Orthology and Phyletic Patterns
(Exercise 8)

8.1 Getting to OrthoMCL from EuPathDB databases
Note: For this exercise use [http://www.microsporidiadb.org](http://www.microsporidiadb.org)

a. Go to the gene page for the *Encephalitozoon cuniculi* gene with the ID: ECU07_0290

b. What does this gene do? It is annotated as hypothetical!

c. Scroll down to the table labeled “Orthologs and Paralogs within MicrosporidiaDB”. Does this gene have orthologs in other *Encephalitozoon* species? What about other organisms? (hint: click on the link below the table that take you to OrthoMCL).

d. Does this protein have orthologs in other organisms? Does it have any orthologs in bacteria or archaea? (hint: mouse over the colorful boxes in the tables to reveal the full species and phylum names – see image below).

e. Take a look at the PFAM domain architectures – do all the proteins in this group have similar domain architecture?
f. Based on the orthologs, what do you think this protein might be doing? Or if you had to give this gene a name, what would you call it?

8.2 Using the phyletic pattern tool in OrthoMCL
Note: For this exercise use [http://www.orthomcl.org](http://www.orthomcl.org)

a. How many protein groups in OrthoMCL do not have any orthologs in bacteria or archaea? (hint: go to the search “by phyletic pattern” option that can be accessed from multiple location in OrthoMCL, see image below)

![OrthoMCL DB](image)

b. How many protein groups do not contain orthologs from eukaryotes? (hint: Click on the icon to specify which taxa or species to include or exclude in the profile).

*Key:* ⊗ = no constraints | 🏀 = must be in group | ✗ = must not be in group | 🐢 = at least one subtaxon must be in group | 🌟 = mixture of constraints
8.3 Using the orthology transform tool to identify apicoplast targeted genes in *Toxoplasma* and *Neospora*.

Note: For this exercise use http://eupathdb.org

a. Start by finding genes in *Plasmodium* that are predicted to target to the apicoplast (hint: click on “Cellular Location” then on “P.f. Subcellular Localization” – see image below).

![Cellular Location](image1.png)

b. Transform the results of the above search to their *Toxoplasma* orthologs (hint: add a step, then select “Transform by Orthology”. On the search page select all *T. gondii* strains and *N. caninum*).

![Transform by Orthology](image2.png)

c. Although *Cryptosporidium* is an apicomplexan parasite it has actually lost its apicoplast! Can you use this fact to refine your results from the above search? (hint: try subtracting out any orthologs present in *Cryptosporidium* – you will need to use a nested strategy).

![Subcell Loc](image3.png)
8.4 (Optional) Using the orthology transform tool to identify as many *P. falciparum* genes containing signal peptides as possible.  
Note: For this exercise use [http://www.plasmodb.org](http://www.plasmodb.org)

a. How many genes in *P. falciparum* are annotated with signal peptides (use default settings)?

b. How many *P. vivax* genes are annotated with signal peptides use default settings)?

c. How many genes on these two lists are in common (hint, use the ortholog query to transform between organisms)?

d. How many *P. falciparum* orthologs of vivax genes with signal peptides do not themselves contain signal peptides? Why might this be the case? Look at a couple of these using the synteny viewer to generate some hypotheses.

e. Generate the most comprehensive list of *P. falciparum* genes using PlasmoDB that may contain signal peptides. How many did you find?

**8.4 (optional) Integrated searches in OrthoMCL**

Find all plant proteins that are likely phosphatases that do not have orthologs outside of plants.

- Use the keyword search to find groups that contain the word “phosphatase”
- Run a phyletic pattern search for groups that contains any plant protein but do not contain any other organism outside plants. (hint: make sure everything has a red x on it except for plants that should be a grey circle).
- Now you need to combine the above two searches with an intersect operation. (hint: from under the search menu item go to “Query History – Groups”, select the two results and click on intersect)
- How many groups did you return? If we have time explore the multiple sequence alignments from some of these groups.