# Comparative Genomics Practical 06
# Orthology Prediction
# Group 6: Tianlin He Xueqing Wang

## Summary

Orthologs are genes in different species that evolve from a common ancestor after a speciation event. In this practical, we aim at predicting the orthologues of three E.*coli* gene using three methods/databases, namely **InParanoid**, **PhylomeDB** and **OMA.** In order to perform ortholog search in these databases, the first step is to find out the correct identifier of these three genes. This is achieved by blasting the selected protein sequence against its database, which was downloaded from InParanoid. With the identifier available, it is able to search the orthologs in these three databases by typing the name of identifier. Finally, we group the search results in tables for comparison. In this session we only look for ortholog in S.*cerevisae*, A. *thaliana*, C. *albican* and S. *pombe*, as they are covered in all of these three databases.

## Objectives

The report should be formatted in one PDF file and sent in by the end of the week. It should cover all points listed below. Missing or faulty items will result in a reduced grade for this practical.
1. Short summary of what you have done (e.g. how did you find protein identifiers etc.).
2. Describe algorithms used in databases you are comparing.
3. How predictions differ (missing/same orthologs) ?
4. Detailed discussion of the results achieved with different methods and the differences between their predictions (pairs, ortholog groups).

## Activity

Perform the following steps in this order
You want to compare the orthology predictions from one database with other methods for three of your genes. **Since you want genes present in at least two methods** (i.e TreeFam and InParanoid) you should restrict your gene selection to a species that is present in selected databases.
1. To search for orthologs you first need correct protein identifiers for your predicted genes (instead of orf1234..):
a. One way to find correct identifiers is to do a local blast search with the sequence of your protein against the source files of the InParanoid database. The source files can be found in http://inparanoid.sbc.su.se/download/current/sequences/processed/

1. Download the proteome of target species (e.g. E.coli) from InParanoid in the directory

```
wget http://inparanoid.sbc.su.se/download/current/sequences/processed/226186.fasta
```

2. Convert the fasta file into a protein database

```
makeblast db -in 83333.fasta -dbtype protein
```

3. Conduct a local blast search of the protein sequence (e.g. 09.fa.txt_orf00002) against the protein database generated

```
blastp -query 09.fa.txt_orf00002 -db 226186.fasta -out geneA.out.txt
```

4. Record the name of the best hit as identifier (e.g. P00561)

| | ORF name | Identifier | E-value |
|---|---|---|---|
| Protein seq 1 | E.coli./09.fa.txt_orf00002 | P00561 | 0 |

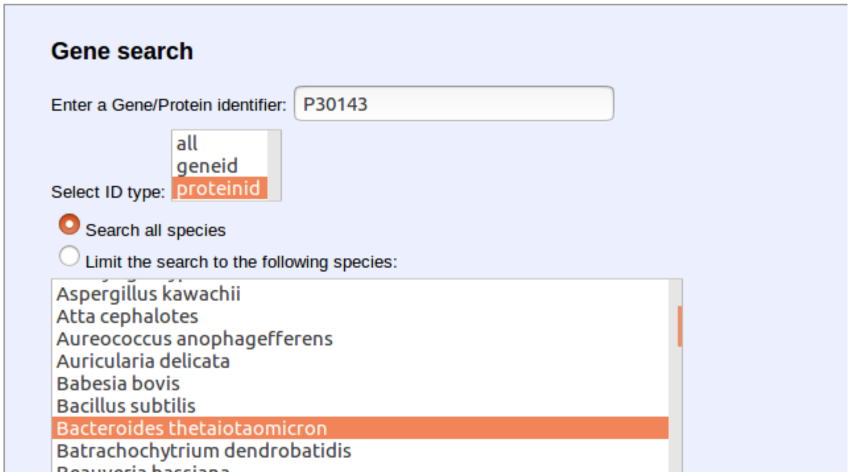
| | | | |
|---|---|---|---|
| Protein seq 2 | E.coli./09.fa.txt_orf00011 | P0A867 | 0 |
| Protein seq 3 | E.coli./09.fa.txt_orf00003 | P00547 | 0 |

b. You can also do an online blast search (e.g. at ncbi or ensembl).
2. Once you have the correct identifiers you need to find three genes which are also present in other databases. For each of the three genes pick at least 3 species for comparison. Compare the predictions (i.e GeneTree in TreeFam vs InParanoid pairs) of selected methods.

