Practical 6: Orthology Prediction
Group number: 2
Group members: Maximilian Senftleben, Zhong Hao Daryl Boey

Summary: The practical involved searching across the various available databases for suitable genes to perform the orthology test, followed by searching for suitable databases for the genes selected. The same genes were then searched-for across the various databases, with the hits compared and quality assessed. We also reviewed the concepts behind the various databases, and their relevance to various purposes. In addition the results were used as a benchmark to compare OMA, InParanoid, and DB Phylome, and their ability to accurately predict orthologs.

KEY QUESTIONS
1. Summarize shortly this practical.

The practical involved searching across the various available databases for suitable genes to perform the orthology test, followed by searching for suitable databases for the genes selected. The same genes were then searched-for across the various databases, with the hits compared and quality assessed. We also reviewed the concepts behind the various databases, and their relevance to various purposes.

2. Pick at least three databases that store orthologs for three of your selected genes (links provided under Material & Tools); describe the used algorithms of the databases you are comparing and motivate your choice of databases.

Gene 1: Q9ZZW7, present in OMA, InParanoid, and Tree Explorer (DB Phylome).
Gene 2: P59932, present in OMA, InParanoid, and Tree Explorer.
Gene 3: G4FFG1, present in OMA, InParanoid, and Tree Explorer.

Therefore we used these 3 databases, as they contained the genes we selected. In addition, OMA uses a comprehensive search algorithm which first involves a Smith-Waterrnan alignment, using evolutionary distance to identify close homologs, followed by clustering. This results in a broad but specific search.

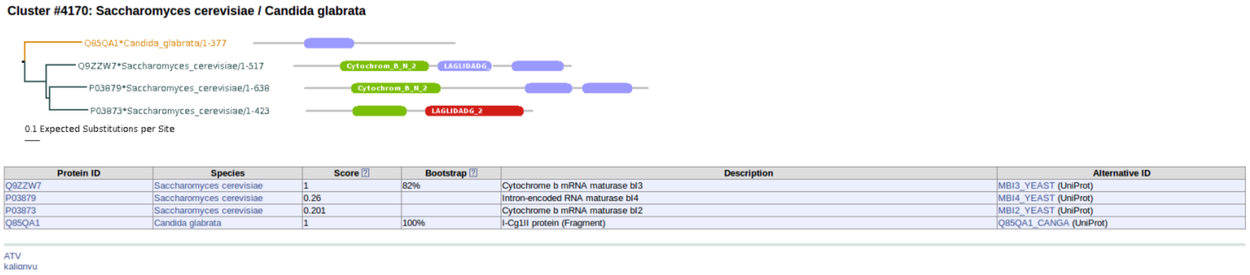
InParanoid uses a different method to find orthologs, by using pair-wise similarity scores from NCBI BLAST, followed by the creation of a seed pair of orthologs, adding sequences from the 2 reference proteomes that are similar to the seed pair, and members of this ortholog group are called inparalogs. This clustering method is faster than phylogenomic methods, and can also provide confidence values for hits, which is useful in assessing quality of the predictions.

DB Phylome employs a phylogenomic method to search for orthologs, using evolutionary tree analysis.

3. Discuss the achieved results with the different algorithms, especially the differences between their predictions (pairs, ortholog groups):

a. How do the predicted orthologs differ? Which are missing or are the same?

Gene 1, did not show any ortholog hits in OMA, returned a comprehensive tree for DB phylome (26 leaves), and multiple clusters for InParanoid.



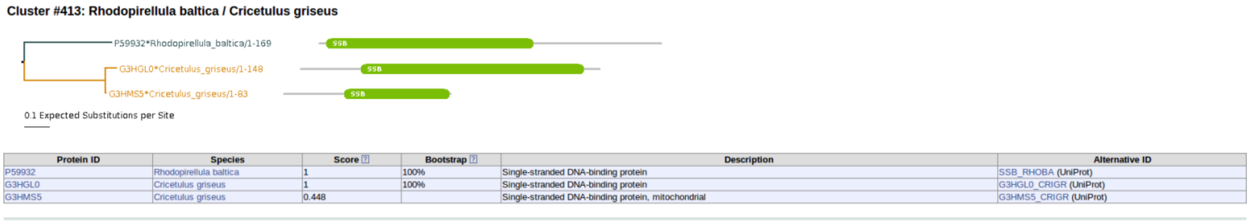
Cluster #4170: Saccharomyces cerevisiae / Candida glabrata

