Phylogenetic- based gene function prediction in the Gene Ontology Consortium

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

[1]University of Southern California, USA

[2]Swiss Institute of Bioinformatics, Switzerland

[3] Lawrence Berkeley National Laboratory, USA

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

1. INTRODUCTION

Gene Ontology (GO) is a community resource that represents biological knowledge of gene functions through the use of a structured ontology (1). Currently, there are two main approaches to capturing gene functions. The first is manual annotation by trained biocurators based on experimental evidence from published literature. The second is automated algorithmic predictions mainly based on sequence homology. Because most sequences have not been experimentally characterized, most available annotations in GO are provided through these predictive methods and therefore it is crucial to use the most accurate prediction method possible.

Evolutionarily related genes that evolved from a common ancestor (orthologs) tend to preserve their functions. Thus inferences based on such information are more accurate than simple sequence homology. The PANTHER (Protein ANnotation THrough Evolutionary Relationship) classification system (http://www.pantherdb.org/) is a comprehensive system that combines gene function, ontology, pathways and statistical analysis tools to facilitate the analysis of large-scale, genome-wide data from sequencing, proteomics or gene expression experiments (2). The system is built with 104 complete genomes organized into gene families and subfamilies, and their *evolutionary relationships* are captured in phylogenetic trees, multiple sequence alignments and statistical models (hidden Markov models or HMMs).

A curation tool, called Phylogenetic Annotation and INference Tool (or PAINT), has been developed to help biocurators to infer annotations among members within a gene family in context of the PANTHER phylogenetic trees (Figure 1) (3). If an extant protein has been previously annotated based on experimental evidence then, based on the phylogenetic information provided in PANTHER trees, the biocurator can make a precise assertion as to which ancestral protein this function first arose in, and furthermore their assertion can then be automatically propagated to all genes that evolved from that ancestor and thereby have also inherited that same function. PAINT enables a biocurator to construct and record a (generally) parsimonious model of the evolution of function in the family that can then be tested against, and modified by, new experimental data as it emerges. Preliminary studies show that PAINT is able to make more accurate inferences, especially to non-model organism genes. It also serves as a Quality Assurance process to validate previous annotations. By viewing individual annotations in the context of annotations from the related genes in the same family, non-concordant annotations can be revealed, reviewed, and corrected. Currently, over 4000 PANTHER families have been curated, and 1.7 millions new annotations have been added. Nearly 32,000 of those new annotations are to human genes, among which 1867 were later validated by experimental evidence from the literature annotation effort in the GO Consortium (Table 1). The new biocuration paradigm, which combines evolutionary relationship, sequence information and curators' expertise during the annotation process, greatly improved the efficiency and quality of GO annotation.

2. FIGURES

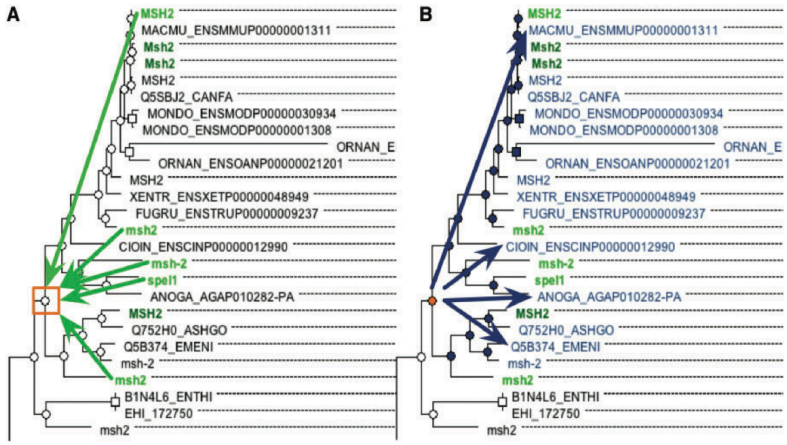

Figure 1. The concept of PAINT. This example shows a MutS homolog family with experimental evidence in GO terms. (A) Primary experimentally based annotations to one term (green labels) are used to infer the most recent common ancestor that the function evolved. The biocurator would annotate the node with the term (orange box). (B). Subsequently, PAINT propagates the annotation to all the descendant leaves of the node (blue labels).

3. TABLES

Table 1. Summary of functional annotations by phylogenetic-based gene function prediction

	Genes	Annotations	Validated annotations
Total	410492	1772005	8636
human	7573	31615	1867
mouse	8863	38896	1255
fruit fly	4536	17963	574
worm	5333	21082	285
YEAST	2154	6373	832
E. coli	758	1687	258

Columns of the table:
Genes – Number of genes that receive at least one annotation.
Annotations – Number of annotations for the genome. An annotation is a gene-GO term combination. For example, if a gene is annotated to two GO terms, it is counted as two annotations.
Validated annotations – Number of predicted annotations that are validated by experimental evidence in literature curation effort in GO.

4. REFERENCES

1. The Gene Ontology Consortium. 2017. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acid Research,* 45 (D1): D331-D338

2. Mi H., Huang X., Muruganujan A., Tang H., Mills C., Kang D., Thomas PD. 2017. PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acid Research,* 45 (D1): D183-D189

3. Gaudet P, Livstone MS, Lewis SE, Thomas PD. 2011. Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium. *Brief Bioinform.* 12(5):449-462