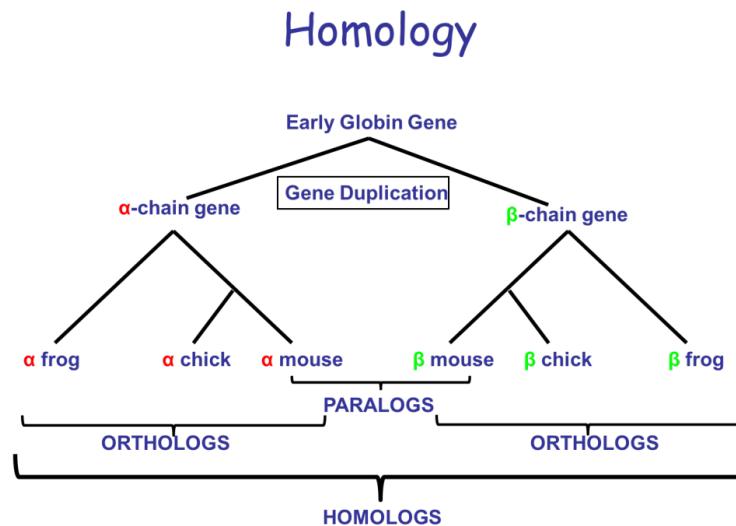# FungiDB & OrthoMCL: Orthology and Phyletic Patterns

Learning objectives:
- Run searches in OrthoMCL.
- Run phyletic pattern searches using check boxes or an expression.
- Combine searches using the strategy system.
- Explore individual ortholog group pages.
- Explore the group cluster graphs.

Homology

About OrthoMCL.

OrthoMCL is a genome-scale algorithm for grouping orthologous protein sequences. It provides not only groups shared by two or more species/genomes, but also groups representing species-specific gene expansion families. Thus, it serves as an important utility for automated eukaryotic genome annotation. OrthoMCL starts with reciprocal best hits within each genome as potential in-paralog/recent paralog pairs and reciprocal best hits across any two genomes as potential ortholog pairs. Related proteins are interlinked in a similarity graph. Then, MCL (Markov Clustering algorithm; Dongen 2000; www.micans.org/mcl) is invoked to split mega-clusters. This process is analogous to the manual review in COG construction. MCL clustering is based on weights between each pair of proteins, so to correct for differences in evolutionary distance the weights are normalized before running MCL.

Background on Orthology and Prediction

Orthologs are homologs separated by speciation events. Paralogs are homologs separated by duplication events. Detection of orthologs is becoming much more important with the rapid progress in genome sequencing (Glover et al. 2019).

OrthoMCL is a genome-scale algorithm for grouping orthologous protein sequences. It provides not only groups shared by two or more species/genomes, but also groups representing species-specific gene expansion families. Thus, it serves as an important utility for automated eukaryotic genome annotation. OrthoMCL starts with reciprocal best hits within each genome as potential in-paralog/recent paralog pairs and reciprocal best hits across any two genomes as potential ortholog pairs. Related proteins are interlinked in a similarity graph. Then, MCL (Markov Clustering algorithm; Dongen 2000; www.micans.org/mcl) is invoked to split mega-clusters. This process is analogous to the manual review in COG construction. MCL clustering is based on weights between each pair of proteins, so to correct for differences in evolutionary distance the weights are normalized before running MCL.

OrthoMCL is similar to the INPARANOID algorithm (Remm et al. 2001) but is extended to cluster orthologs from multiple species. OrthoMCL clusters are coherent with groups identified by EGO (Lee et al. 2002), and an analysis using EC number suggests a high degree of reliability (Li et al. 2003).

We evaluated the performance of seven widely-used orthology detection algorithms that use three general prediction strategies: phylogeny-based, evolutionary distance-based and BLAST-based (Chen, et al. 2007). Specifically, we used Latent Class Analysis (LCA), a statistical technique appropriate for testing large data sets when no gold standard is available. Our results show an overall trade-off between sensitivity and specificity among these algorithms, with INPARANOID and OrthoMCL performing best with False Positive (FP) and False Negative (FN) error rates lower than 20%.