Q85QA1*Candida_glabrata/1-377
Q9ZZW7*Saccharomyces_cerevisiae/1-517
Cytochrom_B_N_2
LAGLIDADG_
P03879*Saccharomyces_cerevisiae/1-638
Cytochrom_B_N_2
P03873*Saccharomyces_cerevisiae/1-423
LAGLIDADG_2
0.1 Expected Substitutions per Site

Protein ID	Species	Score	Bootstrap	Description	Alternative ID
Q9ZZW7	Saccharomyces cerevisiae	1	82%	Cytochrome b mRNA maturase bI3	MBI3_YEAST (UniProt)
P03879	Saccharomyces cerevisiae	0.26		Intron-encoded RNA maturase bI4	MBI4_YEAST (UniProt)
P03873	Saccharomyces cerevisiae	0.201		Cytochrome b mRNA maturase bI2	MBI2_YEAST (UniProt)
Q85QA1	Candida glabrata	1	100%	I-Cg1II protein (Fragment)	Q85QA1_CANGA (UniProt)

ATV
kalignvu

Gene 2 returned 1632 orthologs for OMA, 44 leaves for DB phylome, and multiple clusters for InParanoid.

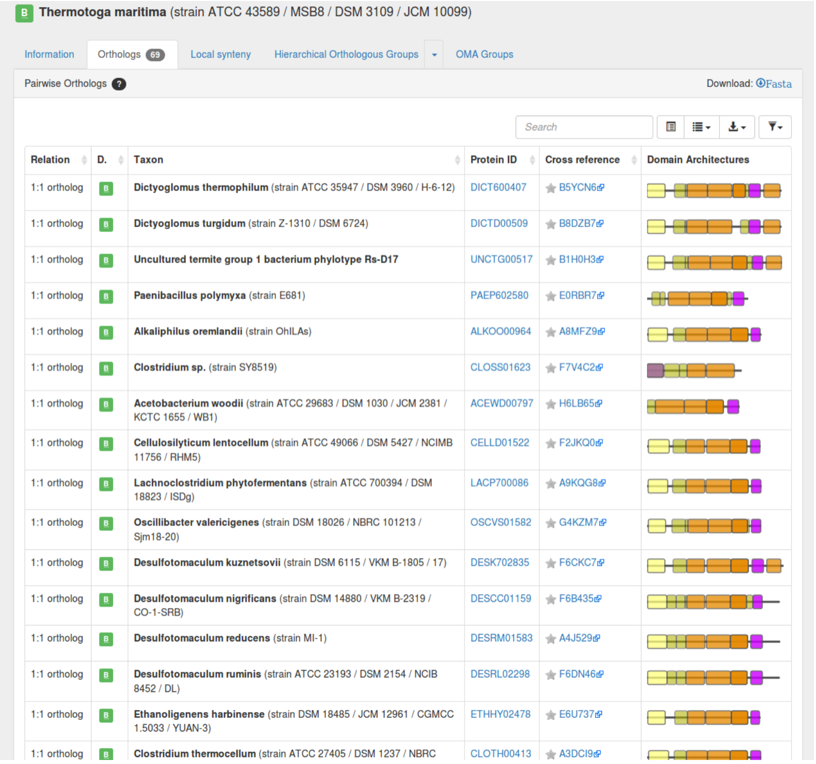


Cluster #413: Rhodopirellula baltica / Cricetulus griseus

P59932*Rhodopirellula_baltica/1-169
SSB
G3HGL0*Cricetulus_griseus/1-148
SSB
G3HMS5*Cricetulus_griseus/1-83
SSB
0.1 Expected Substitutions per Site

