Comparative Genomics 2018

Practical 6: Orthology Prediction

Assistants: Stefanie Friedrich, Miguel Castresana, Deniz Secilmis

All forms of plagiarism are forbidden, and if detected it will result in a lower grade.

PURPOSE

This practical is about comparison and interpretation of results of different ortholog detection methods for three of your genes. Prerequisite is to link identified protein sequences in your genome(s) to protein identifiers.

This practical is also meant to understand the information different databases offer, how this information is retrieved in order to classify results and draw conclusion correctly.

KEY QUESTIONS

1. Summarise shortly this practical
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
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?
    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?
    c. How big are the ortholog groups for your selected genes in the databases you compare?
    d. What can you say about the quality of orthology predictions with the databases you compare?

MATERIAL & TOOLS

1. Your genomes with predicted genes (orf….) from Practical 2
2. Databases
    a. InParanoid http://inparanoid.sbc.su.se
    b. TreeFam http://www.treefam.org/download
    c. OMA http://omabrowser.org/oma/home/

d. PhylomeDB http://phylomedb.org/

e. Metaphors http://metaphors.phylomedb.org/

f. Hieranoid http://hieranoidb.sbc.su.se

ACTIVITIES

1. Choose 3 genes: you want to compare the orthology predictions from one database with the prediction of other databases for three of your genes. Since you want that the genes are present in at least two databases (i.e TreeFam and InParanoid) you should restrict your gene selection to a species that is present in the selected databases. Some databases provide files with trees reporting of the species that were used.

2. Get the protein identifier (gene symbol): to search for orthologs you first need the 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/

      You can even try to find a gene that occurs in two of your genomes to find orthologs for (you will need to adapt the blast search pipeline: makeblastdb.. | blastp .. |); make sure that both organisms are present in at least two databases to get more interesting results.

   b. You can also do an online blast search to find the identifiers (gene symbol)

   c. Once you have the correct identifiers you need to find three genes which are also present in other databases.

3. Search for orthologs and compare the results: for each of the three genes pick at least 3 species for your comparison