**1. InParanoid:**

1. Open the web server with InParanoid
2. Search the identifier of a protein sequence (e.g. P00561 of E.coli) against proteome of one of the three species (e.g. B. thetaiotaomicron).
3. It generates the ortholog cluster of the target protein

InParanoid8
Ortholog Groups with Inparalogs
Home | Browse | Gene search | Text search | Blast | Downloads | Summary | FAQ | Help



Gene search

Enter a Gene/Protein identifier: P30143

Select ID type: all / geneid / proteinid (selected: proteinid)

● Search all species

○ Limit the search to the following species:

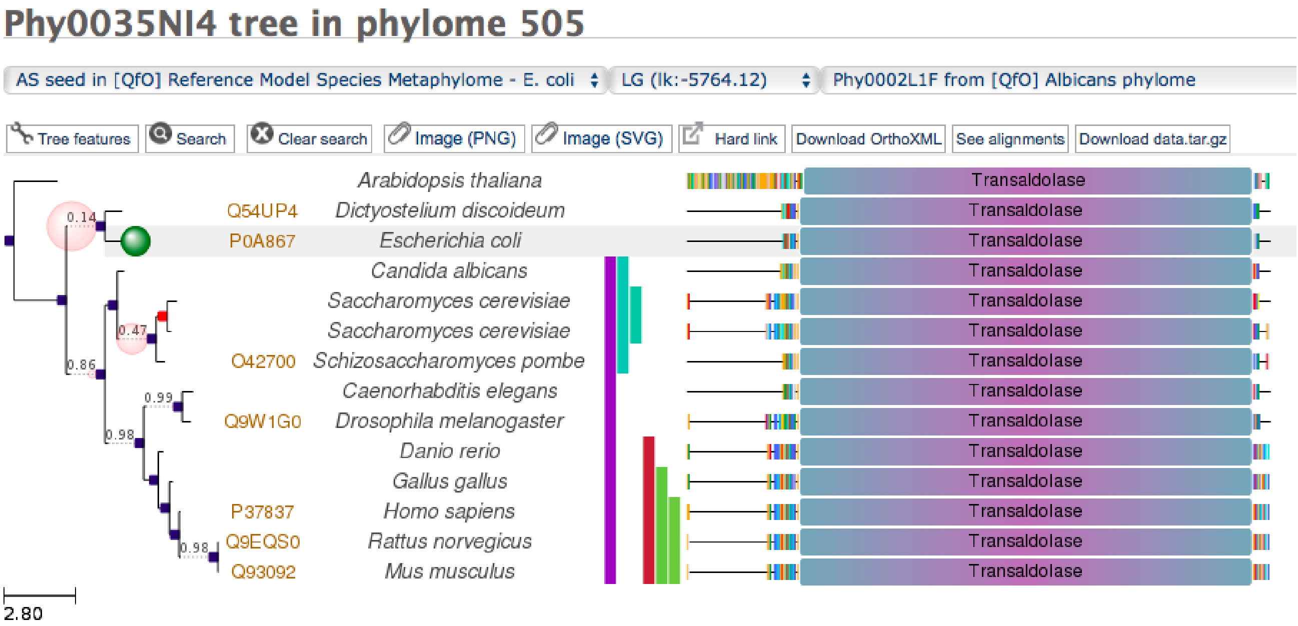
Aspergillus kawachii
Atta cephalotes
Aureococcus anophagefferens
Auricularia delicata
Babesia bovis
Bacillus subtilis
Bacteroides thetaiotaomicron
Batrachochytrium dendrobatidis

| Target protein | S.cerevisae | A.thaliana | S.pombe | c.albican |
|---|---|---|---|---|
| P0A867 | P15019 | Q5A017 | O42700 | Q5A017 |
| P00547 | P17423 | M4B3R1 | O43056 | Q92209 |
| P00561 | P10869 | O81852 | O60163 | - |

