

Orthology Prediction \2017

Practical 6 - Comp Gen 2017

Assistants: Daniel Morgan and Mateusz Kaduk

PURPOSE	1
OBJECTIVES	1
MATERIALS NEEDED	1
ACTIVITY	2

PURPOSE

Compare and interpret the results of different ortholog detection methods for three of your genes.

Pick at least three databases that have orthologs for three of your selected genes.

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).

MATERIALS NEEDED

1. InParanoid <http://inparanoid.sbc.su.se/download/current/>
2. TreeFam <http://www.treefam.org/download>
3. OMA <http://omabrowser.org/oma/home/>
4. PhylomeDB <http://phylomedb.org/>
5. Metaphors <http://metaphors.phylomedb.org/>
6. Panther <http://pantherdb.org/>

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.

Some databases provide files with trees reporting which species were used

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/>
 - 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.
 - a. How does predicted orthologs differ (missing or same)?
 - 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 ?
 - c. How big are ortholog groups for your selected genes in your compared databases ?
 - d. Can you say something about quality of predictions ?