Method for Forming and Expanding Ortholog Groups in OrthoMCL.
Proteins are placed into Ortholog Groups by the following steps:

1. The OrthoMCL algorithm (see below) is employed on proteins from a set of 150 Core species to form Core ortholog groups. These species were carefully chosen based on proteome quality and widespread placement across the tree of life. Each Core protein is placed by the algorithm into a Core ortholog group consisting of one or more proteins. Core group names have the format OG6_xxxxxx (e.g., OG6_101327). OG6 refers to OrthoMCL release 6; for each sub-release (e.g., 6.1, 6.2, etc), the Core species and the Core ortholog group names will remain constant.

2. The proteins from hundreds of additional organisms, termed Peripheral organisms, are mapped into the Core groups. To do this, NCBI BLASTP is used to compare each Peripheral protein to each Core protein in the Core groups. (Note that Peripheral proteins that were previously added to the Core group are NOT used in the BLASTP.) Then, each Peripheral protein is assigned to the Core group containing the Core protein with the best BLAST score, but only if the E-Value is <1e-5 and the percent match length is >=50%.

3. All Peripheral proteins that fail to map to a Core group are collected and subjected to independent OrthoMCL analysis, forming Residual groups consisting of one or more proteins. Residual group names have the format OG6r1_xxxxxx (e.g., OG6r1_101327), where OG6 refers to release 6 and r1 refers to sub-release 1.

4. For each subsequent sub-release (which will occur every ~3 months along with other VEuPathDB sites), proteomes from additional Peripheral organisms will be processed as in steps 2 and 3 above. However, step 3 will differ slightly because the previous set of Residual groups will be disassembled, leaving the previous unmapped Peripheral proteins to be combined with the new unmapped Peripheral proteins. All of these proteins will be used to form new Residual groups (e.g., OG6r2_xxxxxx).

5. During a sub-release, the proteomes of some species will be updated to the latest version. This can be easily done for a Peripheral species: the old set of proteins are removed from ortholog groups and then the new set is mapped into groups as above. However, this is not possible for Core species because these proteins are used to define Core groups. Thus, the Core species with the older proteome remains on the site but is superficially retired by appending its abbreviation with -old (e.g., aaeg becomes aaeg-old). Then, the latest version of the proteome is mapped in as a peripheral species and obtains the original species abbreviation (e.g., aaeg is a peripheral with a more recent proteome than aaeg-old). These retired species will be eliminated fully when a new set of Core species is defined, as described in the next point.

6. On occasion, the set of Core species will be re-defined, as more appropriate proteomes become available and/or when a large number of Core species are retired. In this case, new

Core groups (e.g., OG7_xxxxxx) and Residual groups (e.g., OG7r1_xxxxxx) will be formed from the latest version of proteomes from a carefully-chosen set of core species.

This design allows for the addition of proteomes at every sub-release (e.g., 6.1, 6.2, etc). Note that Core groups (e.g., OG6_101327) will remain between sub-releases, though these groups will expand as Peripheral proteins are mapped in. In contrast, Residual groups will exist only for that sub-release; thus, Residual groups are useful in allowing the user to find proteins related to their

Examining OrthoMCL output on gene record pages in FungiDB

- **Go to the gene record page for the CGB_L0350W, a hypothetical protein CNBL0590.**

    a. What is the function of this gene? How can you infer its function?
        i. Click on the "Orthology and Synteny" link in the Contents menu on the left. Does this gene have orthologs in other *Cryptococcus* species?

    b. Examine evidence in the "Function prediction" section.
    c. What about other organisms <u>outside fungi</u>? (Hint: click on the Ortholog Group OG6_106189).
    d. The OrthoMCL group page is divided into 5 sections:
        1. Phyletic distribution
        2. Group summary
        3. List of proteins
        4. PFam domains
        5. Cluster graph

- **Does this protein have othologs in Archae and Bacteria?**

Phyletic distribution: Numbers refer to the number of proteins in that organism or taxonomic group. In order to see organisms and taxonomic groups without proteins in this ortholog group, uncheck 'Hide zero counts.'

