Comparative and Functional Genomics. Practical session: Orthology

Note that the process of installing bioinformatics software and making it to work in your computer is often a time consuming process (specific libraries and dependencies needed, etc.). For this practical session, there is the possibility to run and produce the output files by yourself. However, for this particular case, **we recommend** students **to jump directly to the provided output files**.

In our experience from previous years, installing and running these tools can cause significant delays in the development of this practical session. Also, results can change slightly between distinct software versions, and for this reason:

IMPORTANT_1: Independently of whether you run the code or not, questions must be answered based on the output files provided by us

IMPORTANT_2: Submit your answers with your ESCI-UPF gmail account. You can only submit your answers once.

IMPORTANT_3: Wednesday 15th of May, 2024, is the last day to submit your answers.

carmen.samedi@alum.esci.upf.edu Switch accounts

Draft saved

* Indicates required question

Email *

☑ Record **carmen.samedi@alum.esci.upf.edu** as the email to be included with my response

1) BRH (Best Reciprocal Hits, or best bidirectional hits):

BRH is a fairly simple method: based on paired protein sequence alignments, it assumes that if two sequences are each other's best hits then they are **orthologs**. (For more information, check the theory slides shared for this practical session).

Download the file '**ps_orthology24.zip**' from the moodle and uncompress it.

Go to the folder named exercise1. There you will see two fasta files, called ACYPI.fa and MYZPE.fa, which contains sequences from two aphid species: *Acyrthosiphon pisum* (Pea aphid) and *Myzus persicae* (Green peach aphid). Each of the files contains ~400 proteins and we will assume they represent a whole proteome.

In order to obtain the best recirpocal hit pairs between ACYPI and MYZPE, we ran BLAST and then used the script get_BRH.py (we did it for you), which produced two output files: **TRY1.BRH.txt** contains all pairs of best reciprocal hits and **TRY1.unpaired.txt** provides a list of proteins for which the script did not find an ortholog. Results for the BLAST alignment which were required by the script correspond to **.blast files.**

Code for this section (optional. Because results can slightly change depending on the software version used, questions must be answered based on the output files provided by us):

```
#1) Install Anaconda if it is not installed in your computer:
wget https://repo.anaconda.com/archive/Anaconda3-2023.03-1-Linux-x86_64.sh
bash Anaconda3*Linux-x86_64.sh

#2) Create a conda environment for this session. You'll probably need to open a new
terminal to access conda if you just installed it.
conda create -n session4 python=2.7
conda activate session4
conda install -c bioconda blast
conda install -c bioconda orthofinder
# To deactivate conda at the end of the session:
conda deactivate

#3) We start by running local blast search to compare each proteome to the other. In order
to run a local blast search we first need to build a blast database which can be done with
makeblastdb. Then we run blastp. For detailed options you can consult the different
manuals using -help.

makeblastdb -dbtype prot -in ACYPI.fa

makeblastdb -dbtype prot -in MYZPE.fa

blastp -db ACYPI.fa -query MYZPE.fa -outfmt 6 -evalue 0.01 -out MYZPE2ACYPI.blast
blastp -db MYZPE.fa -query ACYPI.fa -outfmt 6 -evalue 0.01 -out ACYPI2MYZPE.blast
```

#4) Run the get_BRH.py script to obtain the pairs of best reciprocal hits:

```
python get_BRH.py -s1 ACYPI.fa -s2 MYZPE.fa -h1 ACYPI2MYZPE.blast -h2
MYZPE2ACYPI.blast -t TRY1
```

1.1) By looking at the output files from ***get_BRH.py***, how many orthologous relationships did you find?

● 344

○ 172

○ 144

○ 168

Clear selection

1.2) Which is the ortholog of Phy0042233_ACYPI?

○ Phy0042233_MYZPE

● Phy007ATFF_MYZPE

○ Phy0042233_ACYPI

○ None

Clear selection

1.3) Does Phy0042233_ACYPI have any other homologs in MYZPE? (Hint: this information may not be found in the output files of the script)

○ One other homolog

○ It does not have more homologs

○ Two other homologs

● Three other homologs

Clear selection

1.4) If there were sequences from ACYPI that did not align with any other sequence from MYZPE, what could they be?

