Genes		Curated	
Baseline	Baseline	Yes	0.197	17%	0.202	19%	0.053	6%	0.053	3%
TF-IDF (Unigrams)	TF-IDF	Yes	0.465	100%	0.473	100%	0.100	59%	0.114	55%
TF-IDF (Unigrams & Bigrams)	TF-IDF	Yes	0.473	100%	0.482	100%	0.104	61%	0.120	57%
TF-IDF (Plant Article Unigrams)	TF-IDF	Yes	0.462	100%	0.471	100%	0.096	59%	0.110	59%
NOBLE Coder (Precise)	Annotation	Yes	0.364	91%	0.370	91%	0.072	33%	0.079	25%
NOBLE Coder (Partial)	Annotation	Yes	0.372	100%	0.380	100%	0.082	41%	0.094	33%
LDA (50 Topics)	Topic Modeling	Yes	0.365	91%	0.376	93%	0.080	11%	0.088	12%
LDA (100 Topics)	Topic Modeling	Yes	0.346	83%	0.356	86%	0.073	9%	0.083	10%
NMF (50 Topics)	Topic Modeling	Yes	0.452	100%	0.464	100%	0.089	39%	0.103	47%
NMF (100 Topics)	Topic Modeling	Yes	0.413	100%	0.423	100%	0.086	46%	0.102	46%
Doc2Vec (Wikipedia)	ML (Embeddings)	Yes	0.313	83%	0.321	83%	0.063	22%	0.068	17%
Doc2Vec (Plants)	ML (Embeddings)	Yes	0.233	45%	0.237	48%	0.059	20%	0.060	16%
Word2Vec (Wikipedia)	ML (Embeddings)	Yes	0.276	98%	0.284	98%	0.079	14%	0.096	20%
Word2Vec (PubMed)	ML (Embeddings)	Yes	0.320	93%	0.327	98%	0.086	15%	0.108	19%
Word2Vec (Plants)	ML (Embeddings)	Yes	0.445	93%	0.453	98%	0.097	29%	0.111	35%
BERT	ML (Embeddings)	Yes	0.289	88%	0.296	88%	0.078	14%	0.096	18%
BioBERT	ML (Embeddings)	Yes	0.310	86%	0.317	86%	0.080	11%	0.098	14%
Word2Vec (Wikipedia)	ML (Word Replacement)	Yes	0.387	98%	0.396	100%	0.093	47%	0.107	47%
Word2Vec (PubMed)	ML (Word Replacement)	Yes	0.417	100%	0.425	100%	0.098	51%	0.111	50%
Word2Vec (Plant Phenotypes)	ML (Word Replacement)	Yes	0.473	98%	0.482	100%	0.101	55%	0.114	56%
Baseline	Baseline	No	0.465	74%	0.476	76%	0.082	18%	0.097	39%
TF-IDF (Unigrams)	TF-IDF	No	0.544	95%	0.554	100%	0.097	36%	0.108	53%
TF-IDF (Unigrams & Bigrams)	TF-IDF	No	0.540	95%	0.551	98%	0.097	39%	0.107	54%
TF-IDF (Plant Article Unigrams)	TF-IDF	No	0.555	98%	0.565	100%	0.094	34%	0.106	49%
NOBLE Coder (Precise)	Annotation	No	0.458	81%	0.467	81%	0.086	8%	0.103	30%
NOBLE Coder (Partial)	Annotation	No	0.509	98%	0.519	100%	0.090	24%	0.102	45%
NMF (50 Topics)	Topic Modeling	No	0.489	86%	0.498	88%	0.087	6%	0.099	34%
NMF (100 Topics)	Topic Modeling	No	0.497	91%	0.508	93%	0.087	14%	0.100	37%
LDA (50 Topics)	Topic Modeling	No	0.499	98%	0.510	100%	0.086	8%	0.099	32%
LDA (100 Topics)	Topic Modeling	No	0.499	98%	0.509	100%	0.092	15%	0.104	38%
Doc2Vec (Wikipedia)	ML (Embeddings)	No	0.519	100%	0.530	100%	0.096	27%	0.107	50%
Doc2Vec (Plants)	ML (Embeddings)	No	0.558	93%	0.568	95%	0.095	27%	0.104	50%
Word2Vec (Wikipedia)	ML (Embeddings)	No	0.521	98%	0.532	98%	0.094	15%	0.106	38%
Word2Vec (PubMed)	ML (Embeddings)	No	0.529	100%	0.540	100%	0.095	19%	0.108	44%
Word2Vec (Plants)	ML (Embeddings)	No	0.550	98%	0.561	98%	0.099	32%	0.111	53%
BERT	ML (Embeddings)	No	0.499	100%	0.510	100%	0.096	21%	0.111	43%
BioBERT	ML (Embeddings)	No	0.517	100%	0.527	100%	0.099	23%	0.113	46%
Word2Vec (Wikipedia)	ML (Word Replacement)	No	0.556	98%	0.566	100%	0.099	32%	0.110	53%
Word2Vec (PubMed)	ML (Word Replacement)	No	0.554	100%	0.566	100%	0.098	32%	0.109	51%
Word2Vec (Plant Phenotypes)	ML (Word Replacement)	No	0.570	95%	0.581	98%	0.098	33%	0.107	49%
GO	Curation				0.249	64%			0.140	41%
PO	Curation				0.215	17%			0.056	9%
EQs	Curation				0.475	76%			0.093	50%

