

Computational Classification of Phenologs Across Biological Diversity

Ian R Braun, Carolyn J Lawrence-Dill
Genetics, Development, and Cell Biology
Iowa State University
Ames, IA, United States
irbraun@iastate.edu

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, Walls et al., 2015). Preliminary results comparing predicted and curated EQ statements are presented. Potential use across datasets to enable automated phenolog discovery are discussed.

Keywords—phenologs; phenotypes; text mining, ontologies

I. 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 [1,2,3]. 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 [4].

Plant PhenomeNET is a phenotype similarity network composed of phenotypes from six different model plant species that demonstrates the utility of this approach [3]. 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.

II. METHODS

A. 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.

B. 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.

C. 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.

D. 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.

III. RESULTS AND DISCUSSION

Random forest classifiers specific to each ontology were evaluated using standard precision and recall curves (Figure 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 1).

For each predicted EQ statement, its similarity to the corresponding curated EQ statement was measured (Figure 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.

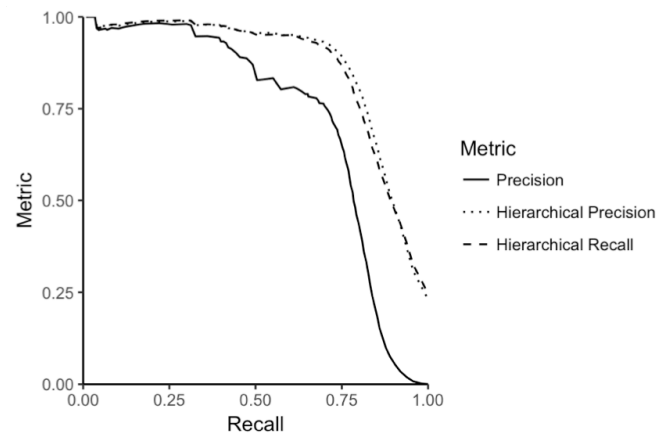


Figure 1. Precision recall curve for predicted PATO terms of holdout atomized-statements from Plant PhenomeNET. Average hierarchical precision and recall are shown between all positive predictions and the closest correct PATO terms.

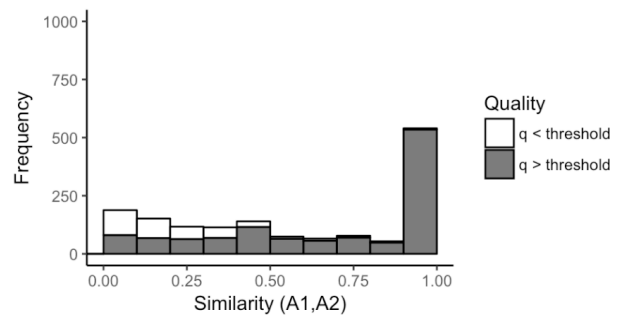


Figure 2. Histogram of similarities (weighted Jaccard) between predicted and curated EQ statements for holdout atomized statements from Plant PhenomeNET. Shaded predictions have quality scores exceeding the learned quality threshold value.

REFERENCES

- [1] McGary KL, Park TJ, Woods JO, Cha HJ, Wallingford JB, Marcotte EM. Systematic discovery of nonobvious human disease models through orthologous phenotypes. *Proc Natl Acad Sci USA*. 2010 Apr 6;107(14):6544-9. doi:10.1073/pnas.0910200107.
- [2] Hoehndorf R, Schofield PN, Gkoutos GV. PhenomeNET: a whole-phenome approach to disease gene discovery. *Nucleic Acids Res*. 2011 Oct;39(18):e119. doi:10.1093/nar/gkr538.
- [3] Oellrich A, Walls RL, Cannon EK, Cannon SB, Cooper L, Gardiner J, Gkoutos GV, Harper L, He M, Hoehndorf R, Jaiswal P, Kalberer SR, Lloyd JP, Meinke D, Menda N, Moore L, Nelson RT, Pujar A, Lawrence CJ, Huala E. An ontology approach to comparative phenomics in plants. *Plant Methods*. 2015 Feb 25;11:10. doi:10.1186/s13007-015-0053-y.
- [4] Braun I, Balhoff J, Berardini TZ, Cooper L, Gkoutos G, Harper L, Huala E, Jaiswal P, Kazic T, Lapp H, Macklin JA, Specht CD, Vision T, Walls RL, Lawrence-Dill CJ. 'Computable' phenotypes enable comparative and predictive phenomics among plant species across domains of life. In: Thessen, AE (Ed.) *Application of Semantic Technologies in Biodiversity Science. Studies on the Semantic Web*, IOS Press/AKA Verlag. To appear.