Comparative Genomics 2018
Practical 6: Orthology Prediction

Group 11
Shuhan Xu, Fuqi Xu, Milda Valiukonyte

Summary
In this practical, we searched for orthologs in different databases and compared the result of different methods. We searched for gene P30143, P0AF03 and P0A6YB using Graph-based method (Inparanoid), tree-based method (PhylomeDB) and hybrid tree/graph method (HieranoiDB). The three methods predicted quite differently both in the number of the orthologs and the orthologs itself. Some orthologs appear in all three predictions. In general, each database has its strengths and weaknesses. If several databases produce the same result, we can be quite confident of the ortholog search.

Key questions

2.
We used Escherichia coli's genes as queries in the different databases because the species is present in all of our selected databases. In addition, we ensured that our selected databases has the species Streptomyces coelicolor since Escherichia coli has many orthologs in Streptomyces coelicolor based on our analysis of the ortholog clusters in practical 4. For our comparison of the different databases, we will search for orthologs in Streptomyces coelicolor and an additional species.

Databases: InParanoid, PhylomeDB, HieranoiDB

Organism of three selected genes: Escherichia coli

Protein identifiers for three selected genes:
P30143 (./09.fa.txt_orf00010_rev)
P0AF03 (./09.fa.txt_orf00012)
P0A6Y8 (./09.fa.txt_orf00018)

Algorithms of the databases:

InParanoid
It is a graph-based method to find ortholog groups between species. It derives pairwise similarity scores based on sequence similarity search results to construct orthology groups between species. These groups are first composed of two seed orthologs (found as best hits in the similarity search), and then inparalogs are added for which confidence values are calculated (how close to seed orthologs those sequences are).

PhylomeDB
It is a tree-based method for ortholog detection. The algorithm uses HMM to search for existing protein families and builds a multiple sequence alignment from the sequences. Then, a phylogenetic tree is constructed based on it and reconciled with NCBI taxonomy tree to infer any gene losses or duplications.

HieranoidDB
It is hierarchichal InParanoid, a hybrid graph and tree based method to find orthologs. The first step is the same as InParanoid, sequence similarity search is performed and ortholog pair is inferred based on the best hit. From these two sequences a consensus sequence is built and then another sequence similarity search is performed

for another derivation of similarity score. This process is iterated until no more significant InParanoid hits are found.

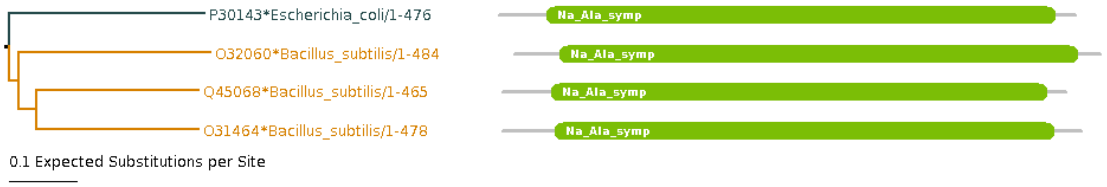
Motivation for choice of databases:
We chose InParanoid, PhylomeDB and HieranoiDB because they are based on different ortholog detection methods (graph-based, tree-based and hybrid graph/tree respectively). Hence, we can analysis how the different methods might lead to different results. In addition, all three databases, especially Inparanoid, contains many species (with both prokaryote and eukaryote), making our search fruitful. In contrast, TreeFam contains only eukaryote species. Hence, we would not be able to search our prokaryotic genes in TreeFam. Finally, all the databases provide some measure of the quality of our searches. For instance, InParanoid has InParanoid score and bootstrap support, PhylomeDB has approximate Likelihood Ratio Tests score (aLRT) and HieranoiDB has InParanoid score

3.
a. and b.

For **Escherichia coli P30143**, we searched for orthologs in Streptomyces coelicolor and Bacillus subtilis.

