Published in final edited form as:

Curr Protoc Bioinformatics. 2011 March ; 0 6: Unit–6.11. doi:10.1002/0471250953.bi0611s33.

Inferring protein function from homology using the Princeton Protein Orthology Database (P-POD)

Michael S. Livstone¹, Rose Oughtred¹, Sven Heinicke¹, Benjamin Vernot^{2,3}, Curtis Huttenhower⁴, Dannie Durand^{2,5}, and Kara Dolinski¹

¹ Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ

² Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, PA

⁴ Department of Biostatistics, Harvard School of Public Health, Boston, MA

⁵ Department of Computer Science, Carnegie Mellon University, Pittsburgh, PA

Abstract

Inferring a protein's function by homology is a powerful tool for biologists. The Princeton Protein Orthology Database (P-POD) offers a simple way to visualize and analyze the relationships between homologous proteins in order to infer function. P-POD contains computationally-generated analysis distinguishing orthologs from paralogs combined with curated published information on functional complementation and on human diseases. P-POD also features an applet, Notung, for users to explore and modify phylogenetic trees and generate their own

Author contact information: Kara Dolinski, kara@genomics.princeton.edu, phone: 609/258-8306, fax: 609/258-8020. Michael S. Livstone, livstone@genomics.princeton.edu, phone: 609/258-9382, fax: 609/258-8020.

³Currently at the Department of Genome Sciences, University of Washington, School of Medicine, Seattle, WA

Internet Resources

<http://ppod.princeton.edu/>

The main P-POD page and search interface.

<http://ppod.princeton.edu/help/>

The P-POD help page contains an overview of the P-POD pipeline, a brief tutorial, and links to additional information.

http://ppod.princeton.edu/help/help_identifiers.html

Valid identifiers for P-POD and sample searches.

http://ppod.princeton.edu/help/help_tech.html

P-POD technical information, including version numbers and setting for all software in the P-POD pipeline.

http://ppod.princeton.edu/help/help_notung_ortho_para.html

A more extensive and illustrated explanation of how Notung infers orthologs and paralogs in P-POD.

<ftp://gen-ftp.princeton.edu/ppod/>

The P-POD ftp site containing all families, support files, and the 48-species PANTHER 7.0 data set. The current release is in the "version4" folder. More detail is available in README's.

http://ppod.princeton.edu/help/help_data_archive.html

Archival technical information for the original 2007 P-POD release only.

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

The Notung application and documentation.

<http://www.ncbi.nlm.nih.gov/omim>

Online Mendelian Inheritance in Man (OMIM).

<http://www.pantherdb.org/>

The PANTHER 7.0 database.

<http://www.yeastgenome.org/>

The *Saccharomyces* Genome Database.

<http://www.geneontology.org/GO.refgenome.shtml>

Homepage of the Gene Ontology Consortium's Reference Genome project.

<http://amigo.geneontology.org/cgi-bin/amigo/go.cgi>

The Gene Ontology Consortium's AmiGO database.

<http://evolution.genetics.washington.edu/phyliip/newicktree.html>

Description of the Newick tree format.

ortholog/paralogs calls. This unit describes how to search P-POD for precomputed data, how to find and use the associated curated information from the literature, and how to use Notung to analyze and refine the results.

Keywords

functional complementation; disease; conservation; phylogenetic analysis; trees; paralogs; Notung

INTRODUCTION

The Princeton Protein Orthology Database (P-POD) offers a simple way to visualize and analyze the relationships between homologous proteins (Heinicke *et al.* 2007). By combining curated functional information with computationally-generated phylogenetic analysis on twelve of the most important and commonly-used model organisms, P-POD provides an integrated resource to answer the most common introductory questions when studying a new protein: What is it, and what does it do? This paper will introduce the reader to the main functions and uses of P-POD.

- Basic Protocol 1 explains how to search for and interpret off-the-shelf computationally-generated information about a protein and its homologs.
- Basic Protocol 2 explains how to access and use curated information about human diseases and cross-species complementation experiments.
- Basic Protocol 3 explains how to use the Notung applet to refine the phylogenetic analysis of a protein family.

For more information on how the data in P-POD were generated, see (Heinicke *et al.* 2007) or the P-POD help page (URL available in Internet Resources, below).

BASIC PROTOCOL 1

SEARCHING FOR HOMOLOGS

P-POD contains information on the proteomes of the twelve organisms – eleven eukaryotes from human to plants, plus *E. coli* – selected by the Gene Ontology (GO) Consortium as the focus of the Reference Genome project (Reference Genome Group 2009) (see figure 6.11.1). A second, archival dataset contains updated versions of the eight proteomes originally analyzed in the first iteration of P-POD (Heinicke *et al.* 2007). A third dataset, available for download only, analyzes the 48 genomes in the PANTHER 7.0 database (Mi *et al.* 2010).

This protocol explains how to find and interpret the family of proteins that contains your protein of interest.

Necessary Resources

Hardware: Any modern desktop computer with an internet connection.