Table 6.6 Comparing F_1 scores for associations and orthologous gene pair relationships.

			Associations (F_1)		Orthologs (F_1)	
Approach	Category	Concat	All Genes	Curated	All Genes	Curated
Baseline	Baseline	Yes	0.031	0.041	0.001	0.001
TF-IDF (Unigrams)	TF-IDF	Yes	0.049	0.068	0.010	0.061
TF-IDF (Unigrams & Bigrams)	TF-IDF	Yes	0.051	0.072	0.016	0.054
TF-IDF (Plant Article Unigrams)	TF-IDF	Yes	0.048	0.067	0.008	0.057
NOBLE Coder (Precise)	Annotation	Yes	0.037	0.049	0.012	0.022
NOBLE Coder (Partial)	Annotation	Yes	0.042	0.060	0.003	0.016
LDA (50 Topics)	Topic Modeling	Yes	0.039	0.053	0.002	0.005
LDA (100 Topics)	Topic Modeling	Yes	0.038	0.053	0.005	0.004
NMF (50 Topics)	Topic Modeling	Yes	0.042	0.060	0.007	0.013
NMF (100 Topics)	Topic Modeling	Yes	0.043	0.061	0.006	0.020
Doc2Vec (Wikipedia)	ML (Embeddings)	Yes	0.033	0.047	0.015	0.029
Doc2Vec (Plants)	ML (Embeddings)	Yes	0.031	0.041	0.001	0.007
Word2Vec (Wikipedia)	ML (Embeddings)	Yes	0.042	0.059	0.003	0.003
Word2Vec (PubMed)	ML (Embeddings)	Yes	0.042	0.060	0.006	0.007
Word2Vec (Plants)	ML (Embeddings)	Yes	0.052	0.070	0.012	0.065
BERT	ML (Embeddings)	Yes	0.045	0.059	0.003	0.003
BioBERT	ML (Embeddings)	Yes	0.046	0.062	0.009	0.020
Word2Vec (Wikipedia)	ML (Word Replacement)	Yes	0.046	0.065	0.006	0.029
Word2Vec (PubMed)	ML (Word Replacement)	Yes	0.048	0.066	0.023	0.133
Word2Vec (Plant Phenotypes)	ML (Word Replacement)	Yes	0.048	0.069	0.018	0.080
Baseline	Baseline	No	0.044	0.067	0.004	0.008
TF-IDF (Unigrams)	TF-IDF	No	0.051	0.072	0.006	0.008
TF-IDF (Unigrams & Bigrams)	TF-IDF	No	0.051	0.072	0.005	0.007
TF-IDF (Plant Article Unigrams)	TF-IDF	No	0.051	0.073	0.004	0.007
NOBLE Coder (Precise)	Annotation	No	0.051	0.068	0.004	0.003
NOBLE Coder (Partial)	Annotation	No	0.048	0.069	0.004	0.005
NMF (50 Topics)	Topic Modeling	No	0.049	0.070	0.004	0.007
NMF (100 Topics)	Topic Modeling	No	0.049	0.071	0.005	0.006
LDA (50 Topics)	Topic Modeling	No	0.047	0.069	0.005	0.006
LDA (100 Topics)	Topic Modeling	No	0.048	0.069	0.006	0.008
Doc2Vec (Wikipedia)	ML (Embeddings)	No	0.051	0.071	0.007	0.008
Doc2Vec (Plants)	ML (Embeddings)	No	0.051	0.071	0.006	0.010
Word2Vec (Wikipedia)	ML (Embeddings)	No	0.049	0.070	0.006	0.005
Word2Vec (PubMed)	ML (Embeddings)	No	0.048	0.071	0.007	0.009
Word2Vec (Plants)	ML (Embeddings)	No	0.053	0.074	0.006	0.008
BERT	ML (Embeddings)	No	0.048	0.070	0.005	0.008
BioBERT	ML (Embeddings)	No	0.048	0.071	0.005	0.008
Word2Vec (Wikipedia)	ML (Word Replacement)	No	0.052	0.073	0.005	0.007
Word2Vec (PubMed)	ML (Word Replacement)	No	0.052	0.073	0.005	0.006
Word2Vec (Plant Phenotypes)	ML (Word Replacement)	No	0.050	0.072	0.006	0.007
GO	Curation			0.094		0.059
PO	Curation			0.048		0.001
EQs	Curation			0.063		0.014