○ Highly diverged inparalogs

○ Highly diverged outparalogs

● New genes

○ All are correct

Clear selection

2) Inparanoid:

Inparanoid adds a layer of complexity to the best reciprocal hits approach. BRH only considers best reciprocal hit relationships and excludes the rest of the alignment data, while Inparanoid also makes the distinction between **inparalogs** and **outparalogs**. Given one pair of ortholog sequences, if there is a sequence in the same species that has a higher blast score than the corresponding ortholog from the other species, then Inparanoid classifies it as inparalog. Additionally, Inparanoid allows using an **outgroup** to distinguish between orthologs, inparalogs and outparalogs. More details on the differences between Inparanoid and BRH can be found in the slides for this session.
Now move to the folder exercise2. Among the output files from Inparanoid, the main tables of interest to us are (1) **table.ACYPI.fa-MYZPE.fa**, which is a list of the orthologous groups, and (2) **Output.ACYPI.fa-MYZPE.fa**, which gives a more detailed overview

Code for this section (optional. Because results can slightly change depending on the software version used, questions must be answered based on the output files provided by us):

#1) We are going to use Inparanoid to predict orthologs between two species: Acyrthosiphon pisum *and* Myzus persicae. *Open inparanoid.pl and adjust parameters if needed ($use_outgroup = 0). Make sure the matrix you are going to use is in the folder as well as the two fasta files. Also make sure the lines for seqstat and blast_parser.pl are pointing to the folder where the programs are.*

#2) Execute inparanoid without an outgroup.

perl inparanoid.pl ACYPI.fa MYZPE.fa

2.1) Look at the results produced by running Inparanoid without an outgroup [exercise2/no_outgroup]. Which of the following statement/s is/are true regarding the difference between Inparanoid and BRH results?

□ The number of groups of orthologs is exactly the same, only the composition (sequences included within each group) can change

☑ In contrast to BRH, groups of orthologs from Inparanoid can include inparalogs

☑ As with the BRH approach, groups of orthologs from Inparanoid can display co-orthologous relationships

□ In contrast to BRH, groups of orthologs from Inparanoid are expected to include outparalogs

2.2) Which of the following statements are true regarding the evolutionary relationships represented in the group 79 (do not pay attention to the % in the output):

☑ Phy0042233_ACYPI is ortholog to Phy007ATFF_MYZPE and also to Phy007AV73_MYZPE

☑ Phy007ATFF_MYZPE and Phy007AV73_MYZPE are co-orthologs of Phy0042233_ACYPI

□ Phy007ATFF_MYZPE and Phy007AV73_MYZPE are inparalogs

□ Phy0042233_ACYPI and Phy007ATFF_MYZPE are outparalogs

Now we are going to run Inparanoid using a third species (SIPHA) as an outgroup

Code for this section (optional. Because results can slightly change depending on the software version used, questions must be answered based on the output files provided by us):

# 1) Move the main results to a different folder so that they are not overwritten

# 2) Run inparanoid with an outgroup (SIPHA.fa). We may need to modify the parameters in the script first ($use_outgroup = 1)

```
sed 's/$use_outgroup\ =\ 0/$use_outgroup\ =\ 1/g' ../inparanoid.pl > inparanoid_outgroup.pl
```

```
perl inparanoid_outgroup.pl ACYPI.fa MYZPE.fa SIPHA.fa
```

# 3) Aligning the sequence Phy00BX1KU_ACYPI against the proteomes of MYZPE and SIPHA (see question 2.4)

```
grep "Phy00BX1KU_ACYPI" ACYPI.fa -A 28 > Phy00BX1KU_ACYPI.fa
```

```
cat MYZPE.fa SIPHA.fa > MYZPE-SIPHA.fa
```

```
blastp -query Phy00BX1KU_ACYPI.fa -subject MYZPE-SIPHA.fa -outfmt 6 -out Phy00BX1KU_ACYPI.vs.MYZPE-SIPHA.outfmt6
```

2.4) Look at the 'exercise2/with_outgroup/**rejected_sequences.SIPHA.fa**' file. This file includes previous ortholog relationships that were discarded thanks to the addition of the outgroup species. Take as an example of 'Phy00BX1KU_ACYPI'. We have aligned this particular sequence against the proteomes of MYZPE and SIPHA using BLASTP ('exercise2/**Phy00BX1KU_ACYPI.vs.MYZPE-SIPHA.outfmt6**').

By looking at this file, write an explanation on why this sequence could have been considered as more closely related to Phy00DEWZG_SIPHA than to Phy007AURM_MYZPE? (BLAST output tabular format column header information: https://www.metagenomics.wiki/tools/blast/blastn-output-format-6). Your answer must have less than 25 words.

The higher the bit score , the better the seque

3) OrthoFinder:

Based on all-against-all sequence alignments, OrthoFinder uses a clustering algorithm (mcl) in order to represent groups of orthologs from different species (orthogroups). This way it is able to extend an analysis limited to a pair of species to multiple species. Once orthogroups have been defined, OrthoFinder reconstructs gene trees for all orthogroups and identifies gene duplication events based on gene-tree species-tree reconciliation algorithms. (A more detailed explanation on how OrthoFinder works can be found in the slides for this session as well as here https://github.com/davidemms/OrthoFinder)

Code for this section (optional. Because results can slightly change depending on the software version used, questions must be answered based on the output files provided by us):

#1) Move to the folder exercise3. There you will find two directories: set_a which contains the data of two species, and set_b which contains the data of three species.

