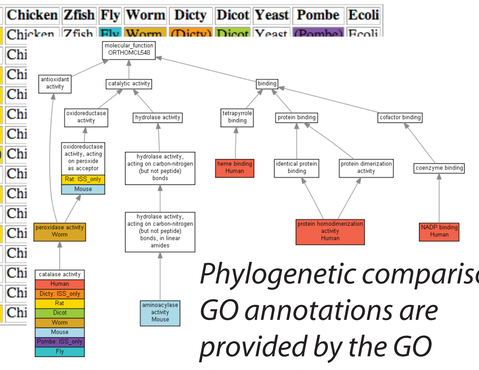
P-POD, the Princeton Protein Orthology Database, as a tool for identifying gene function

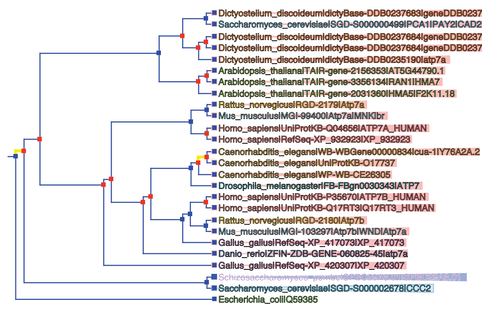
Michael S. Livstone, Rose Oughtred, Sven Heinicke, Fan Kang, Benjamin Vernot, Aiton Goldman, Dannie Durand, David Botstein, Kara Dolinski

P-POD, the Princeton Protein Orthology Database, provides a convenient, centralized resource to help researchers infer protein function. P-POD classifies proteins from the twelve Gene Ontology (GO) Consortium Reference Genome organisms into families of homologs and provides curated information from the literature addressing functional relationships between them. As part of a new effort for the Reference Genome Project, OrthoMCL results from P-POD are being integrated with larger PANTHER protein families to map GO terms from annotated proteins to their unannotated homologs. Several new computational features assist in this effort.

Category	ID	Term	Human	Mouse	Rat	Chicken	Zfish	Fly	Worm	Dicty	Dicot	Yeast	Pombe	Ecoli
Molecular Function	GO:0016209	antioxidant activity	Human	Mouse	Rat	Chicken	Zfish	Fly	Worm	(Dicty)	Dicot	Yeast	(Pombe)	Ecoli
Molecular Function	GO:0005488	binding	Human	Mouse	Rat	Chi								
Molecular Function	GO:0003824	catalytic activity	Human	Mouse	Rat	Chi								
Cellular Component	GO:0005623	cell	Human	Mouse	Rat	Chi								
Cellular Component	GO:0044464	cell part	Human	Mouse	Rat	Chi								
Cellular Component	GO:0043226	organelle	Human	Mouse	Rat	Chi								
Cellular Component	GO:0044422	organelle part	Human	Mouse	(Rat)	Chi								
Biological Process	GO:0065007	biological regulation	Human	Mouse	Rat	Chi								
Biological Process	GO:0009987	cellular process	Human	Mouse	Rat	Chi								
Biological Process	GO:0032502	developmental process	Human	Mouse	Rat	Chi								
Biological Process	GO:0008152	metabolic process	Human	Mouse	Rat	Chi								
Biological Process	GO:0032501	multicellular organismal process	Human	Mouse	Rat	Chi								
Biological Process	GO:0048518	positive regulation of biological process	Human	Mouse	Rat	Chi								
Biological Process	GO:0050789	regulation of biological process	Human	Mouse	Rat	Chi								
Biological Process	GO:0050896	response to stimulus	Human	Mouse	Rat	Chi								



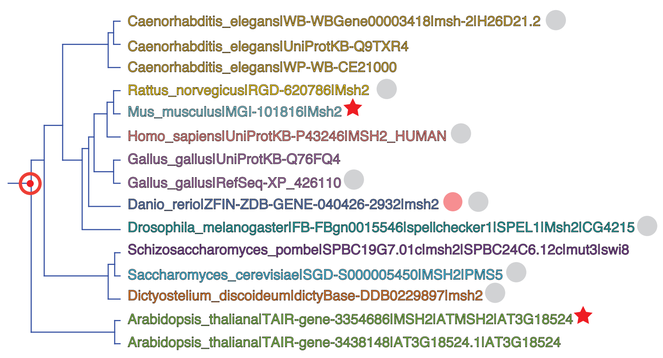
Phylogenetic comparisons of GO annotations are provided by the GO Consortium.



http://www.cs.cmu.edu/~durand/Notung/

The Notung applet (left) allows users to analyze paralogous and orthologous relationships between proteins. Orthologs of yeast *CCC2* are shown in blue, paralogs in pink. The *E. coli* homolog cannot be classified unambiguously due to an edge with weak sequence support, shown in yellow.