Software: Any web browser, such as Firefox (v. 3 or greater), Internet Explorer (v. 7 or greater), or Safari (v. 4 or greater)

1. Connect to P-POD, <http://ppod.princeton.edu/> (See figure 6.11.1).
2. Choose either a gene/protein search (steps 3–4) or a disease search (step 5).

By default, P-POD searches the Reference Genome data set. Alternatively, you may choose to search the set of eight organisms – a subset of the

Reference Genome set, plus Plasmodium falciparum – used in the original 2007 P-POD release. These data have been updated using the current P-POD pipeline and protein sequences. To select this data set, click the yellow ‘Heinicke et al., 2007: update’ tab.

3. Type a gene name, symbol, identifier, or descriptive keyword in the box next to the “Search” button.

You can use any of a wide variety of search terms, including gene names and aliases, UniProt or other database identifiers, or descriptions (Table 6.11.1). Searches are case-insensitive, and wild cards (“?” for a single character, “” for zero or more characters) are supported. If P-POD does not find one or more exact matches, it will then perform partial-string matching. For example, a *S. cerevisiae* search for “sod1” returns the gene SOD1; a search for “sod1*” returns SOD1 and CCS1, which is described as a “Copper chaperone for ... Sod1p”; and a search for “sod” returns the SOD1, SOD2, CCS1, and several genes described as “sodium pump”.*

4. Use the checkboxes to select a combination of organisms to search; by default, P-POD limits the search to *S. cerevisiae* proteins. Click the “Search” button to begin searching.
5. Alternatively, if you know the Online Mendelian Inheritance in Man (Hamosh *et al.* 2002) identifier (OMIM id) of a human disease, you also have the option to search for genes associated with that disease. Type the six-digit OMIM id (without the “#” symbol) in the box at the bottom of the page and click the “Lookup” button.
6. Examine the search results (See figure 6.11.2).

A successful search returns a table of results where each row contains information about a protein matching the search parameters. The left half of the table shows information such as the organism, source database and unique identifier, protein name, and description. The most familiar protein name may appear in either the “Protein” or “Synonyms” column, depending on which database provided the sequence. The string of characters that matched the search parameters is highlighted.

The right half of the table shows how many proteins from each organism are in each of the four types of families: OrthoMCL, MultiParanoid, Jaccard, and Naïve Ensemble. OrthoMCL and MultiParanoid produce families of orthologs and close paralogs; Naïve Ensemble clusters are an effort to achieve consensus between the OrthoMCL and MultiParanoid results. Jaccard analysis produces larger families of more distantly-related proteins.

*Each family is represented as a 4×3 (or 4×2) array of circles that use a two-letter code to show the number of proteins from each of the organisms; for example, “At 2” means that the family contains two proteins from *Arabidopsis thaliana*. The colors are chosen in accordance with GO Consortium conventions. If a protein is not assigned to a family in a particular analysis, the word “orphan” appears.*

7. Click on a set of circles or family name to connect to a Protein Family page (See figure 6.11.3).

A phylogenetic tree showing the relationships between the members of the family appears in the top half of the page. As with the circles on the previous page, proteins are colored by species in accordance with GO

Consortium conventions. In the tree, gene duplication events are shown as red squares labeled with the letter “D,” and speciation events are shown as blue squares.

Below the phylogenetic tree is a table listing the members of the family sorted by organism and with identifiers and descriptions of each member. Typically, the table provides three linkouts to information about the protein: from the designated model organism database, from UniProt or NCBI, and from the GO Consortium’s AmiGO database. These sites contain gene locus data, protein sequences, functional information about the protein, links to literature, and much more.

8. (Optional) Above the phylogenetic tree, click on the “Download files” tab to obtain files used to generate the family.

P-POD provides the files used to generate each family. These include: protein sequences, the multiple sequence alignment, the unrooted PHYML (Guindon and Gascuel 2003) phylogenetic tree, and the rooted phylogenetic tree. An additional file, a table showing the relationship (orthologous or paralogous) between each pair of proteins is also provided; see Basic Protocol 3 for additional details on interpreting this file. All files are in plain text format. Protein sequences and alignments are in FASTA format. Phylogenetic trees are in Newick format. The ortholog/paralog table is in CSV (comma-separated values) format.

NOTE: The four gray tabs above the phylogenetic tree display different aspects of information for the Protein Family, but do not link to new pages. Use the tabs to switch between the four types of information, but do not use your browser’s “Back” button.

BASIC PROTOCOL 2

INVESTIGATING THE CONSERVED FUNCTION AND SIGNIFICANCE OF A PROTEIN

P-POD also displays several types of information curated from the literature about the function of a gene and how it is conserved across species. If, for example, an experiment shows that a human gene can complement its yeast ortholog, then we may infer first that the function of the protein is conserved from yeast to human, and second, that yeast may provide a viable model system for studying the protein. It is also interesting to note if the human gene is known to be involved in a disease process. P-POD integrates these types of information on the Protein Family page.

1. On the Protein Family page (see Basic Protocol 1, step 7), click the “Functional Conservation” tab to display a table showing papers that describe the results of complementation experiments involving the yeast genes in the family (See figure 6.11.4).

