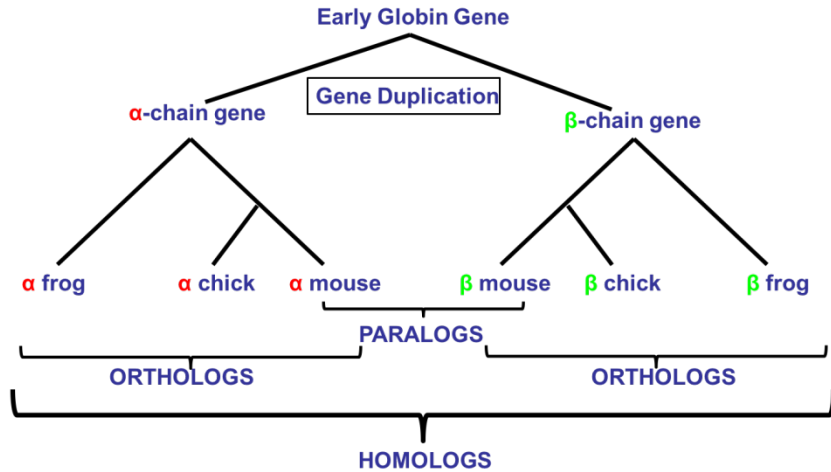# Orthology and Phyletic Patterns

Homology

1. Getting to OrthoMCL from EuPathDB databases
    Note: For this exercise use http://cryptodb.org and http://orthomcl.org/

a. Go to the gene page for the *Cryptosporidium parvum* gene with the ID: cgd7_2290
b. What does this gene do? It is annotated as a hypothetical protein!
c. Scroll down to the table labeled "Orthologs and Paralogs within CryptoDB". Does this gene have orthologs in other *Cryptosporidium* species? What about other organisms? (hint: click

on the link below the table that takes 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 table to reveal the full species and phylum names – see image below).

Group: OG5_127679
(110 sequences)
Add to Basket   Add to Favorites

Sequences & Statistics | PFam domains (graphic) | PFam domains (details) | MSA | Cluster graph

Phyletic Distribution Hide

Legend:
[0] no ortholog
[1] one ortholog
[n] more than one ortholog

FIRM   PROT   OBAC   ARCH
EUGL   AMOE   VIRI   ALVE
FUNG   META   OEUK

☑ show labels

saur	cper	bant	lmon	spne	cbot	bmal	bpse	rsol	yent	sent	cbur	vcho	ypes	sfle	ftul	ecol	cjej	wsuc	rpro	wend	bsui	atum	rtyp	gsul	cpne	mtub	drad	deth	ctep	tmar
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

mlep	syne	rbal	tpal	aaeo	nmar	hbut	smar	ssol	msed	ihos	cmaq	ckor	nequ	halo	tvol	mmar	hwal	mjan	aful	msmi	lbra	tbru	lmex	tviv	tcon	tbrg	lmaj	linf	tcru	einv
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	2	1	1	1	1	1	1	2	1

edis	ddis	ehis	gthe	rcom	atha	osat	micr	ppat	otau	crei	vcar	tpse	cmer	tthe	pviv	pfal	pber	pyoe	pkno	pcha	tpar	tann	bbov	cmur	tgon	ncan	cpar	chom	acry	ylip
1	1	1	1	2	3	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

spom	psti	ncra	scer	egos	dimm	cpos	calb	mgri	klac	dhan	anid	afum	gzea	cgla	ecun	eint	ebie	pchr	lbic	cneg	cneo	isca	dmel	aaeg	bmor	amel	cpip	phum	apis	agam
1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

nvec	tadh	drer	trub	tnig	cint	oana	mor	hsap	mmus	mdom	mmul	clup	ptro	ecab	ggal	cele	bmaa	cbri	sman	mbre	tvag	glae	glab	pram	glam
1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	2	1	3	1	1	1	1	1	1

e. Take a look at the PFAM domain architectures found under the PFam domains (graphic) tab. 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? If you had to give this gene a name, what would you call it?