Phlyetic Distribution of Proteins ❓ ⬇ Download

Numbers refer to the number of proteins in that organism or taxonomic group.

expand all | collapse all ☐ Hide zero counts

▼ **Eukaryota (EUKA)**	**266**
▶ **Alveolates (ALVE)**	**3**
▶ **Amoebozoa (AMOE)**	**1**
▶ **Euglenozoa (EUGL)**	**0**
▶ **Fungi (FUNG)**	**130**
▶ **Metazoa (META)**	**125**
▶ **Other Eukaryota (OEUK)**	**3**
▶ **Viridiplantae (VIRI)**	**4**
▼ **Archaea (ARCH)**	**0**
Nitrosopumilus maritimus (strain SCM1) (nmar)	0
▶ **Crenarchaeota (CREN)**	**0**
▶ **Euryarchaeota (EURY)**	**0**
▶ **Korarchaeota (KORA)**	**0**
▶ **Nanoarchaeota (NANO)**	**0**
▼ **Bacteria (BACT)**	**0**
▶ **Firmicutes (FIRM)**	**0**
▶ **Other Bacteria (OBAC)**	**0**
▶ **Proteobacteria (PROT)**	**0**

Group summary breaks down summary by protein types: A core protein is from one of the 150 core species that were initially used to form 'core' groups. A peripheral protein is from a peripheral species whose entire proteome was mapped into the 'core' groups. Peripheral proteins that do not map into a 'core' group are placed into residuals groups.

- **Do all *Cryptococcus* species currently integrated in FungiDB contain this protein?**

☐ Hide zero counts

Cryptoc ✕ ❓

Eukaryota (EUKA)	**266**
Fungi (FUNG)	**130**
Basidiomycota (BASI)	**33**
Cryptococcus cf. gattii MF34 (ccfg)	1
Cryptococcus depauperatus CBS 7841 (cdep)	1
Cryptococcus depauperatus CBS 7855 (cdcb)	1
Cryptococcus gattii CA1873 (cgac)	1
Cryptococcus gattii EJB2 (cgae)	1
Cryptococcus gattii NT-10 (cgan)	1
Cryptococcus gattii VGII R265 (cdeu)	1
Cryptococcus gattii VGIV IND107 (cgai)	1
Cryptococcus gattii WM276 (cgat)	1
Cryptococcus neoformans strain:VNII (cnev)	1
Cryptococcus neoformans var. grubii H99 (cneq)	1
Cryptococcus neoformans var. grubii KN99 (cnek)	1
Cryptococcus neoformans var. neoformans B-3501A (cnep)	1
Cryptococcus neoformans var. neoformans JEC21 (cneo)	1
Cryptococcus neoformans var. neoformans JEC21 (old build 2016-06-16) (cneo-old)	1

- **What is the most common PFAM domain associated with the proteins in this group?**

4 PFam domains

▼ PFam Legend ⬇ Download

Search this table... ❓

	Accession	Symbol	Description	Count ❓	Legend
	PF01545	Cation_efflux	Cation efflux family	251	
	PF03645	Tctex-1	Tctex-1 family	2	
	PF03102	NeuB	NeuB family	1	
	PF01423	LSM	LSM domain	1	

- **How can you look up protein alignments for *Cryptococcus*?**
  (Hint: In the "List of proteins" section, limit species to "Cryptococcus" and run ClustalOmerga tool).

3 **List of proteins**

▾ List of All Proteins ⬇ Download

To align sequences, select proteins from the table below. Then choose the 'Output format' and click the 'Run Clustal Omega for selected genes' button.

Clustal Omega	Accession	Description	Organism	Taxon	Core/Peripheral	Length	EC Numbers	Previous Groups
☑	cnev|LQV05_001641	unknown	Cryptococcus neoformans strain:VNII	Fungi	Peripheral	855	N/A	N/A
☑	cnek|CKF44_05394	unknown	Cryptococcus neoformans var. grubii KN99	Fungi	Peripheral	853	N/A	N/A
☑	cdcb|L204_05931	Cation:cation antiporter [Source:UniProtKB/TrEMBL;Acc:A0A1E3INZ6]	Cryptococcus depauperatus CBS 7855	Fungi	Peripheral	839	N/A	N/A
☑	cneo-old|CNH00620	cation:cation antiporter, putative	Cryptococcus neoformans var. neoformans JEC21 (old build 2016-06-16)	Fungi	Core	839	N/A	N/A

Check All Uncheck All

Please note: selecting a large number of proteins will take several minutes to align.