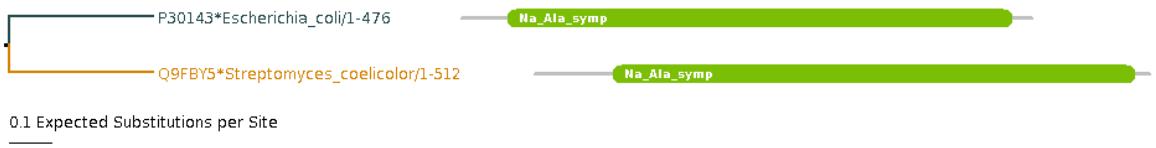
InParanoid:

Cluster #191: Escherichia coli / Bacillus subtilis



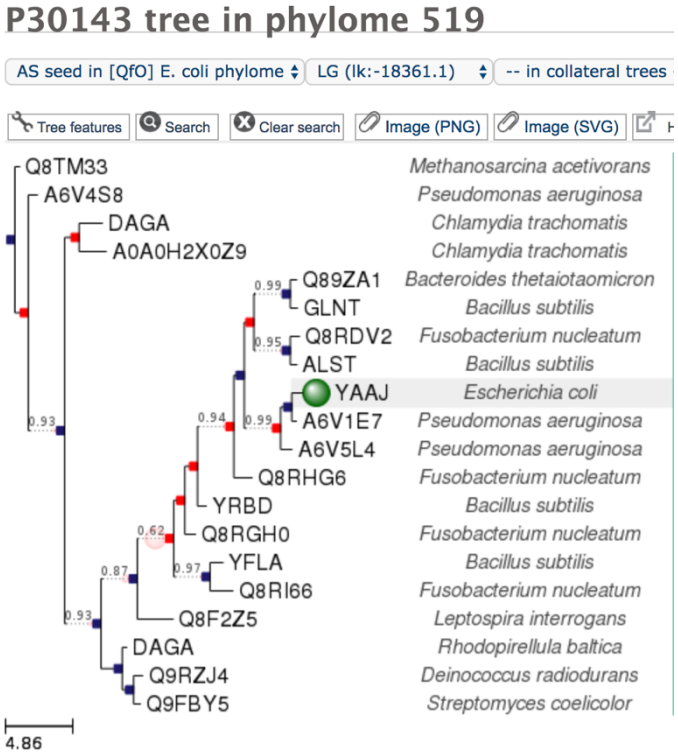
Protein ID	Species	Score	Bootstrap	Descriptio
P30143	Escherichia coli	1	100%	Uncharacterized transporter YaaJ
O31464	Bacillus subtilis	1	99%	Probable sodium/glutamine symporter GlnT
Q45068	Bacillus subtilis	0.238		Amino-acid carrier protein AlsT
O32060	Bacillus subtilis	0.063		Putative sodium/proton-dependent alanine carrier protein YrbD