2. **Using the phyletic pattern tool in OrthoMCL**
   Note: For this exercise use http://orthomcl.org/

How many protein groups in OrthoMCL <u>do not</u> have any orthologs in bacteria or archaea? (Hint: go to the "Phyletic Pattern" search in the Evolution section of the "Identify Ortholog Groups"

OrthoMCL DB
Ortholog Groups of Protein Sequences
Version 5
10 May 13
A EuPathDB Project

Groups Quick Search: synth*   Sequences Quick Search: synth*
About OrthoMCL   Help   Login   Register   Contact Us

Home | New Search | My Strategies | My Basket (0) | Tools | Data Summary | Downloads | Community | My Favorites

Data Summary
- Genomes: 150
- Protein Sequences: 1,398,546
- Ortholog Groups: 124,740

News and Tweets

Community Resources

Education and Tutorials

About OrthoMCL

Identify Ortholog Groups
Text, IDs
   Group ID(s)
   Text Terms
Evolution
   Phyletic Pattern ←
Function
   PFam ID or Keyword
   Enzyme Commission Assignment
Group Stat
   Number
   Number
   Avg % C
   % Pairs
   Avg % Id
   Avg % M
   Avg E-Va

Identify Protein Sequences
Text, IDs
   Sequence ID(s)
   Group ID(s)
   Text Terms
Function
   PFam ID or Keyword
   Enzyme Commission Assignment
   Similarity/Pattern

Tools:
- BLAST
- Assign your proteins to groups
- Download OrthoMCL software
- Web Services
- Publications mentioning OrthoMCL

Identify Groups based on Phyletic Pattern
Find Ortholog Groups that have a particular phyletic pattern, i.e., that include or exclude taxa or species that you specify.

The search is controlled by the Phyletic Pattern Expression (PPE) shown in the text box. Use either the text box or the graphical tree display, or both, to specify your pattern. The graphical tree display is a friendly way to generate a pattern expression. You can always edit the expression directly. For PPE help see the instructions at the bottom of this page.

In the graphical tree display:
- Click on -/+ to show or hide subtaxa and species.
- Click on the ⊗ icon to specify which taxa or species to include or exclude in the profile.
- Refer to the legend below to understand other icons.

Expression: BACT=0T AND ARCH=0T   Get Answer

Key: ⊗ =no constraints | ✔ =must be in group | ✘ =at least one subtaxon must be in group | ✘ =must not be in group | ✷ =mixture of constraints

⊹ Root (ALL):
⊞ ✘ Bacteria (BACT):
⊞ ✘ Archaea (ARCH):
⊞ ⊗ Eukaryota (EUKA):

Get Answer

OrthoMCL 5
©2013 The EuPathDB

Key: ⊗ =no constraints | ✔ =must be in group | ✘ =must not be in group | ✘ =at least one subtaxon must be in group | ✷ =mixture of constraints

category). To specify a phyletic pattern click on the icon next to the taxonomic group or species to include or exclude it.

   a. How many protein groups <u>do not</u> contain orthologs from eukaryotes?

   b. Find all groups that contain orthologs from at least one species of *Cryptosporidium* and *Giardia* but not from bacteria or archaea.

All EuPathDB sites also have a phyletic pattern search that uses OrthoMCL data under Genes -> Evolution -> Orthology Phylogenetic Profile. This search is very useful 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 but not present in the host as these genes may make good drug targets or vaccine candidates. Optional: go to your favorite EuPathDB site and run this search to identify all genes that are not present in human or mouse.

3. **Using the orthology transform tool to identify apicoplast targeted genes in *Toxoplasma* and *Neospora*.**
   **Note: For this exercise use http://eupathdb.org**
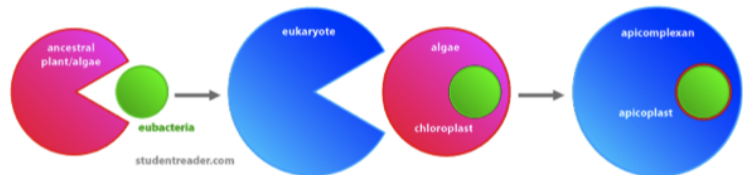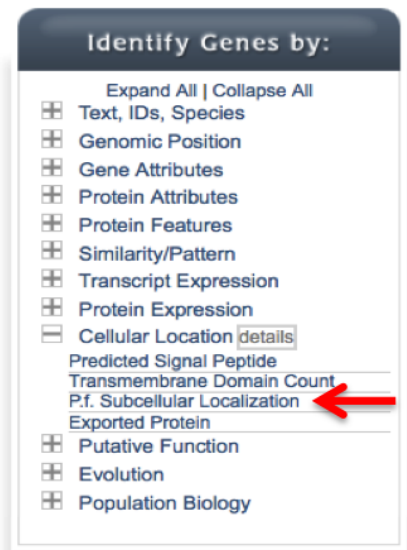
a. Start by finding genes in *Plasmodium* that are predicted to target to the apicoplast. Hint: click on "Cellular Location" then on "P.f. Subcellular Localization"; see image below.