2. PhylomeDB
By searching the identifier at PhylomeDB, it displays orthologs of our search. It also generates a phylogenic tree from the orthologs between species.

| Target protein | S.cerevisae | A.thaliana | C.albican | S.pombe |
|---|---|---|---|---|
| P0A867 | Phy000CXK5<br>Phy000CYXI | Phy004E1TS | Phy0002L1F | Phy000D1YS |

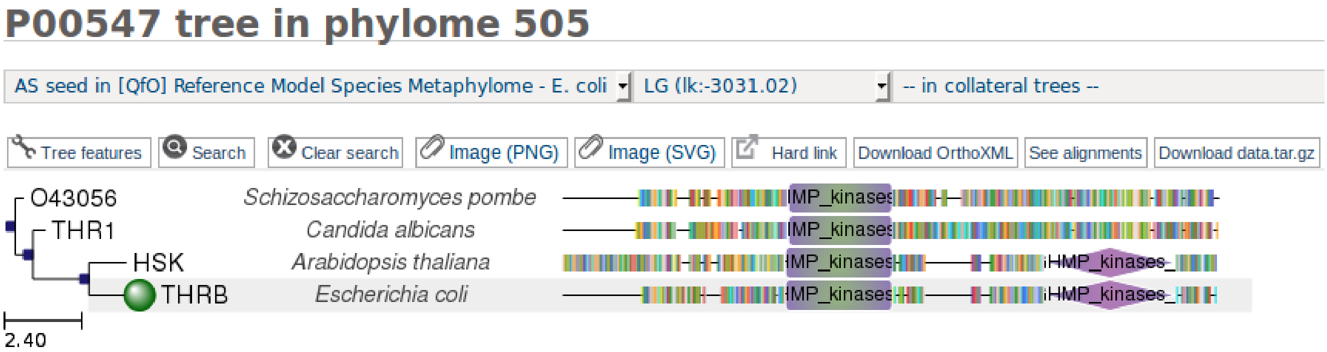
| | | | | |
|---|---|---|---|---|
| P00547 | Phy0035NH9 | Phy00018PE | Phy0002O07 | Phy000D16M |
| P00561 | Phy000CYAU | Phy0001J3A<br>Phy0001DLB<br>Phy0001HO2<br>Phy0001QQY<br>Phy0001QQY<br>Phy0001JAO | Phy0002JQ7 | Phy000D0WL |



Phy0035NI4 tree in phylome 505

AS seed in [QfO] Reference Model Species Metaphylome - E. coli | LG (lk:-5764.12) | Phy0002L1F from [QfO] Albicans phylome