Cluster #462: Escherichia coli / Streptomyces coelicolor

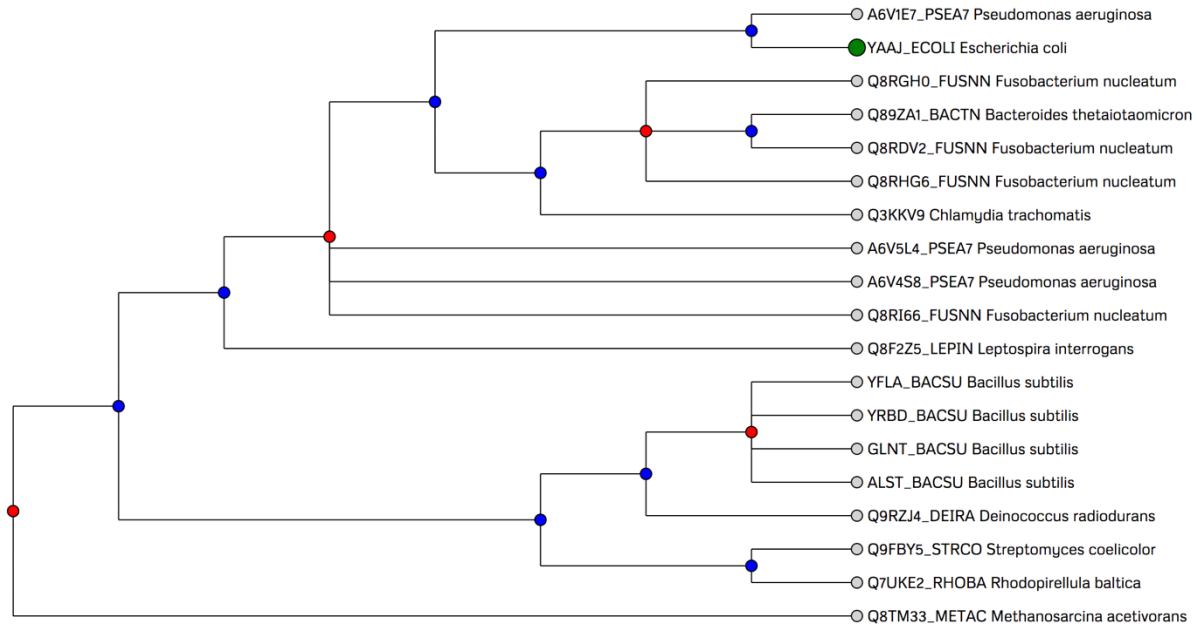


Protein ID	Species	Score	Bootstrap	De
P30143	Escherichia coli	1	100%	Uncharacterized transporter YaaJ
Q9FBY5	Streptomyces coelicolor	1	100%	Putative amino acid transport integral membrane protein

PhylomeDB:



HieranoidDB:



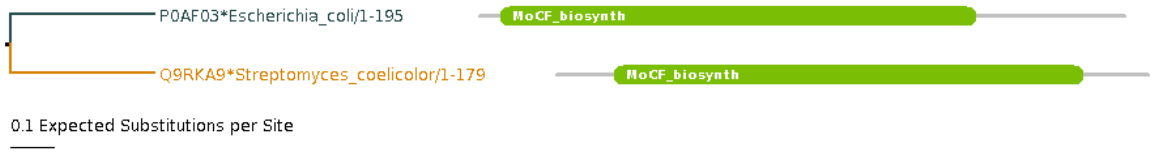
For Streptomyces coelicolor, all three databases identified Q9FBY5 as the ortholog of Escherichia coli P30143 (YAAJ).

For Bacillus subtilis, InParanoid identified O31464 (GLNT), Q45068 (ALST), O32060 (YRBD) as co-orthologs of Escherichia coli P30143(YAAJ). However, PhylomeDB identified O31464 (GLNT) and Q45068 (ALST) as co-orthologs and O32060 (YRBD) and O34708 (YFLA) as out-paralogs. HieranoidDb identified O31464 (GLNT), Q45068 (ALST), O32060 (YRBD) and O34708 (YFLA) as co-orthologs. For PhylomeDB, the additional information from the multiple sequence alignments of sequences from multiple species enables it to resolve out-paralogs from in-paralogs.

For **Escherichia coli P0AF03**, we searched for orthologs in Streptomyces coelicolor and Caenorhabditis elegans.

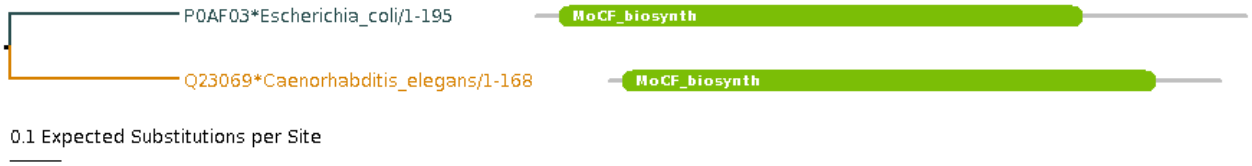
InParanoid:

Cluster #1039: Escherichia coli / Streptomyces coelicolor



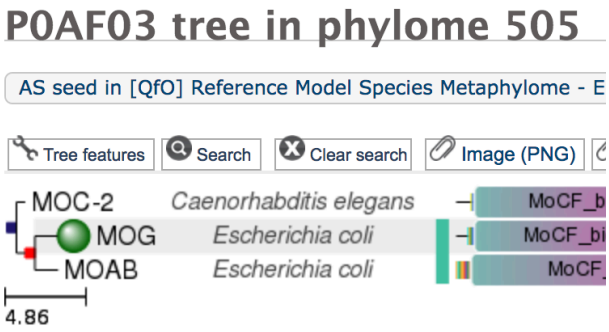
Protein ID	Species	Score	Bootstrap	Description
P0AF03	Escherichia coli	1	70%	Molybdopterin adenylyltransferase
Q9RKA9	Streptomyces coelicolor	1	100%	Molybdenum cofactor biosynthesis protein (Putative secreted protein)

Cluster #357: Escherichia coli / Caenorhabditis elegans

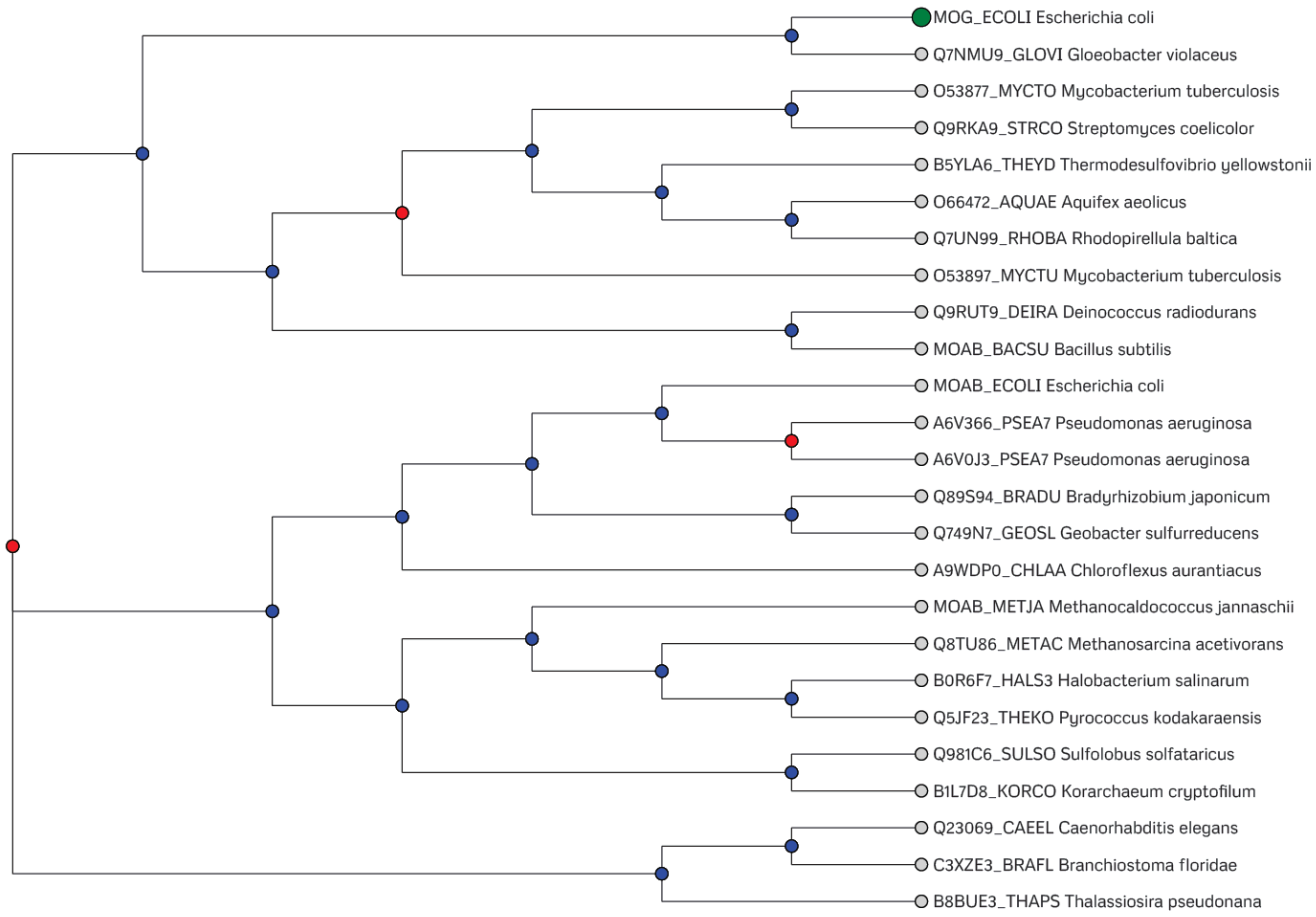


Protein ID	Species	Score	Bootstrap	
P0AF03	Escherichia coli	1	82%	Molybdopterin adenylyltransferase
Q23069	Caenorhabditis elegans	1	100%	Protein MOC-2

PhylomeDB:



HieranoidDB:



For Streptomyces coelicolor, only InParanoid and HieranoiDB identified Q9RKA9 as the ortholog of Escherichia coli P0AF03 (MOG). PhylomeDB did not identify any ortholog in Streptomyces coelicolor. The poor result of PhylomeDB could be caused by