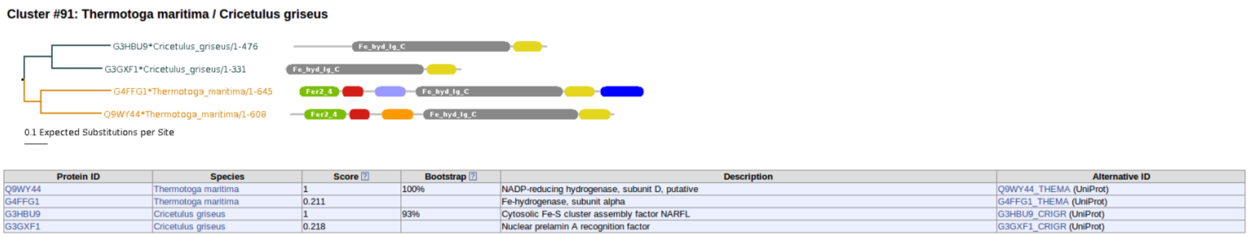
Protein ID	Species	Score	Bootstrap	Description	Alternative ID
P59932	Rhodopirellula baltica	1	100%	Single-stranded DNA-binding protein	SSB_RHOBA (UniProt)
G3HGL0	Cricetulus griseus	1	100%	Single-stranded DNA-binding protein	G3HGL0_CRIGR (UniProt)
G3HMS5	Cricetulus griseus	0.448		Single-stranded DNA-binding protein, mitochondrial	G3HMS5_CRIGR (UniProt)

Gene 3 returned 69 orthologs in OMA, 72 leaves for DB phylome, and multiple clusters for InParanoid.



B **Thermotoga maritima** (strain ATCC 43589 / MSB8 / DSM 3109 / JCM 10099)

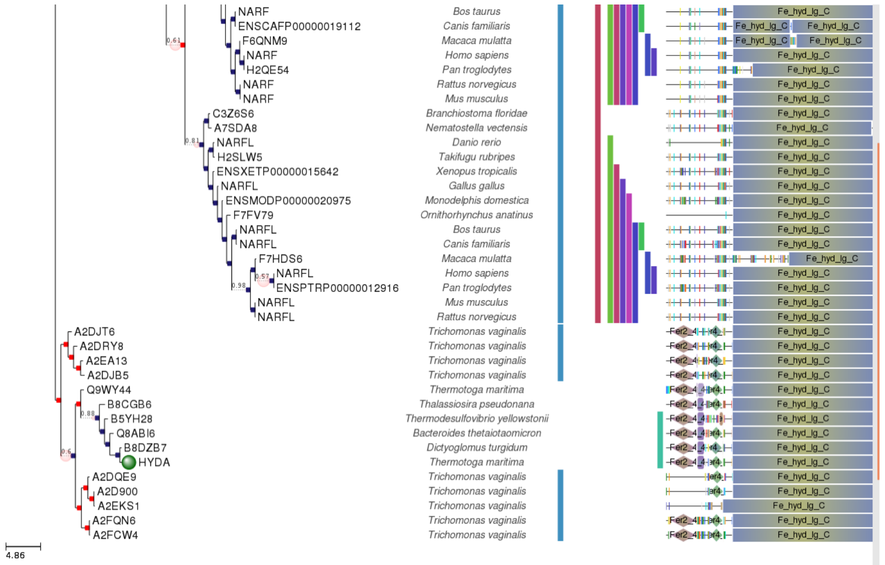
Relation	D.	Taxon	Protein ID	Cross reference	Domain Architectures
1:1 ortholog	B	**Dictyoglomus thermophilum** (strain ATCC 35947 / DSM 3960 / H-6-12)	DICT600407	B5YCN6	
1:1 ortholog	B	**Dictyoglomus turgidum** (strain Z-1310 / DSM 6724)	DICTD00509	B8DZB7	
1:1 ortholog	B	**Uncultured termite group 1 bacterium phylotype Rs-D17**	UNCTG00517	B1H0H3	
1:1 ortholog	B	**Paenibacillus polymyxa** (strain E681)	PAEP602580	E0RBR7	
1:1 ortholog	B	**Alkaliphilus oremlandii** (strain OhILAs)	ALKOO00964	A8MFZ9	
1:1 ortholog	B	**Clostridium sp.** (strain SY8519)	CLOSS01623	F7V4C2	
1:1 ortholog	B	**Acetobacterium woodii** (strain ATCC 29683 / DSM 1030 / JCM 2381 / KCTC 1655 / WB1)	ACEWD00797	H6LB65	
1:1 ortholog	B	**Cellulosilyticum lentocellum** (strain ATCC 49066 / DSM 5427 / NCIMB 11756 / RHM5)	CELLD01522	F2JKQ0	
1:1 ortholog	B	**Lachnoclostridium phytofermentans** (strain ATCC 700394 / DSM 18823 / ISDg)	LACP700086	A9KQG8	
1:1 ortholog	B	**Oscillibacter valericigenes** (strain DSM 18026 / NBRC 101213 / Sjm18-20)	OSCVS01582	G4KZM7	
1:1 ortholog	B	**Desulfotomaculum kuznetsovii** (strain DSM 6115 / VKM B-1805 / 17)	DESK702835	F6CKC7	
1:1 ortholog	B	**Desulfotomaculum nigrificans** (strain DSM 14880 / VKM B-2319 / CO-1-SRB)	DESCC01159	F6B435	
1:1 ortholog	B	**Desulfotomaculum reducens** (strain MI-1)	DESRM01583	A4J529	
1:1 ortholog	B	**Desulfotomaculum ruminis** (strain ATCC 23193 / DSM 2154 / NCIB 8452 / DL)	DESRL02298	F6DN46	
1:1 ortholog	B	**Ethanoligenens harbinense** (strain DSM 18485 / JCM 12961 / CGMCC 1.5033 / YUAN-3)	ETHHY02478	E6U737	
1:1 ortholog	B	**Clostridium thermocellum** (strain ATCC 27405 / DSM 1237 / NBRC	CLOTH00413	A3DCI9	



