Phylogenetic- based gene function prediction in the Gene Ontology Consortium

Huaiyu Mi[1*], Pascale Gaudet[2], Marc Feuermann[2], Anushya Muruganujan[1], Suzanna Lewis[3], Paul D. Thomas[1]

[1]University of Southern California, USA

[2]Swiss Institute of Bioinformatics, Switzerland

[3] Lawrence Berkeley Laboratory, USA

*To whom correspondence should be addressed: huaiyumi@usc.edu

Gene Ontology (GO has been using a phylogenetic-based gene function prediction approach to annotate gene function since 2010. A curation tool, called Phylogenetic Annotation and INference Tool (or PAINT), has been developed to help curators to infer annotations among members within a gene family that is defined in PANTHER. Evolutionarily related genes that evolved from a common ancestor (orthologs) tend to preserve their functions. If a gene has been previous annotated with experimental evidence, the curator can make precise assertion as when the function is evolved from based on the phylogenetic information of the family, and propagate the function to that ancestor. All genes evolved from that ancestor would then inherit the same function. PAINT enables a biocurator to construct and record a (generally) parsimonious model of the evolution of function in the family that can be tested against, and modified by, new experimental data as it emerges. As of December 2017, over 5000 families have been curated. A total of 460K genes from 112 organisms have been annotated with new function, and a total of 2.1 million new annotations have been added to the GO. GO experimental curation effort added new experimental annotations to nearly 28K of above genes subsequently. Our analysis shows that two thirds of those annotations were predicted by the phylogenetic annotation, and one third of the phylogenetic annotations were validated by the experimental curation effort. Our results show that PAINT is able to make more accurate inferences, especially to non-model organism genes. It also serves as a QA process to validate the previous annotations by viewing annotations from many related genes. The new curation paradigm greatly improved the efficiency and quality of GO annotation, and will greatly help our user community to utilize the GO knowledge in their data analysis.