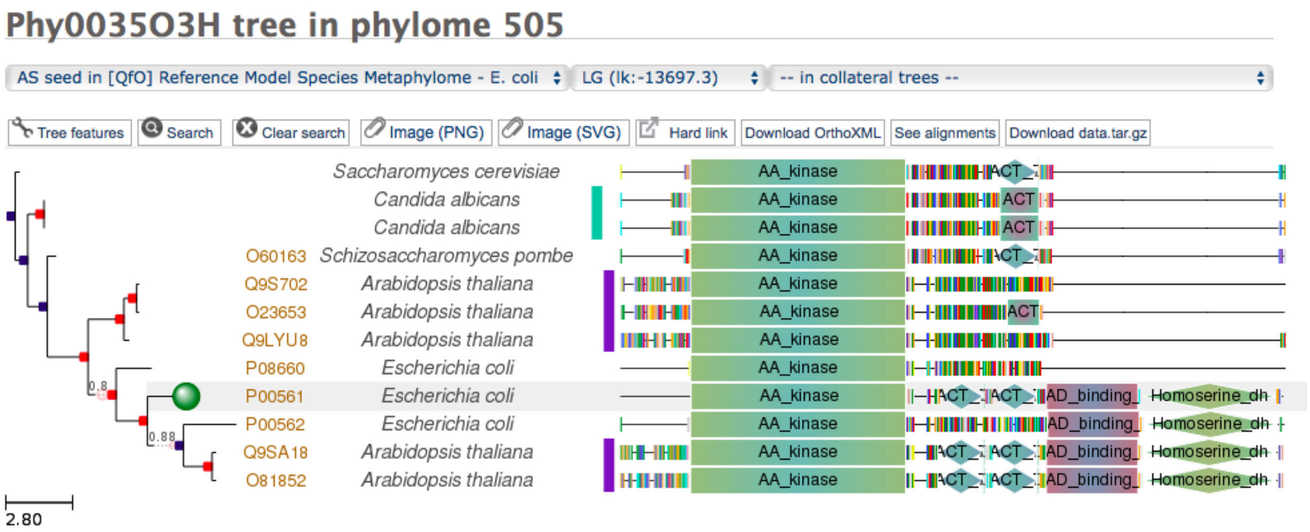
Fig. 1 Tree of P0A867



P00547 tree in phylome 505

AS seed in [QfO] Reference Model Species Metaphylome - E. coli | LG (lk:-3031.02) | -- in collateral trees --

Fig. 2 Tree of P00547



Phy0035O3H tree in phylome 505

AS seed in [QfO] Reference Model Species Metaphylome - E. coli

LG (lk:-13697.3)

-- in collateral trees --

Fig. 3 Tree of P00561

### 3. OMA

| Target protein | S.cerevisae | A.thaliana | C.albican | S.pombe |
|---|---|---|---|---|
| P0A867 | YEAST04250 | -- | CANAW03234 | SCHPO04445 |
| P00547 | YEAST02875 | ARATH05271 | CANAW04609 | SCHPO03580 |
| P00561 | - | ARATH02098<br>O81852 | - | - |

a. How does the predicted orthologs differ (missing or same)?

We use three orthologues-searching method, namely InParanoid, PhylomeDB and OMA to predict the orthologues of three E.coli genes (P0A867, P00547 and P00561) in four species (S.cerevisiae, A. thaliana, C. albican and S.pombe)

These three methods generate different results, which are listed in three tables above.

The reason why these four species are selected is that they present in all databases. However, from these three tables it is observed that they produce different orthologues pair.

For example, using phylomeDB we are able to identify orthologues of P00561 in all 4 species, but we can only find one using OMA.

For the sake of comparison, we converted the ProteinID obtained from OMA to accession number, and only the overlapping blocks are highlighted in yellow. It is observed that there exist disagreement between the results from OMA and InParanoid. It is found that only about half of blocks show mutual agreement between the two methods.

| Target protein | S.cerevisae | A.thaliana | C.albican | S.pombe |
|---|---|---|---|---|
| P0A867 | P15019 | -- | C4YLJO | O42700 |
| P00547 | P14723 | Q8L7R2 | C4YQW2 | O43056 |
| P00561 | - | O81852 | - | - |

| Target protein | S.cerevisae | A.thaliana | c.albican | S.pombe |
|---|---|---|---|---|
| P0A867 | P15019 | Q5A017 | Q5A017 | O42700 |
| P00547 | P17423 | M4B3R1 | Q92209 | O43056 |
| P00561 | P10869 | O81852 | - | O60163 |

b. Can you find orthologs in one database that are not orthologous in another database but appear as different pairs? Why do you think this happens?