poor multiple sequence alignment for the homologs of Escherichia coli P0AF03 (MOG) which prevents it from constructing a good phylogeny tree and predicting the orthologs in other species.

For Caenorhabditis elegans, all 3 databases identified Q23069 (MOC-2) as ortholog of Escherichia coli P0AF03 (MOG). In addition, PhylomeDB also identified Escherichia coli P0AEZ9 (MOAB) as in-paralog of Escherichia coli P0AF03 (MOG).

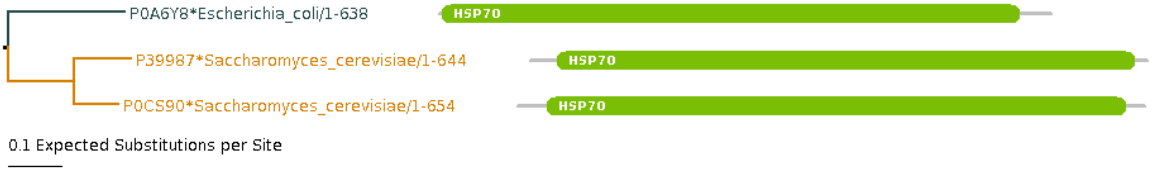
For **Escherichia coli P0A6Y8**, we searched for orthologs in Streptomyces coelicolor and Saccharomyces cerevisiae.

InParanoid:

Cluster #36: Escherichia coli / Streptomyces coelicolor

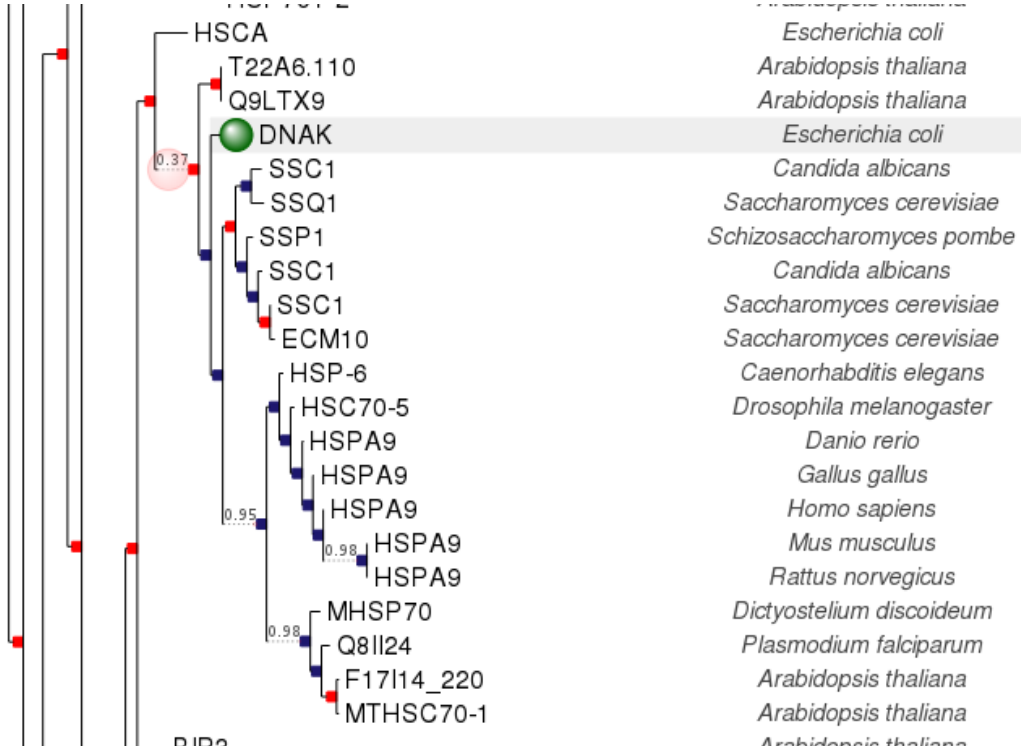
Protein ID	Species	Score	Bootstrap	
P0A6Y8	Escherichia coli	1	100%	Chaperone protein DnaK
Q05558	Streptomyces coelicolor	1	100%	Chaperone protein DnaK