Table 6.7 Comparing F_1 scores for pathways for intraspecies and interspecies gene pairs.

			Pathways, All Genes (F_1)		Pathways, Curated (F_1)	
Approach	Category	Concat	Intraspecies	Interspecies	Intraspecies	Interspecies
Baseline	Baseline	Yes	0.053	0.051	0.054	0.049
TF-IDF (Unigrams)	TF-IDF	Yes	0.107	0.067	0.116	0.094
TF-IDF (Unigrams & Bigrams)	TF-IDF	Yes	0.111	0.069	0.123	0.097
TF-IDF (Plant Article Unigrams)	TF-IDF	Yes	0.102	0.067	0.112	0.092
NOBLE Coder (Precise)	Annotation	Yes	0.078	0.055	0.082	0.073
NOBLE Coder (Partial)	Annotation	Yes	0.088	0.058	0.097	0.072
LDA (50 Topics)	Topic Modeling	Yes	0.084	0.060	0.089	0.076
LDA (100 Topics)	Topic Modeling	Yes	0.078	0.060	0.086	0.065
NMF (50 Topics)	Topic Modeling	Yes	0.092	0.073	0.104	0.097
NMF (100 Topics)	Topic Modeling	Yes	0.091	0.068	0.104	0.081
Doc2Vec (Wikipedia)	ML (Embeddings)	Yes	0.069	0.051	0.070	0.062
Doc2Vec (Plants)	ML (Embeddings)	Yes	0.060	0.055	0.062	0.049
Word2Vec (Wikipedia)	ML (Embeddings)	Yes	0.089	0.053	0.102	0.067
Word2Vec (PubMed)	ML (Embeddings)	Yes	0.095	0.056	0.114	0.074
Word2Vec (Plants)	ML (Embeddings)	Yes	0.105	0.071	0.115	0.107
BERT	ML (Embeddings)	Yes	0.087	0.052	0.102	0.059
BioBERT	ML (Embeddings)	Yes	0.089	0.051	0.104	0.060
Word2Vec (Wikipedia)	ML (Word Replacement)	Yes	0.100	0.063	0.110	0.088
Word2Vec (PubMed)	ML (Word Replacement)	Yes	0.105	0.062	0.114	0.088
Word2Vec (Plant Phenotypes)	ML (Word Replacement)	Yes	0.106	0.071	0.115	0.108
Baseline	Baseline	No	0.091	0.051	0.101	0.049
TF-IDF (Unigrams)	TF-IDF	No	0.102	0.067	0.109	0.099
TF-IDF (Unigrams & Bigrams)	TF-IDF	No	0.102	0.069	0.109	0.093
TF-IDF (Plant Article Unigrams)	TF-IDF	No	0.100	0.066	0.107	0.098
NOBLE Coder (Precise)	Annotation	No	0.093	0.057	0.106	0.081
NOBLE Coder (Partial)	Annotation	No	0.096	0.058	0.105	0.069
NMF (50 Topics)	Topic Modeling	No	0.094	0.054	0.102	0.070
NMF (100 Topics)	Topic Modeling	No	0.093	0.058	0.103	0.070
LDA (50 Topics)	Topic Modeling	No	0.091	0.056	0.102	0.069
LDA (100 Topics)	Topic Modeling	No	0.098	0.070	0.107	0.077
Doc2Vec (Wikipedia)	ML (Embeddings)	No	0.103	0.056	0.110	0.070
Doc2Vec (Plants)	ML (Embeddings)	No	0.101	0.063	0.106	0.077
Word2Vec (Wikipedia)	ML (Embeddings)	No	0.100	0.055	0.108	0.069
Word2Vec (PubMed)	ML (Embeddings)	No	0.103	0.060	0.112	0.082
Word2Vec (Plants)	ML (Embeddings)	No	0.106	0.072	0.113	0.104
BERT	ML (Embeddings)	No	0.104	0.057	0.115	0.069
BioBERT	ML (Embeddings)	No	0.106	0.057	0.116	0.079
Word2Vec (Wikipedia)	ML (Word Replacement)	No	0.104	0.070	0.112	0.102
Word2Vec (PubMed)	ML (Word Replacement)	No	0.104	0.064	0.111	0.090
Word2Vec (Plant Phenotypes)	ML (Word Replacement)	No	0.103	0.073	0.108	0.108
GO	Curation				0.137	0.191
PO	Curation				0.057	0.107
EQs	Curation				0.097	0.049