P-POD contains both complementation experiments, in which an exogenous gene from one organism was expressed in a mutant of another organism to determine whether the exogenous gene can substitute for the mutated gene, and simple exogenous expression experiments where functional substitution was not tested. In the table, the “Curator notes” column describes in free text whether it was a complementation experiment and tells whether the exogenous gene substituted for the mutant. For technical and historical reasons, all the experiments involve one *S. cerevisiae* gene/protein; in most cases, the yeast gene is mutated,

but some experiments were performed with a yeast protein in another organism.

Each row in the table shows one experiment. A single paper may contain multiple experiments and consequently appear in several rows. The total number of experiments from all papers is shown in parentheses in the “Functional Conservation” tab.

The table also shows identifiers for the yeast gene/protein used in the experiment and the citation for the paper in which the experiment appeared. Click on the PubMed ID or icon for a paper to connect to its PubMed citation.

2. (Optional) Download the complete list of curated complementation experiments from the P-POD ftp site (see Internet Resources).
3. Click the “Disease References” tab to display information linking these genes to human diseases (See figure 6.11.5).

This tab shows information from two sources. First, when they are available, all human proteins in the family associated with disease descriptions at OMIM are linked via the six-digit OMIM id for that disease. (Note that the OMIM id can also be copied and pasted into the OMIM search box on the Search Page, See Basic Protocol 1.)

Second, the *Saccharomyces* Genome Database (SGD) identifies and curates papers containing information on yeast genes with human homologs involved in disease. P-POD provides links to these “Disease Gene Related” papers for all yeast proteins in the family. Click on the PubMed ID or icon for a paper to connect to its PubMed citation.

BASIC PROTOCOL 3

USING THE NOTUNG APPLLET TO EXAMINE HOMOLOGY RELATIONSHIPS IN GREATER DETAIL

The P-POD pipeline uses Notung (Durand *et al.* 2006) to fit protein phylogenetic trees to a known species tree, allowing identification of orthologs and paralogs. Notung is also used to identify the best root by minimizing the weighted cost of gene duplications and losses, known as the D/L score.

You can use the Notung applet to explore alternate trees and evaluate ortholog and paralog inferences as a function of the D/L score. Notung will enable you to choose a new root for the tree and/or rearrange the tree around edges with weak sequence support. (In this context, the term “edges” refers to the lines connecting parts of the tree and can be thought of as branches; in contrast, the “leaves” of the tree, the proteins in the family, are a type of “node.”) To ensure that all inferences are strongly supported by the data, Notung does not make a determination of orthology or paralogy for relationships that change after rearranging a weak edge.

Steps 1–3 describe how to get basic information from Notung, and steps 4–7 describe how to use Notung to reconfigure trees and obtain new orthology and paralogy assignments.

Additional Necessary Resources (see Basic Protocol 1)

Software: Java 1.4 or greater for the Notung applet

1. On the Protein Family page (see Basic Protocol 1, step 7), click the “Protein Family” tab and then the “Notung tree analysis” link.

Two windows will launch, one with the applet itself and one with information on Notung and instructions on how to close the applet. You may minimize or hide this second window, but do not close it until you are ready to quit Notung.

The larger Java applet window (figure 6.11.6) contains the Notung applet itself. You may resize or hide this window as you would any window on your desktop, but it will only close when you quit Notung. This window has its own set of menus at the top; menu commands in Basic Protocol 3 refer to the Notung menus.

You may wish to adjust your display options using the choices under the “Font” and “Zoom” menus.

2. Examine the protein tree.

The tree shown is the same tree from the Protein Family page. Additional leaf nodes, showing proteins lost during the evolution of this family, are also shown in gray; you may hide these “LOST” nodes by unchecking the “Display Loss Nodes” checkbox in the bottom pane or in the “Display Options” menu.

Edge weights are shown in green. Weakly supported edges, i.e. those with weights below the “Edge Weight Threshold,” are highlighted in yellow.

3. Examine the homology relationships in the family. To see the orthologs and paralogs of your protein of interest, mouse over it. For a complete table of relationships for this family, select a table from the “About This Tree” menu.

Mouse over a protein in the tree to highlight orthologs in blue and paralogs in pink. A legend in the upper left corner of the tree has buttons for more information and to hide the legend. Use the “Orthologs/Paralogs” button in the bottom pane to turn this feature on and off.

Tables in three formats showing ortholog/paralog relationships for all family members are available under the “About This Tree” menu. Since the Notung applet cannot write files to the hard drive, use the button provided to copy the tables as text and save them in a separate application.

(NOTE: The pink triangles may appear to vary in color due to the superposition of gene names in colored text. The effect can be mitigated by increasing the size of the Notung applet window.)

4. Re-root the tree. Start by clicking the “Rooting” tab in the bottom pane. You may then perform the following three actions in any order and as many times as you like:

- a. Click the “Run Rooting Analysis” button; Notung will calculate the rooting score for each edge in the tree.

The rooting score for an edge is the D/L score when a root is placed on that edge in an unrooted tree. In the P-POD pipeline, Notung chose a root with a minimal D/L score (shown in the lower left corner of the window), but you may wish to root the tree differently. After a Rooting Analysis, the edge or edges with the lowest D/L score are highlighted in red, and close alternatives

are highlighted in pink. You can see the D/L scores due to rooting on each edge by selecting the “Display root score” checkbox.