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 *Toxoplasma* and *Neospora*.
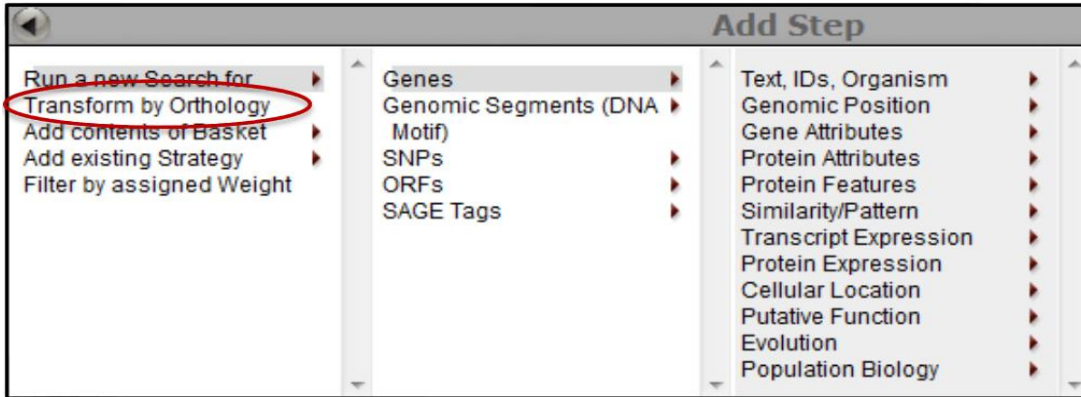
c. Although *Cryptosporidium* is an apicomplexan parasite it has actually lost its apicoplast! Can you use this fact to refine your results from the above search?

Hint: try subtracting out any orthologs present in *Cryptosporidium*. You will need to use a nested strategy.
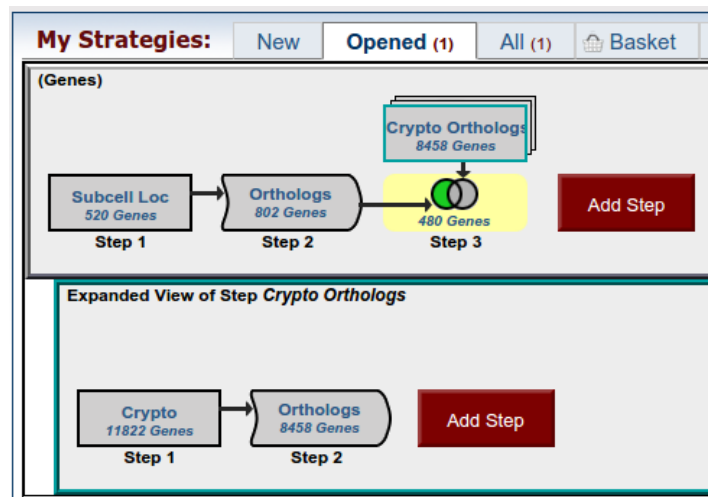
4. Combining searches in OrthoMCL (Use http://orthomcl.org for this exercise).

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

a. Use the text search to find OrthoMCL groups that contain the word "*phosphatase*" (note that the search should be run without the quotation marks but with the asterisks).

b. Add a step and run a phyletic pattern search for groups that contain 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 (Viridiplantae (VIRI)), which should be a grey circle).

c. How many groups did you return? Explore the multiple sequence alignments from some of these groups. (Hint: click on a group ID and open the MSA tab).

5.   **Exploring a specific OrthoMCL group - examining the cluster graph.** (Use
http://orthomcl.org for this exercise).

   a. Visit the orthomcl group OG5_127676. You can either type the ID in the group
      quick search option at the top of the page of follow this link:
      http://orthomcl.org/group/OG5_127676
   b. *Examine the "Sequences & Statistics" tab:* Based on the EC description and the
      product descriptions of the members of this group, what kind of a protein does
      this group represent? What is the phylogenetic distribution of the members of
      this group?
   c. *Examine the "PFam Domains (graphic)" tab:* How many PFam domains are
      represented in this group? What is the most common on? Which one is the least
      common one?
   d. *Examine the "Cluster Graph" tab:* Modify the E-value cutoff slider. What happens
      when you increase the E-value? What happens when you decrease the E-value?
      Can you identify subclusters?