Cluster #91: Thermotoga maritima / Cricetulus griseus

G3HBU9*Cricetulus_griseus/1-476
Fe_hyd_Ig_C
G3GXF1*Cricetulus_griseus/1-331
Fe_hyd_Ig_C
G4FFG1*Thermotoga_maritima/1-645
Fer2_4
Fe_hyd_Ig_C
Q9WY44*Thermotoga_maritima/1-608
Fer2_4
Fe_hyd_Ig_C
0.1 Expected Substitutions per Site

Protein ID	Species	Score ?	Bootstrap ?	Description	Alternative ID
Q9WY44	Thermotoga maritima	1	100%	NADP-reducing hydrogenase, subunit D, putative	Q9WY44_THEMA (UniProt)
G4FFG1	Thermotoga maritima	0.211		Fe-hydrogenase, subunit alpha	G4FFG1_THEMA (UniProt)
G3HBU9	Cricetulus griseus	1	93%	Cytosolic Fe-S cluster assembly factor NARFL	G3HBU9_CRIGR (UniProt)
G3GXF1	Cricetulus griseus	0.218		Nuclear prelamin A recognition factor	G3GXF1_CRIGR (UniProt)



In the instance of gene 3, all the databases returned a close ortholog for B8DZB7, a gene from *D. turgidum*.

It is immediately apparent that the 3 databases produce results that are reported in very different fashions, due to varying conceptions.

b. Can you find orthologs in one database that are either missing or appear as out-paralogs in another database? Why do you think this happens?

While conducting a search for homologs in gene 3, the next evolutionarily closest ortholog reported by Phylome DB was Q8ABI6 from the organism *B. thetaiotaomicron*. However scanning the other 2 databases for this ortholog did not return a hit for this search. This is highly likely due to the varying construction methods used to produce the databases, as well as different reference genomes and databases used in construction. Therefore some databases may contain references or hits that other databases may have excluded.

c. How big are the ortholog groups for your selected genes in the databases you compare?

Gene 1, did not show any ortholog hits in OMA, returned a comprehensive tree for DB phylome (26 leaves), and multiple clusters for InParanoid. Gene 2 returned 1632 orthologs for OMA, 44 leaves for

DB phylome, and multiple clusters for InParanoid. Gene 3 returned 69 orthologs in OMA, 72 leaves for DB phylome, and multiple clusters for InParanoid.

d. What can you say about the quality of orthology predictions with the databases you compare?

It is hard to compare the quality of the predictions across the databases as they are completely differently developed tools, but InParanoid very handily clusters the ortholog hits into groups, with bootstrap scores and scores reported, allowing easy understanding of the quality of the hits, while OMA and Phylome DB are devoid of such a function.