- b. Click on any edge of the tree to place the root on it.
 - c. Adjust the costs of gene duplications and losses, shown in the lower right corner of the window, by clicking the “Edit Values” button. Repeat the Rooting Analysis (step 4a) to update the root scores. When satisfied with the root you have chosen, continue below.
5. Rearrange the tree to minimize the impact of weak edges, *i.e.* edges with weights below the Edge Weight Threshold. First, click the “Rearrange” tab. You may perform the following three actions in any order and as many times as you like:
 - a. Select the “Highlight weak edges” checkbox to see weak edges highlighted in yellow.

It may be necessary to uncheck and re-check the checkbox to highlight the weak edges. It is also possible that there are no weak edges.
 - b. Change the Edge Weight Threshold, shown in the lower right corner of the window, by clicking the “Edit Values” button.
 - c. Click the “Perform Rearrangement” button to rearrange the tree.

Notung will rearrange weak edges to minimize the D/L score; edges with weights above the threshold will not be changed. Unlike Rooting, it is not possible to rearrange the topology of the tree manually.
6. If you wish to make a change in your analysis or correct a mistake, you may return to any earlier step. Click the “History” tab, select any step from the scrolling list, and continue as above.

Notung keeps track of all analysis steps and changes in parameter values. You may click on any previous step to return to that state. If you make any subsequent changes, all steps originally resulting from that step will be replaced with your new analysis.
7. Save the resulting tree file and ortholog/paralog table as separate files. To save the tree file, select an option from File → View Tree in Text Format; to save the table, select an option from the “About This Tree” menu. For both, follow the on-screen directions to copy the text files, then paste the text into a new document in a text editor and save the resulting file.

The Notung applet cannot save files to the hard drive, so it is necessary to copy the tree file and ortholog/paralog table as text into a new application. Note that control-C is used to copy, even on a Macintosh.

If you would like to generate a graphical version of the revised tree, the tree file can be used as input for the full Notung application which can save both image (png) and text files. The Notung application is freely distributed (see URL in Internet Resources).

The ortholog/paralog table can be pasted into a spreadsheet program for easier reading. This is simplest with the tab-delimited format, but may work for the CSV or HTML tables too. The table includes a key, which states that orthologous and paralogous relationships are denoted by “O”

and “P,” respectively. Notung does not make a determination of orthology or paralogy for those relationships which change after rearranging a weakly-supported edge.

The “About This Tree” menu also has options to display statistics on the tree.

8. Quit Notung by selecting File → Exit or by following the instructions in the accessory Notung window.

File → Close Window closes the tree window, not the applet.

GUIDELINES FOR UNDERSTANDING RESULTS

Proteins in P-POD were assigned to families and, when possible, classified as orthologs and paralogs in a reviewed, automated pipeline. We assessed critical parameters for producing the best overall results in each analysis. This included ensuring that well-understood families of proteins sampling a spectrum of characteristics appeared as expected. By necessity, this is a balancing act, and it is not possible to optimize results individually. Therefore, P-POD provides multiple forms of analysis and provides tools so users can make their own assessments of orthology and paralogy.

P-POD has two analyses, OrthoMCL and MultiParanoid, designed to identify families of orthologs and close paralogs. The Naïve Ensemble analysis is an attempt to merge and reconcile the OrthoMCL and MultiParanoid results. The Jaccard analysis is designed to place ortholog-scale results in a broader evolutionary context; this is useful for a sense of perspective, but is not recommended as a primary source of ortholog identification.

Users are encouraged to use the ortholog results in P-POD as-is for many applications, especially large-scale comparisons involving multiple proteins or families. For more focused work, users are further encouraged to use the Notung applet (Basic Protocol 3) to explore the justification for assignments within their family of interest.

If you have any questions about P-POD, please refer to the help pages listed below or use the email link provided on every page to post a question to the P-POD staff.

COMMENTARY

Background Information

Homology inference and P-POD—Inferring a protein’s function by homology is a powerful tool for biologists. However, while it is simple to search for proteins with similar sequences, placing the results in an evolutionary context to distinguish orthologs (i.e., homologs related by descent and usually assumed to share function) from “mere” paralogs (i.e., homologs related by a gene duplication and therefore capable of diverging in function) is less straightforward. While there are a number of resources that identify orthologs, P-POD combines computational ortholog identification with tools to allow users to make or refine their own ortholog and paralogs determinations, plus curated information addressing the functional relationships between homologs. This integrated suite of computational analysis, user tools, and curated, focused data provides a powerful resource for bench scientists to draw inferences about protein function.