CHAPTER 7. GENERAL CONCLUSION

7.1 Summary of Findings

The preceding chapters have presented a discussion on how computational methods can be used to both represent and compare phenotypes in a scalable manner, using both biological ontologies and natural language processing approaches that account for semantics in free text descriptions of phenotypes. Chapter 3 and Chapter 4 presents a computational pipeline for producing EQ statement annotations given input text descriptions of phenotypes, and discuss how while this set of predicted annotations is comparable to the curated dataset of annotations in producing similarity values that reflect biology, it does not outperform even simple natural language processing approaches for representing and comparing the phenotype descriptions. While this analysis relied on specific relationships such as similarities between anthocyanin biosynthesis genes to evaluate the performance of each computational or curation approach, Chapter 6 presents an analyses of how these results generalize both across relevant biological relationships like orthology, protein associations, biochemical pathways, and phenotypes, but also for a variety of natural language processing approaches for representing and comparing text, ranging from simple bag-of-words approaches to leveraging transformer models to account for context-specific semantics. These results determined that orthology and protein associations are not generally represented through phenotype description similarity in the existing dataset in plants, while phenotype categorizations are, even when relying solely on computation and not curation. While phenotype description similarity is not predictive of whether two genes are involved in a shared pathway in the general case, a percentage of pathways represented in the dataset do contain genes with similarities that are accounted for purely through computation and without curation. This analysis also demonstrates how the predictive ability of text-based similarities are improved by accounting for semantic relationships between words, with both general and domain-specific word embedding models. These results informed the creation of a webtool enabling researchers to query with genes, terms, keywords, and free text descriptions to obtain groups of plant genes with related phenotypes, and that chapter presents examples and illustrations of its utility.

7.2 Future Work

The work described here was primarily focused on characterizing the utility of the existing dataset of phenotype descriptions in plants, (i.e., What biological relationships can we recover with it? What techniques are most effective in representing it computationally?). As a result, the future of this area of research will largely depend on how this dataset is expanded. Because of the utility we have demonstrated for computing on and querying with text descriptions in comparison to ontology term annotations, we advocate for community databases to consider including this datatype as an additional field when generating annotations, or for including unprocessed text from papers as a short-term solution while high-quality datasets of annotations are curated for this data. In addition, extracting phenotype descriptions from academic papers is an active area of research that has the potential to greatly expand this dataset in plants (e.g., Collier et al. (2015); Xing et al. (2018)). As the dataset is grown through any one of these developments, the types of computational approaches that are most applicable for representing and comparing these descriptions will likely change from what is described here, and tools that enable researchers to work with this data will of course require corresponding improvements. Semantic similarity between free text phenotype descriptions in general also has applications outside of the type of mutant phenotype datasets discussed here. As described in Chapter 5, this research forms one of the foundations of a pipeline in progress for using voice recordings from researchers in the field to extract plant traits that can be defined as semantic clusters and related to genotypes as an input to genome-wide association studies (discussed in Yanarella et al. (2020)). This is the work of Colleen Yanarella (Lawrence-Dill Lab, Iowa State University), and is a promising direction for the field of comparing phenotype descriptions computationally in plants.

7.3 References

Collier, N., Groza, T., Smedley, D., Robinson, P. N., Oellrich, A., and Rebholz-Schuhmann, D. (2015). Phenominer: from text to a database of phenotypes associated with omim diseases. *Database*, 2015.

Xing, W., Qi, J., Yuan, X., Li, L., Zhang, X., Fu, Y., Xiong, S., Hu, L., and Peng, J. (2018). A gene–phenotype relationship extraction pipeline from the biomedical literature using a representation learning approach. *Bioinformatics*, 34(13):i386–i394.

Yanarella, C. F., Braun, I. R., Lawrence-Dill, C. J., et al. (2020). Computing on phenotypic descriptions for candidate gene discovery and crop improvement. *Plant Phenomics*, 2020:1963251.