Yes. We can see this from the inconsistency in the size of ortholog groups obtained from three different databases, and the mismatch between the results. As these three databases use the same proteomes, the difference is probably due to the threshold. For example, Inparanoid uses 0.05 as cut-off, while PhylomeDB uses a specific E-value.

c. How big are ortholog groups for your selected genes in your compared databases?

Ortholog groups obtained from three databases have different sizes, because they contain different number of species,  use different cut-off value, and display the results in different ways (one-to-one ortholog, inparalog, outparalog…)
Size of Ortholog groups are displayed below:

| Target protein | InParanoid | PhylomeDB | OMA |
|---|---|---|---|
| P0A867 | 220 | 132 | 632 |
| P00547 | 123 | 203 | 1131 |
| P00561 | 98 | 214 | 431 |

d. Can you say something about quality of predictions?

- The advantage of Inparanoid is that it produces a cluster of mutually best-matching hit between 2 species by NCBI-blast, effectively separate the inparalog (which are also ortholog) from outparalog. Inside this cluster, the best-hit score is usually 1 (i.e. identical to the seed inparalog), which indicates a high match between seed ortholog and the result, and it is further confirmed with bootstrapping (usually 100%). By using Blast, it is both sensitive and fast.[i]
- PhylomeDB generates reliable results as it makes use of Smith-waterman algorithm for homology search, applying a specific E-value and overlap cut-off. The consistent score (CS) and evidence level are also displayed in the ortholog list.[ii]
- Similar to Phylome DB, OMA also uses all-against-all Smith-waterman algorithm for alignment. Furthermore, it also takes into account the probability of differential gene loss, and distance-inferred uncertainty, so that the accuracy is higher.[iii]

## Discussion

Identifying the orthologs of a target protein is crucial to the construction of phylogenic tree and understanding of evolution. Nowadays, there are numerous programmes which enables us to perform a orthology search, such as InParanoid, Treefam, Panther, OMA, PhylomeDB……In this practical, we made use of three of them: Paranoid, OMA and PhylomeDB. Each of them have their own advantages: InParanoid adopts the mutual best-hit strategy with BLAST, therefore, it is capable of distinguishing the inparalog from less functionally-related outparalog, and BLAST requires less computational power; PhylomeDB and OMA are both Smith-waterman based, therefore they are sensitive in detecting distant homolog, and they also show the one-to-one, one-to-many, or even many-to-many orthologues. Besides that, PhylomeDB displays the inconsistency (if any) between several databases, and is able to construct a simple phylogenic tree based on the search research. On the other head, OMA is a powerful database because it covers many organisms and display the hierarchical relationship between groups.

Although they all use the published NCBI genomes databases, due to the difference in algorithm, organism coverage and search parameters, these three methods give us different results. It can be revealed in their discrepancy in ortholog group size, missing of ortholog in same databases and inconsistency in the ortholog found.

There are two major obstacles that we encountered during the exercise. The first one is the difference in coverage between these databases: for instance, Treefam does not include bacteria therefore we cannot search the orthologs for E.*coli*. Secondly, they adopt different nomenclature for protein: for example, InParanoid displays accession number while PhylomeDB shows the search results in Protein ID, which makes it difficult to compare and draw a final conclusion.

## iReference

Maido Remm, Christian E. Stirm, Erik L. Sonnhammer. 2001. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *Journal of Molecular Biology* 314: 1041-1052.

ii J Huerta-Cepas, S Capella-Gutierrez, LP Pryszcz, M Marcet-Houben, Gabaldon. 2014. PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome. *Nucleic Acids Research* 42(Database issue): 897-902.

iii Adrian M. Altenhoff, Nives Škunca, Natasha Glover, Clément-Marie Train, Anna Sueki, Ivana Piližota, Kevin Gori, Bartlomiej Tomiczek, Steven Müller, Henning Redestig, Gaston H Gonnet and Christophe Dessimoz. 2015. The OMA orthology database in 2015: function predictions, better plant support, synteny view, and other improvements. *Nucleic Acids Research* 43: 240-249.