To build P-POD, the 201,332 proteins from the twelve Reference Genome organisms were BLASTed (Altschul *et al.* 1990) against each other and divided into families of related proteins at two levels of resolution using three different clustering techniques (See figure 6.11.7 for pipeline). OrthoMCL (Li *et al.* 2003) and MultiParanoid (Alexeyenko *et al.* 2006)

were used to generate families of orthologs and close paralogs; larger families including more distantly related proteins were assembled using a Jaccard clustering algorithm. We also generated “naïve ensemble” clusters in an effort to achieve a consensus between the OrthoMCL and MultiParanoid results. For each family, we used MAFFT (Katoh and Toh 2008) to generate a multiple sequence alignment of its members, PHYML (Guindon and Gascuel 2003) to generate phylogenetic trees, and Notung (Durand *et al.* 2006) to reconcile the protein tree with the species tree. These computationally-generated data are combined with information from the literature and user tools in an easy-to-use web interface.

Notung and P-POD—The P-POD pipeline uses the standalone Notung application to root and rearrange the trees in each family. Notung also classifies homologous relationships between proteins as orthologous or paralogous, providing insight into the function of uncharacterized proteins. Notung parsimoniously fits a protein tree to a species tree by minimizing the penalty due to gene duplications and losses (the D/L score) and allowing flexibility around edges where the sequence data provide ambiguous information (“weak” edges). However, to ensure that all inferences are strongly supported by the data, Notung does not make a determination of orthology or paralogy for relationships that change after rearranging a weak edge.

P-POD also hosts a Notung applet to allow users to explore and alter trees, generating their own revised ortholog/paralogs inferences. The Notung applet uses the Notung GUI with a reduced set of features since it is only necessary for users to be able to re-root the tree and rearrange it around weak edges. One additional limitation of the applet is that, for security reasons, it cannot save files to the hard drive; the applet provides a workaround to allow users to save their work as described above.

Critical Parameters and Troubleshooting

No data found: Failed searches and orphans—If P-POD does not recognize the search string used in the search interface, it will return the user to the search interface with the message “Search failed.” Make sure that you have selected your organism(s) of interest and are using a valid identifier for that organism(s). Alternatively, search for descriptions (“kinase”) rather than identifiers; and consider using a partial word to cast a wider net, *e.g.* “nucleo” for “nucleotide,” “nucleolar,” “ribonucleoprotein,” *etc.*

Proteins that are not assigned to families are not uncommon; for example, approximately 20% of Reference Genome proteins were not assigned to families in the OrthoMCL analysis. Such proteins are referred to as “orphans” in the search results, and no Protein Family page is available.

Partial data found: Small families and extremely large families—One cannot produce meaningful phylogenetic trees or multiple sequence alignments (MSA’s) for families with only two members. Therefore, there are no trees or alignments available for display or download for these families.

For a small number of extremely large families, it was not possible to produce MSA’s and/or phylogenetic trees. In both cases, P-POD is not able to produce a tree image, as indicated on the Protein Family page. P-POD does provide whatever files were produced for these individual families under the Download Files tab.

Acknowledgments

This work was supported by the National Institute of General Medical Sciences Center (NIGMS) for Quantitative Biology/National Institutes of Health grant P50 GM-071508 to PI David Botstein; and National Human Genome

Research Institute (NHGRI) (P41 HG02273 to GO PIs Judith Blake., Michael Ashburner, J. Michael Cherry, Suzanna Lewis, subcontract to Kara Dolinski).

Literature Cited

- Alexeyenko A, Tamas I, Liu G, Sonnhammer ELL. Automatic clustering of orthologs and inparalogs shared by multiple proteomes. *Bioinformatics*. 2006; 22:e9–e15. [PubMed: 16873526]
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990; 215(3):403–10. [PubMed: 2231712]
- Durand D, Halldórsson BV, Vernot B. A hybrid micro-macroevolutionary approach to gene tree reconstruction. *J Comput Biol*. 2006; 13(2):320–35. [PubMed: 16597243]
- Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*. 2003; 52(5):696–704. [PubMed: 14530136]
- Hamosh A, Scott AF, Amberger J, Bocchini C, Valle D, McKusick VA. Online Mendelian Inheritance in Man (OMIM): A knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*. 2002; 30:52–55. [PubMed: 11752252]
- Heinicke S, Livstone MS, Lu C, Oughtred R, Kang F, Angiuoli SV, White O, Botstein D, Dolinski K. The Princeton Protein Orthology Database (P-POD): a comparative genomics analysis tool for biologists. *PLoS One*. 2007; 22(2):1, e766.
- Katoh K, Toh H. Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform*. 2008; 9(4):286–98. [PubMed: 18372315]
- Li L, Stoeckert CJ Jr, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res*. 2003; 13:2178–2189. [PubMed: 12952885]
- Mi H, Dong Q, Muruganujan A, Gaudet P, Lewis S, Thomas PD. PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. *Nucleic Acids Res*. 2010; 38(Database issue):D204–10. [PubMed: 20015972]
- Reference Genome Group of the Gene Ontology Consortium. The Gene Ontology's Reference Genome Project: a unified framework for functional annotation across species. *PLoS Comput Biol*. 2009; 5(7):e1000431. [PubMed: 19578431]

P-POD: Princeton Protein Orthology Database

[Home](#) [Help](#)[Reference Genome Annotation Project](#) [Heinicke et al, 2007: update](#)

Search by gene/protein name or description:

 E.G. 'ABC transporter', 'Q5F433', 'CCC2';☐ Select/deselect all

- ☐ *Arabidopsis thaliana*
- ☐ *Caenorhabditis elegans*
- ☐ *Danio rerio*
- ☐ *Dictyostelium discoideum*
- ☐ *Drosophila melanogaster*
- ☐ *Escherichia coli*
- ☐ *Gallus gallus*
- ☐ *Homo sapiens*
- ☐ *Mus musculus*
- ☐ *Rattus norvegicus*
- ☒ *Saccharomyces cerevisiae*
- ☐ *Schizosaccharomyces pombe*

Or search by OMIM ID:

Send questions, suggestions, and comments to: yfgdb@genomics.princeton.edu

Figure 6.11.1.