Cluster #7: Escherichia coli / Saccharomyces cerevisiae

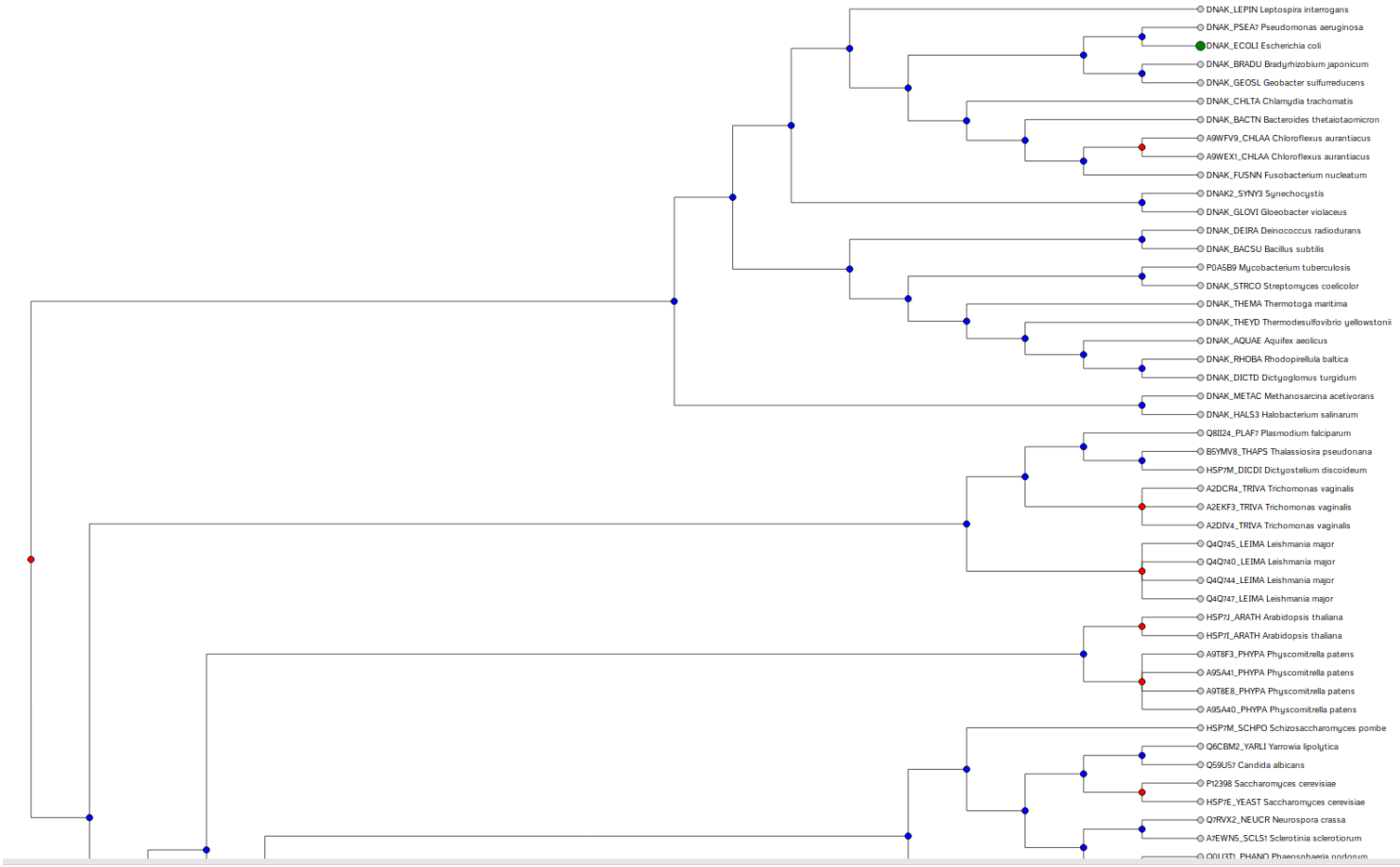


Protein ID	Species	Score	Bootstrap	
P0A6Y8	Escherichia coli	1	100%	Chaperone protein DnaK
P0CS90	Saccharomyces cerevisiae	1	99%	Heat shock protein SSC1, mitochondrial
P39987	Saccharomyces cerevisiae	0.593		Heat shock protein SSC3, mitochondrial

PhylomeDB:



HieranoidDB:



For Streptomyces coelicolor, only InParanoid and HieranoiDB identified Q05558 (DNAK_STRCO) as the ortholog of Escherichia coli P0A6Y8 (DNAK_ECOLI). PhylomeDB did not identify any ortholog in Streptomyces coelicolor. This may due to poor multiple sequence alignment in PhylomeDB.

For Saccharomyces cerevisiae, while InParanoid and PhylomeDB identified P0CS90 (P12398/SSC1) and P39987 (ECM10) as co-orthologs of Escherichia coli P0A6Y8 (DNAK_ECOLI), HieranoiDB identified P0CS90 (P12398/SSC1) and P39987 (ECM10) as out-paralogs instead. It is difficult to say which method is more accurate as there is a huge number of homologs for DNAK. In PhylomeDB, this huge group of homologs makes accurate multiple sequence alignment difficult. In HieranoiDB, the many iterations of consensus sequence generations may make the method less reliable.

c.

Size of ortholog groups for selected genes in databases

	InParanoid	PhylomeDB	HieranoiDB
P30143	11 clusters	20 genes in tree	19 genes in tree