Output format: Mismatches highlighted ▾

Run Clustal Omega for selected proteins

- Examine the cluster graph for this group (it can be accessed at the bottom of the page)

Cluster Graph: OG6_106189 (266 proteins)
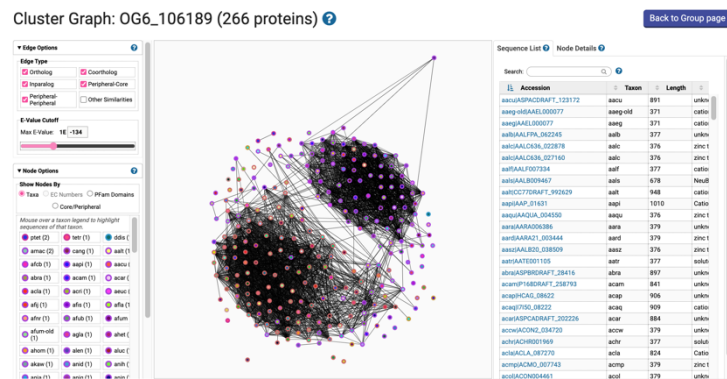
You can interact with the cluster graph. For example, move the slide to increase the E-value cutoff stringency (e.g., to a more negative number). Can you identify subclusters? Click on the nodes in the graph – notice how the organism is updated on the right.

On the left of the page in the *Node Options* panel, click on PFam Domains to see which proteins have the various PFam domains.
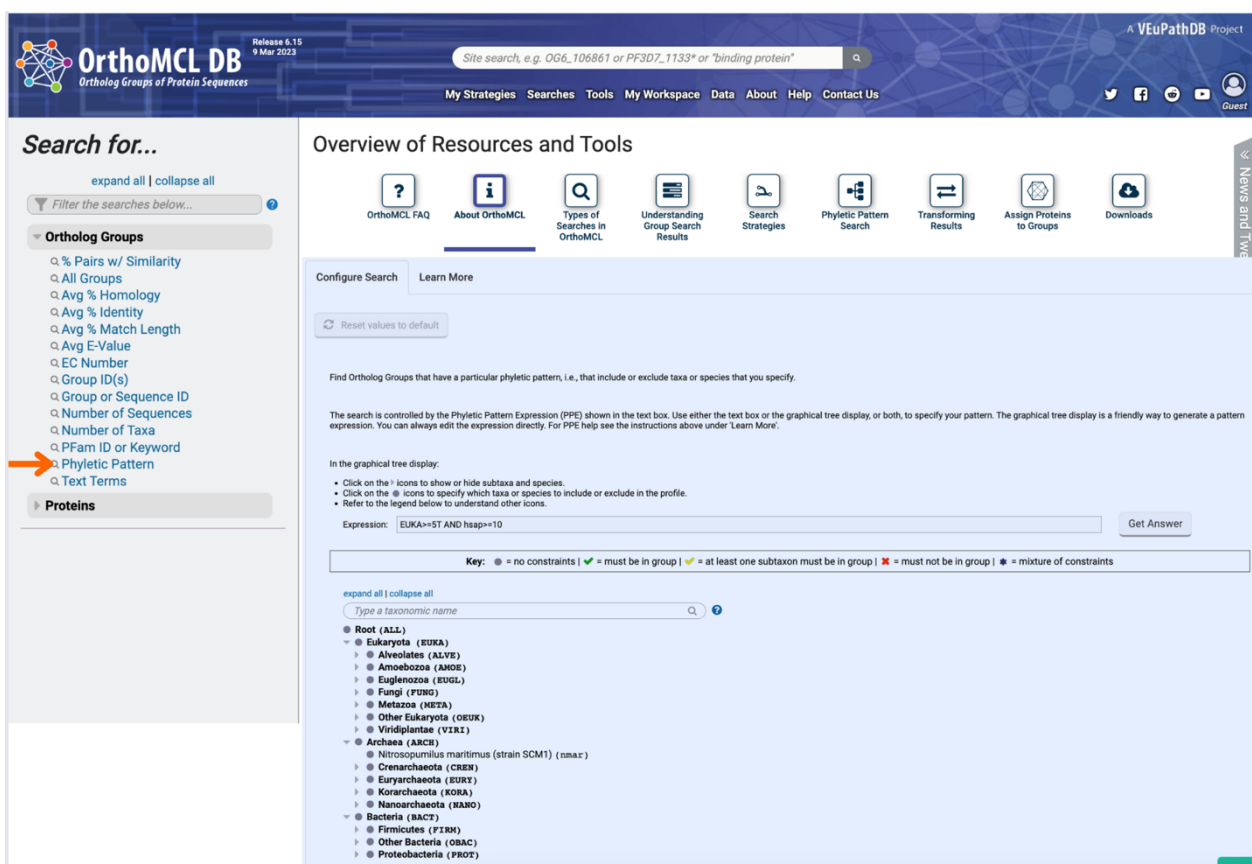
In the *Node Options* panel, you can click on *Core/Peripheral* to observe which proteins were derived from Core species and which proteins were derived from Peripheral species. Proteins from Core species were used in the initial OrthoMCL algorithm to form Core ortholog