The P-POD search interface at <http://ppod.princeton.edu/>.

Figure 6.11.2. Sample P-POD search results. In this example, a search for *S. cerevisiae* proteins matching “2678” returned the SGD database identifier S000002678, the UniProt identifier P26784, and seven other proteins (not shown). Matching strings are highlighted in pink. The right half of the table shows the compositions of the four types of families: OrthoMCL, MultiParanoid, Jaccard, and Naïve Ensemble. Each family is represented as a 4×3 array of circles that use a two-letter code to show the number of proteins from each of the organisms; for example, “At 2” means that the family contains two proteins from *Arabidopsis thaliana*. If a protein is not assigned to a family in a particular analysis, the word “orphan” appears. If the icon for a family does not load properly, the family name – for example, “OrthoMCL884” – still appears in the relevant table entry.

P-POD: Princeton Protein Orthology Database: GO3/OrthoMCL884

[Home](#) [Help](#) [About](#)

This family has 18 members: 2 *Arabidopsis thaliana*, 1 *Caenorhabditis elegans*, 2 *Danio rerio*, 1 *Dictyostelium discoideum*, 1 *Drosophila melanogaster*, 1 *Escherichia coli*, 2 *Gallus gallus*, 2 *Homo sapiens*, 2 *Mus musculus*, 2 *Rattus norvegicus*, 1 *Saccharomyces cerevisiae*, 1 *Schistosoma mansoni*.

[Protein Family](#) [Functional Conservation \(10\)](#) [Download Files \(5\)](#) [Disease References \(23\)](#)

Interactive Java Applets: [Notung Tree Analysis](#) · [Jalview Alignment](#)

Image generated with Notung 2.6, on Dec 17, 2009

Organism	Protein (Synonyms)	18 members.	Description	AmiGO
<i>A. thaliana</i>	NCBI:NP_176333.1 · TAIR:locus:2031361 (F2K11.18 · AT1G63440 · F2K11_18 · HMA5)		HEAVY METAL ATPASE 5	Link
<i>A. thaliana</i>	NCBI:NP_199292.1 · TAIR:locus:2156354 (K23L20.14 · RAN1 · AT5G44790 · K23L20_14 · HMA7)		RESPONSIVE-TO-ANTAGONIST 1	Link
<i>C. elegans</i>	WB:WBGene00000834 (Y76A2A.2 · cua-1) · UniProtKB:O17737			Link
<i>D. rerio</i>	ENSEMBL:ENSDARP00000029666			Link
<i>D. rerio</i>	UniProtKB:Q4F8H5 · ZFIN:ZDB-GENE-060825-45 (atp7a)		ATPase, Cu++ transporting, alpha polypeptide	Link
<i>D. discoideum</i>	UniProtKB:Q54Q77 · dictyBase:DBD_G0284141 (atp7a)		P-type ATPase, copper-transporting ATPase	Link
<i>D. melanogaster</i>	FB:FBgn0030343 (CG1886 · dmATP7 · ATP7 · DmATP7) · UniProtKB:Q9VYT4			Link
<i>E. coli</i>	UniProtKB:Q59385 · ECOLI-G6260-MONOMER (b0484 · ECK0478 · B0484 · YbaR · CopA · ybaR · YbaR Cu++-exporting ATPase · copper transporter · Cu++-exporting ATPase			Link
<i>G. gallus</i>	NCBI:XP_420307 · ENTREZ:422333			
<i>G. gallus</i>	ENTREZ:418879 · NCBI:XP_417073			
<i>H. sapiens</i>	ENSEMBL:ENSG00000165240 · UniProtKB:Q04656 (IP00215614 · IP00215617 · ATP7A_HUMAN · IP00228610 · IP00215619 · MCL · Q04656 · IP00215616 · ATP7A · MNK · IP00215615)		Copper-transporting ATPase 1	Link
<i>H. sapiens</i>	ENSEMBL:ENSG00000121191 · UniProtKB:P35670 (ATP7B_HUMAN · IP00216296 · WC1 · IP00658175 · PWD · IP00515019 · P35670 · ATP7B · WND · IP0020058)		Copper-transporting ATPase 2	Link
<i>M. musculus</i>	UniProtKB:Q64430 · MGI:MGI:99400 (tr · MNK · Atp7a)		Menkes protein · ATPase, Cu++ transporting, alpha polypeptide	Link
<i>M. musculus</i>	UniProtKB:Q64446 · MGI:MGI:103297 (WND · Atp7a · Atp7b)		Wilson protein · ATPase, Cu++ transporting, beta polypeptide	Link
<i>R. norvegicus</i>	UniProtKB:P70705 · RGD:2179 (Atp7a)		ATPase, Cu++ transporting, alpha polypeptide	Link
<i>R. norvegicus</i>	RGD:2180 (Atp7b) · UniProtKB:Q64535 (ATP7B_RAT · Pina · IP100760123 · IP100324094 · IP100230869 · Q64535 · Atp7b · Wnd)		ATPase, Cu++ transporting, beta polypeptide · Copper-transporting ATPase 2	Link
<i>S. cerevisiae</i>	UniProtKB:P38995 · SGD:S000002678 (CCC2 · YDR270W)		Cu(+2)-transporting P-type ATPase, required for export of copper from the cytosol into an extracytosolic compartment	Link
<i>S. pombe</i>	UniProtKB:O59666 · GeneDB:SPBC29A3.01 (SPBC29A3.01 · ccc2)		copper transporting ATPase Ccc2	Link