ORTHOMCL2648 (MSH2 clade, see example 2):

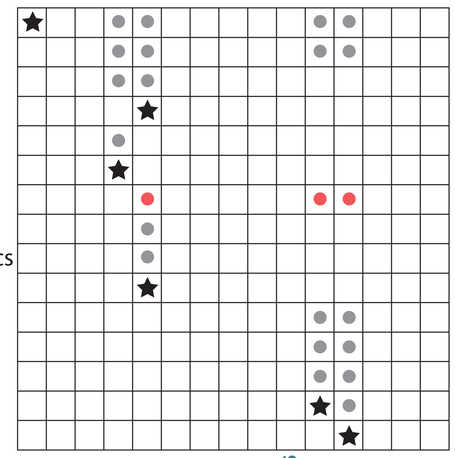


GO:0030983: mismatched DNA binding

★ **Direct annotation** ● **Annotation to child** ● **IEA (direct or to child)**

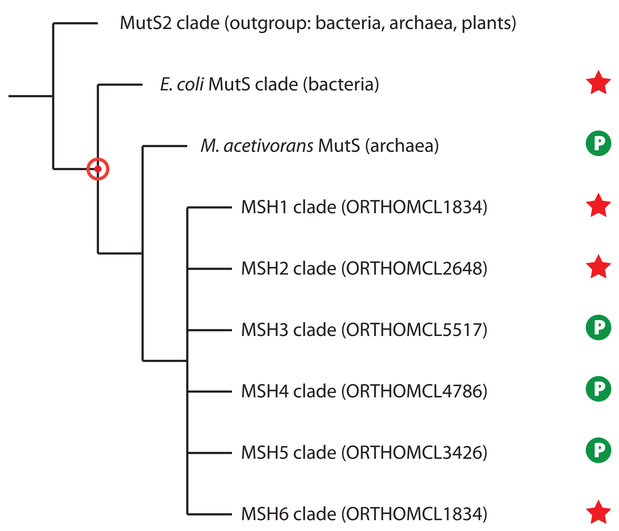
Example 1: Propagation of a GO term from and to multiple members of a homolog family. Three *MSH2* homologs are annotated either directly to the molecular function "mismatched DNA binding" or to one of its children. The common ancestor of these three proteins is the common ancestor of all homologs, so the term can be propagated to all family members. Eight family members already have IEA annotations to this term or a child term, but five have none at all.

- **reproduction**
  - reproductive process
    - gamete generation
      - **germ cell development**
      - male gamete generation
        - **spermatogenesis**
    - reproductive developmental process
      - development of primary sexual characteristics
        - development of primary male sexual characteristics
          - **male gonad development**
      - sex determination
        - mating type determination
          - mating type switching
            - **gene conversion at mating-type locus**
              - **removal of nonhomologous ends**



★ **Direct annotation**
● ● **Annotation to child**

Clades of the MSH2 superfamily, from PANTHER family PTHR11361:



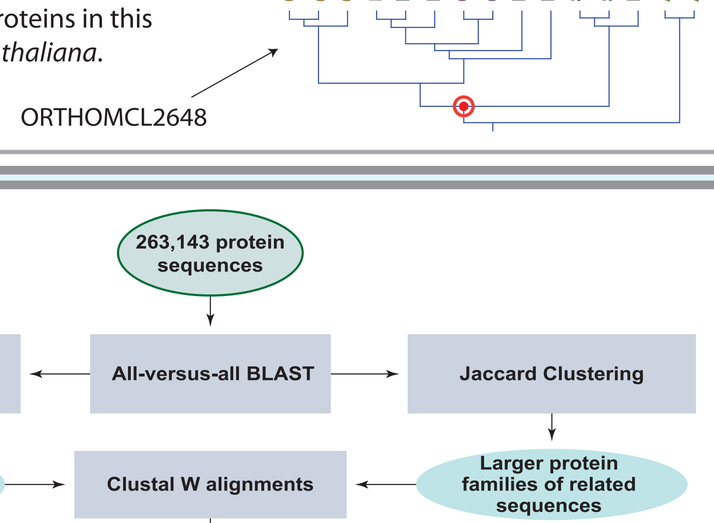
GO:0016887: ATPase activity

★ **Existing annotation** Ⓟ **Propagate annotation**

Transferring GO annotations from homolog to homolog

Example 2: Propagation of a term to families with no members annotated to the term. At least one member each of the MSH1, MSH2, MSH6, and MutS clades is annotated to the molecular function "ATPase activity." By similar reasoning as in example 1, this GO term can be propagated to all members of the MSH3, MSH4, MSH5 and archaeal MutS clades. "ATPase activity" can also be propagated to the remaining members of the MSH1, MSH2, and MSH6 clades.

Example 3: Propagation of a term not explicitly used for annotation. Both *S. cerevisiae* and *S. pombe MSH2* homologs play roles in mating type switching, a child of the biological process "sex determination." Mouse *Msh2* plays a role in male gonad development. All of these terms are children of "reproductive developmental process" (GO:0003006), which can be propagated to all proteins in this clade except those from *A. thaliana*.



ORTHOMCL2648

To address cross-species functional conservation, P-POD provides cross-references to papers describing complementation experiments, disease-related papers from SGD, and OMIM disease information.

Papers describing cross-species expression

Publication	Curator Notes
9889267: Clark AB, *et al.* Functional analysis of human MutSalpha and MutSbeta complexes in yeast. *Nucleic Acids Res.* 1999 Feb 1;*27(3)*:736-42.	The H. sapiens protein ENSP00000233146 does not complement the orthologous mutation in *S. cerevisiae*. Introduction of human *MSH2* (ENSP00000233146), *MSH3* (ENSP00000265081) or *MSH6* (ENSP00000234420), either alone or in combination, does not decrease the high reversion rate of a yeast *msh2* mutant.

Disease-related papers for *MSH2*

11133819: Campbell MR, *et al.* (2000) Candidate mutator genes in mismatch repair-deficient thymic lymphomas: no evidence of mutations in the DNA polymerase delta gene. *Carcinogenesis 21*(12):2281-5

11555625: Ellison AR, *et al.* (2001) Functional analysis of human *MLH1* and *MSH2* missense variants and hybrid human-yeast *MLH1* proteins in *Saccharomyces cerevisiae*. *Hum Mol Genet 10*(18):1889-900

15947132: Clodfelter JE, *et al.* (2005) *MSH2* missense mutations alter cisplatin cytotoxicity and promote cisplatin-induced genome instability. *Nucleic Acids Res 33*(10):3323-30

17720936: Gammie AE, *et al.* (2007) Functional Characterization of Pathogenic Human *MSH2* Missense Mutations in *Saccharomyces cerevisiae*. *Genetics 177*(2):707-21

OMIM Phenotype Information for *MSH2* (P43246)

#120435 LYNCH SYNDROME I
#158320 MUIR-TORRE SYNDROME; MTS
#608089 ENDOMETRIAL CANCER

P-POD uses a modular analysis pipeline and the Generic Model Organisms Database schema (www.gmod.org). Our implementation can accommodate multiple analyses simultaneously, and users can choose from multiple sets of results.

All the data in P-POD are freely and publicly available through the web and by downloading the entire database system via the URL:

http://ortholog.princeton.edu/

For more information on the Reference Genome Project, be sure to attend Pascale Gaudet's talk on Saturday.

References:
(1) Heinicke S, Livstone MS, Lu C, Oughtred R, Kang F, Angiuoli SV, White O, Botstein D, and Dolinski K, "The Princeton Protein Orthology Database (P-POD): a comparative genomics analysis tool for biologists," *PLoS ONE 22*;2(1):e766 (2007).
(2) Durand D, Halldorsson BV, Vernot B, "A hybrid micro-macroevolutionary approach to gene tree reconstruction," *J Comput Biol 13*(2):320-35 (2005).
(3) Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, and Narechania A, "PANTHER: a library of protein families and subfamilies indexed by function," *Genome Res 13*: 2129-2141(2003).
(4) Mi H, Guo N, Kejariwal A, and Thomas PD, "PANTHER version 6: protein sequence and function evolution data with expanded representation of biological pathways," *Nucl Acids Res 35*: D247-D252 (2007).

Nature Precedings : doi:10.1038/npre.2009.3119.1 : Posted 21 Apr 2009