groups. Proteins from Peripheral species were mapped into these Core groups by sequence similarity (determined by BLAST score).

Using the Phyletic Pattern search in OrthoMCL

The "Phyletic Pattern" search is an ortholog group search – look under the ortholog groups category and explore the available searches.

- **Find the "Phyletic Pattern" search.**

There are two ways to specify a phyletic pattern:
1. <u>Using the expression box.</u>

- **Run the default search for EUKA>=5T AND hsap>=10.**

- Use the "Learn More" tab to decipher the expression used above.

Description

Find Ortholog Groups by phyletic pattern.

Phyletic Pattern Expression is a flexible and powerful way to identify ortholog groups with a certain conserva proteins from specific taxa are present or absent. Also, the pattern finds groups with a certain copy number (

Examples

These expressions find ortholog groups in which...

`hsap>=5`	there are five or more human sequences
`hsap+ecol=2T`	both human and E. coli are present.
`hsap+ecol=1T`	only one species of human or E. coli is present.

2. <u>Using the selectable tree menu.</u>
You can click on the circle next to the taxon you want to include or exclude it from the search.

- **Using the "Phyletic pattern" search, identify how many eukaryotic protein groups <u>do not</u> contain orthologs from bacteria and archaea.**

  Hint: leave EUKA class with no constraints.

Strategy URL:
https://orthomcl.org/orthomcl/app/workspace/strategies/import/eebc49abcff1d99f

- **Find all groups that contain orthologs from <u>at least one species</u> of *Basydiomycota fungi* (1T) but <u>not from</u> bacteria, archaea or metazoan (0T).**

Strategy URL:
https://orthomcl.org/orthomcl/app/workspace/strategies/import/0e3f159d10e70af8

- **Examine your results and learn how to interpret the graphical representation for each group.**
  Scroll to the right of the results table examine graphical representation of the results. You can hover over each graph to learn more about phyletic distribution for each class.

Archaea ✖	Bacteria ✖	Alveolata ✖	Amoeba ✖	Euglenozoa ✖	Fungi ✖	Metazoa ✖	Viridiplantae ✖
0 / 27 (0%)	0 / 47 (0%)	0 / 137 (0%)	0 / 14 (0%)	0 / 73 (0%)	7 / 309 (2%)	0 / 124 (0%)	0 / 14 (0%)
0 / 27 (0%)	0 / 47 (0%)	0 / 137 (0%)	0 / 14 (0%)	0 / 73 (0%)	1 / 309 (0%)	0 / 124 (0%)	0 / 14 (0%)
0 / 27 (0%)	0 / 47 (0%)	109 / 137 (80%)	4 / 14 (29%)	27 / 73 (37%)	59 / 309 (19%)	0 / 124 (0%)	1 / 14 (7%)
		132 / 137 (96%)	14 / 14 (100%)	72 / 73 (99%)	1 / 309 (0%)	0 / 124 (0%)	1 / 14 (7%)
		1 / 137 (1%)	0 / 14 (0%)	4 / 73 (5%)	1 / 309 (0%)	0 / 124 (0%)	1 / 14 (7%)
0 / 27 (0%)	0 / 47 (0%)	0 / 137 (0%)	0 / 14 (0%)	0 / 73 (0%)	1 / 309 (0%)	0 / 124 (0%)	0 / 14 (0%)