Send questions, suggestions, and comments to: ygd@genomics.princeton.edu

Figure 6.11.3.

The Protein Family Page contains a phylogenetic tree and a table of the members of the family. Blue text and symbols in the table represent linkouts. The four gray tabs near the top of the page allow the user to switch back and forth to other types of data. The “Notung Tree Analysis” link activates the Notung applet.

Protein Family	Functional Conservation (10)	Download Files (5)	Disease References (23)	
Protein		Publication		Curator Notes
UniProtKB:P38995 · SGD:S000002678		11373292 Voskoboinik I, et al. <i>The regulation of catalytic activity of the menkes copper-translocating P-type ATPase. Role of high affinity copper-binding sites.</i> J Biol Chem. 2001 Jul 27;276(30):28620-7.		The H. sapiens protein Q04656 complements a homologous mutation in S. cerevisiae.
UniProtKB:P38995 · SGD:S000002678		15963506 Portmann R, et al. <i>Purification and functional reconstitution of the human Wilson copper ATPase, ATP7B.</i> FEBS Lett. 2005 Jul 4;579(17):3589-95.		The H. sapiens protein P35670 complements a homologous mutation in S. cerevisiae.
UniProtKB:P38995 · SGD:S000002678		9261163 Hung IH, et al. <i>Biochemical characterization of the Wilson disease protein and functional expression in the yeast Saccharomyces cerevisiae.</i> J Biol Chem. 1997 Aug 22;272(34):21461-6.		The H. sapiens protein P35670 complements a homologous mutation in S. cerevisiae.
UniProtKB:P38995 · SGD:S000002678		9354393 Sambongi Y, et al. <i>Caenorhabditis elegans cDNA for a Menkes/Wilson disease gene homologue and its function in a yeast CCC2 gene deletion mutant.</i> J Biochem. 1997 Jun;121(6):1169-75.		The C. elegans protein Y76A2A.2 complements a homologous mutation in S. cerevisiae.
UniProtKB:P38995 · SGD:S000002678		14962673 Hsi G, et al. <i>Functional assessment of the carboxy-terminus of the Wilson disease copper-transporting ATPase, ATP7B.</i> Genomics. 2004 Mar;83(3):473-81.		The H. sapiens protein P35670 complements a homologous mutation in S. cerevisiae.
UniProtKB:P38995 · SGD:S000002678		9452509 Payne AS, et al. <i>Functional expression of the menkes disease protein reveals common biochemical mechanisms among the copper-transporting P-type ATPases.</i> J Biol Chem. 1998 Feb 6;273(6):3765-70.		The H. sapiens protein Q04656 complements a homologous mutation in S. cerevisiae.
UniProtKB:P38995 · SGD:S000002678		10942420 Forbes JR, et al. <i>Copper-dependent trafficking of Wilson disease mutant ATP7B proteins.</i> Hum Mol Genet. 2000 Aug 12;9(13):1927-35.		The H. sapiens protein P35670 was expressed in S. cerevisiae, but complementation was not directly tested. This paper shows the localization of human variant ATP7B proteins in yeast.
	16976819 Bouhem M, et al. <i>Evidence that translation substitution leads to a neurodegeneration</i>		

Figure 6.11.4.
Click the “Functional Conservation” tab to display a list of curated experiments describing complementation and exogenous expression experiments. The right-hand column contains curator notes describing the experiment and its results. Linkouts in the middle column connect to the PubMed entry for the paper. This list has been truncated due to space considerations.