P0AF03	36 clusters	3 genes in tree	25 genes in tree
P0A6Y8	266 clusters	150 genes in tree	75 genes in tree

d.
We can have an idea of the quality of the predictions based on the different evaluation metrics given by the different databases.

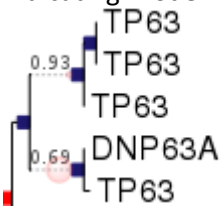
InParanoid evaluates different orthologs based on the Inparalog score and the bootstrap result. Inparanoid uses the reciprocally best-matching ortholog pairs as a seed-ortholog pair, and the Inparalog score shows how identical members in the cluster are with the seed-inparalog (score = 1.0 means identical). Inparanoid also checked prediction confidence based on the bootstrap, checking how many times the same result shows up in different sampling.

Cluster 717

Protein ID	Species	Score	Bootstrap
P30143	Escherichia coli	1	100%
B9THA8	Ricinus communis	1	100%

ref: http://inparanoid.sbc.su.se/cgi-bin/faq.cgi#clusters

In phylomeDb, it calculates an approximate Likelihood Ratio Tests (aLRT) support value instead of bootstrap value for each branch since the computation time for the former is less. A high support value (with 1.0 being highest) means high confidence in the branch. If the value is too low, the branch is marked with an icon indicating 'Node inconsistency'.



ref: http://phylomedb.org/?q=faq

HieranoiDB uses the Inparanoid score for evaluation.