Tooltip overlay:

ALVEOLATA
Ciliates:	0 / 2
Apicomplexa	
Haemosporida:	60 / 60
Coccidia:	48 / 51
Piroplasmida:	17 / 17
Other apicomplexa:	4 / 4
Other alveolata:	3 / 3

- **Revise your search to find groups that:**
  - <u>do not</u> contain orthologs from Alveolates, Amoebozoa, archaea, bacteria and Ascomycetes.
  - contain <u>at least one ortholog group</u> from *Mucor circinelloides* f. *circinelloides* 1006PhL (mcic) **AND** *Mucor circinelloides* f. *lusitanicus* CBS 277.49 (mcir).

Hint: You cannot answer this question by using the check boxes alone. For Mucor, use the expression field to finish the parameter set up manually.

If you are getting frustrated trying to figure this one out, you have a right to be! If your results look different, hover over the search step and click to revise the parameter search. The cool thing about OrthoMCL is that has an added feature that allows you to enter an expression to define the phyletic pattern. This option provides additional flexibility. Can you figure out what expression to use to answer this question? (hint: start by assigning the "do not contain" parameter (x) using check boxes to Alveolates, Amebozoa, archaea, bacteria and Ascomycetes. Next, use the expression window to add "AND" followed by specific criteria for Mucor spp. Use the learn more tab for more information.

Strategy URL: https://orthomcl.org/orthomcl/app/workspace/strategies/import/88e60b823cb2c959

If you ran a search using just check boxes, the search will be configured to look for groups that:

- do not contain orthologs from Alveolates, Amoebozoa, archaea, bacteria and Ascomycetes.
- contain ortholog groups from both *Mucor circinelloides* f. *circinelloides* 1006PhL (mcic) **AND** *Mucor circinelloides* f. *lusitanicus* CBS 277.49 must be present
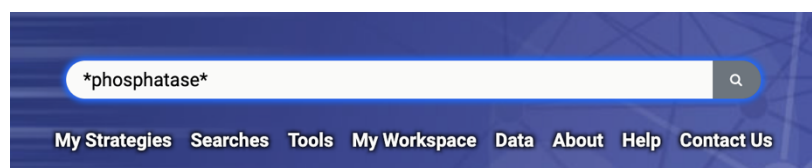
Strategy URL: https://orthomcl.org/orthomcl/app/workspace/strategies/326574153/430551723

Useful information:

All VEuPathDB genomics sites (e.g., FungiDB) have an integrated phyletic pattern search that uses OrthoMCL to return lists of genes. For example, you use the "Orthology Phylogenetic Profile" search to identify genes in your organism of interest that are restricted in their profile. For example, you frequently want to identify genes that are conserved among organisms in your genus of interest but not present in the host as these genes may make good drug targets or vaccine candidates.
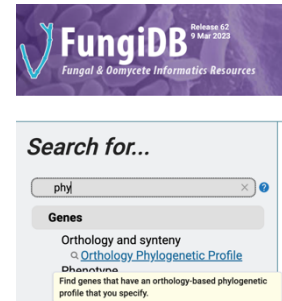
Combining searches in OrthoMCL

- **Find all fungal proteins that are likely to be phosphatases and that do not have orthologs outside of fungal kingdom.**

  a. Use the site search to look for *phosphatase* (use asterisks to find any combination of the word "phosphatase").

How many protein sequences were identified?
How many ortholog groups did you identify?

b. Display the ortholog groups containing the word phosphatase and export the results as a search strategy.
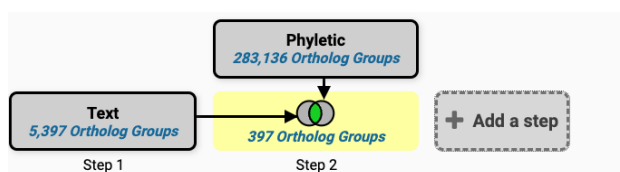

c. Add a step and run a phyletic pattern search for groups that contain any fungi proteins but do not contain any other organism outside fungi. (hint: make sure everything has a red x on it except for fungi, which should be a grey circle (no constrains)).

How many groups did the search return?

Strategy URL: https://orthomcl.org/orthomcl/app/workspace/strategies/import/56ef105e373a7ccc