# 2) Run OrthoFinder on two species:

```
orthofinder -f set_a/
```

3.1) Go to exercise3/set_a. Based on the output files produced by running OrthoFinder on two species (ACYPI and MYZPE), how many sequences from ACYPI and MYZPE, respectively, are in the orthogroups?

● 389, 394

○ 187, 198

○ In contrast to Inparanoid, OrthoFinder does not produce orthogroups

○ Other

Clear selection

3.2) Look at the different output files produced by OrthoFinder, many of which contain valuable information. How many orthologs have a one-to-one relationship? (your answer must be a single numeric value)

321

3.3) Based on OrthoFinder, which proteins are orthologous to Phy0042233_ACYPI? Is this result congruent with the previous results from Inparanoid? (Note that NOT all orthogroup members are orthologs among them. You may need to review the theory slides for this practical session before answering this question)

○ Phy0042233_ACYPI, Phy00BWVQD_ACYPI, Phy007ATFF_MYZPE, Phy007AV73_MYZPE. This result is not congruent with Inparanoid's results

○ Phy00BWVQD_ACYPI, Phy007ATFF_MYZPE, Phy007AV73_MYZPE. This result is not congruent with Inparanoid's results

● Phy007ATFF_MYZPE, Phy007AV73_MYZPE. This result is congruent with Inparanoid's results

○ Phy007ATFF_MYZPE, Phy007AV73_MYZPE. This result is not congruent with Inparanoid's results

Clear selection

3.4) How many duplication events are shared between ACYPI and MYZPE, and how many duplications are species-specific?

○ 21, 15

○ 10, 16

● 10, 36

○ Other

Clear selection

Now go to the folder exercise3/set_b, which includes the results produced by running OrthoFinder on three species (ACYPI, MYZPE and SIPHA)

Code for this section (optional. Because results can slightly change depending on the software version used, questions must be answered based on the output files provided by us):
#1) Run OrthoFinder on three species:

```
orthofinder -f set_b/
```

3.5) How many duplication events are shared between the three species? (Hint: There is at least a file among the OrthoFinder outputs showing the labels corresponding to each internal node of the phylogeny)

○ 13

○ 79

○ 15

● 2

Clear selection

3.6) Check which of the three species is less represented in the orthogroups. Which of the following statements are possible explanations, if any?

□ This species has more predicted genes than the other two species but some genes could be highly mutated

□ The genome of this species could have not been properly annotated, some genes may be missing in the annotations

☑ It make sense. We will always observe that the most distantly related species have fewer genes in common with the other species of the dataset

4) Phylogeny-based orthology/paralogy assignment:

For this section, we are going to use a reconstructed gene tree ('exercise4/Phy007AWWE_MYZPE.tree.txt') which you can find at the **exercise4 folder**. To build this tree we have performed a blast search using a given protein as a starting point, then we have filtered the blast results to keep only those that have an E-value below 1e-05. The multiple sequence alignment was done with mafft and the resulting fasta alignment was converted into phylip format using readAl. Finally RAxML was used to build the phylogenetic tree using the PROTGAMMALG model.

4.1) Look at the **gene tree** file, which we have printed in PDF (**exercise4/Phy007AWWE_MYZPE.tree.txt.pdf**) thanks to the software FigTree.

Following the explanation of the reconciliation and the species-overlap algorithms that can be found at the final theory slides provided for this practical session, count the number of orthologs and paralogs inferred for the **sequence Phy007AYVO_MYZPE** in the gene tree based on (i) species overlap and (ii) reconciliation. (For reconciliation, you will need the species tree file **exercise4/species_tree.txt.pdf** to be compared with the gene tree). Note that every sequence in the tree will be either ortholog or paralog to Phy007AYVO_MYZPE, we do not need to distinguish between inparalogs and outparalogs here.

Then, compare the ortholog and paralog counts retrieved by each of these two methods with those obtained for this same sequence by BRH, Inparanoid (with outgroup) and OrthoFinder (with 3 species).

Choose the correct answer:

○ According to OrthoFinder, it does not have any ortholog

● In this case, the results from BRH and Inparanoid were more informative than those from retrieved from the species overlap and the reconciliation methods

○ According to the species overlap method, Phy007AYVO_MYZPE has two orthologs

○ According to the reconciliation method, Phy007AYVO_MYZPE has two orthologs

Clear selection

Submit

Clear form