

Enabling Cross-Species Computation on Phenotype in Plants

Carolyn J. Lawrence

Associate Professor

Department of Genetics, Development and Cell Biology

Iowa State University

triffid@iastate.edu

Plant phenotype datasets can be found in a range of formats including free text and species-specific or knowledge domain-specific controlled vocabularies. While this enables some limited comparison of phenotype data across a single species or within a knowledge domain such as crop breeding, queries or analyses that span a broader set of species are not possible in the absence of a common vocabulary for describing phenotypes. To enable cross-species and cross-domain phenotype comparisons and analyses in plants, we have launched an effort to convert existing phenotype datasets for 8 plant species, encompassing both model species and crops, into a common format using taxonomically broad ontologies representing plant anatomical parts and developmental stages (Plant Ontology), biological processes (Gene Ontology), chemicals (ChEBI), and phenotypic qualities (PATO). Our effort focuses on mutant and overexpression phenotypes associated with genes of known sequence in Arabidopsis, tomato, potato, pepper, maize, rice, soybean and Medicago. Shared use of ontologies, annotation standards, formats and best practices across these eight plant species ensures that the resulting dataset will produce valid results for cross-species querying and semantic similarity analyses. Additionally, the dataset will enable us to explore the relationship between sequence similarity and phenotypic similarity across a range of plant species to make predictions about gene function as well as to predict allele combinations useful for crop improvement.