Protein Family	Functional Conservation (10)	Download Files (5)	Disease References (23)
OMIM (3)			
ENSEMBL:ENSG00000165240 · UniProtKB:Q04656	#304150 CUTIS LAXA, X-LINKED; OCCIPITAL HORN SYNDROME; OHS; EHLERS-DANLOS SYNDROME, OCCIPITAL HORN TYPE, FORMERLY; EDS IX, FORMERLY; EDS9, FORMERLY		
ENSEMBL:ENSG00000123191 · UniProtKB:P35670	#509400 MENKES DISEASE; MK; MNK; MENKEA SYNDROME; KINKY HAIR DISEASE; STEELY HAIR DISEASE; COPPER TRANSPORT DISEASE		
UniProtKB:P38995 · SGD:S000002678	#277900 WILSON DISEASE; WND; WD; HEPATOENTERIC DEGENERATION		
SGD Disease Papers (20)			
			9920665 Borjigin J, et al. (1999) A novel pineal night-specific ATPase encoded by the Wilson disease gene. <i>J Neurosci</i> 19(3):1018-26
			9837819 Forbes JR and Cox DW (1998) Functional characterization of missense mutations in ATP7B: Wilson disease mutation or normal variant? <i>Am J Hum Genet</i> 63(6):1663-74
			9654149 Iida M, et al. (1998) Analysis of functional domains of Wilson disease protein (ATP7B) in <i>Saccharomyces cerevisiae</i> . <i>FEBS Lett</i> 428(3):281-5
			9546033 Dancis A (1998) Genetic analysis of iron uptake in the yeast <i>Saccharomyces cerevisiae</i> . <i>J Pediatr</i> 132(3 Pt 2):S24-9
			9520490 Gaxiola RA, et al. (1998) The yeast CLC chloride channel functions in cation homeostasis. <i>Proc Natl Acad Sci U S A</i> 95(7):4046-50
			9452509 Payne AS and Gillin JD (1998) Functional expression of the menkes disease protein reveals common biochemical mechanisms among the copper-transporting P-type ATPases. <i>J Biol Chem</i> 273(6):3765-70
			9325307 Yuan DS, et al. (1997) Restriction of copper export in <i>Saccharomyces cerevisiae</i> to a late Golgi or post-Golgi compartment in the secretory pathway. <i>J Biol Chem</i> 272(41):25787-93
			9261163 Hung IH, et al. (1997) Biochemical characterization of the Wilson disease protein and functional expression in the yeast <i>Saccharomyces cerevisiae</i> . <i>J Biol Chem</i> 272(34):21461-6
			8628392 Bassett DE Jr, et al. (1996) Yeast genes and human disease. <i>Nature</i> 379(6566):589-90
			7785328 Fu D, et al. (1995) Sequence, mapping and disruption of CCC2, a gene that cross-complements the Ca^{2+} -sensitive phenotype of <i>exg1</i> mutants and encodes a P-type ATPase belonging to the Cu^{2+} -ATPase subfamily. <i>Yeast</i> 11(3):283-92
UniProtKB:P38995 · SGD:S000002678			7708696 Yuan DS, et al. (1995) The Menkes-Wilson disease gene homologue in yeast provides copper to a ceruloplasmin-like oxidase required for iron uptake. <i>Proc Natl Acad Sci U S A</i> 92(7):2632-6
			19745976 Tera T, et al. (2008) Clinical outcome in Menkes disease patients with a nonsense mutation, ATP7B mutation, C937B. <i>Mol Genet</i>

Figure 6.11.5.

Click the “Disease References” tab to display linkouts to OMIM diseases associated with human members of the protein family (top panel) and a list of papers from SGD containing information about yeast genes with human homologs involved in disease. This list has been truncated due to space considerations.

Figure 6.11.6.

The Notung applet window. The top pane shows the protein family phylogenetic tree and legend. The bottom pane has several tabs, each with its own set of buttons and checkboxes, allowing access to functions to modify the tree. Additional functions are accessible through the pulldown menus at the top. The “Edit Values” button in the lower right corner allows the user to change tree parameters.

Figure 6.11.7.

The P-POD pipeline. Protein sequences were assigned to families using several different techniques, and curated information from several sources is displayed with the computational results.

Sample search terms used in P-POD for each of the twelve Reference Genome organisms include a wide variety of gene names, synonyms, descriptions, and database identifiers.

Table 6.11.1

Species	Gene name	Synonym	Description	Database identifier
<i>A. thaliana</i>	AT1G72830	HAP2C	AT1G72830	2032758
<i>C. elegans</i>	cogc-3	Y71F9AM.4	wvp-1	WBGene00022121
<i>D. rerio</i>	fgf8a		fibroblast growth factor 8 a	ZDB-GENE-990415-72
<i>D. discoideum</i>	mrB	Actr10	ribonucleoside-diphosphate reductase	DDB_G0274021
<i>D. melanogaster</i>	dpp	Activin-A-receptor-25D	decapentaplegic	FBgn00000490
<i>E. coli</i>				P0A6C5
<i>G. gallus</i>	PRAF3_CHICK	IP00597483	PRA1 family protein 3	Q5F433
<i>H. sapiens</i>	LAMB4_HUMAN	IP00295437	Laminin subunit	A4D0S4
<i>M. musculus</i>	Abcd1	ALDP	ATP-binding cassette, sub-family D	MGI:1349215
<i>R. norvegicus</i>	Snap25		synaptosomal-associated protein 25	3728
<i>S. cerevisiae</i>	CDC28	YAL003W	Translation elongation factor 1 beta	S000000003
<i>S. pombe</i>	zta1	SPCC285.01c	NADPH quinone oxidoreductase	SPBC28F2.07