

Phylogenomics \2017

Practical 4 - Comp Gen 2017

Assistants: Daniel Morgan and Mateusz Kaduk

PURPOSE	1
MATERIALS NEEDED	1
ACTIVITY	2
Orthologous gene datasets	2
Metagene approach	3
Consensus reconstruction	4

PURPOSE

In this practical you will learn how identify all orthologs for a complete genome using the best BLAST hit. You will further align all ortholog sequences to find the metagene sequence and reconstruct a tree using Belvu.

MATERIALS NEEDED

1. BLAST
2. Target genome
3. Reference genome

ACTIVITY

Perform the following steps in this order

Orthologous gene datasets

In this exercise you will select as reference one of your prokaryotic genomes for which you will be searching for orthologs in remaining prokaryotic genomes. For

this purpose BLAST tool will be used.

1. For all of your prokaryotic genomes (including reference) find multi-fasta files with proteins from one of previous exercises.
 - a. This multi-fasta file contains all protein sequences inferred from your genomes and will be called proteome file.
2. Configure BLAST
 - a. To be able to use blast commands copy configuration file

```
cp /afs/pdc.kth.se/home/a/arnee/.ncbirc ~/
```

3. Create a BLAST database for you proteomes
 - a. In previous exercises you used makeblastdb to create a database for blastn, repeat that for all of your selected proteomes individually (this time indexed file is of protein type).
4. Do a blastp(for proteins) search against your reference proteome (against one selected proteome)

```
blastp -outfmt 5 -query <ref-proteome.fa> -db <query-proteome.fa> -out  
<output file>
```

- a. Added -m5 parameter returns XML output which is easy to parse in Biopython
 - b. Repeat this for all your query proteomes and create output XML files
5. Modify your script from the previous lab to parse the XML output.
 - a. The input for the script should be

```
<xml output> <tag for reference genome> <tag for target genome>
```

- b. Reference genome tag and target genome tag parameters are used to assign the genome name (file name) for query and target sequences.
 - c. The output for the script should be in **one line (lines below are wrapped)** with following four columns:

```
<tag for reference genome> <query protein id in reference genome>  
<tag for target genome> <best hit protein id in target genome>
```

for example

```
human proteinI chicken proteinXV
```

```
human proteinI mouse proteinXX
```

```
human proteinII chicken proteinIX
```

6. Combine the best hits into one cluster file
 - a. Use the output of your BLAST parser as input
 - b. Each query protein can have best hits in different species (i.e proteinI in chicken and mouse from example above)
 - c. Group them in one line as in example below

```
human_proteinI chicken_proteinXV mouse_proteinXX ...
```

- d. Each such line now represents cluster of orthologs
7. Select 10 different ortholog clusters such that all your prokaryotic genomes are included and acquire the corresponding protein sequences from the multifasta files.
 - a. The result should be a number of multifasta files, one for each ortholog cluster
 - b. The multifasta file should include all sequences that appear in cluster of orthologs (including query sequence)

Metagene approach

8. Make a multiple alignment for each of your multifasta cluster files (so you have one alignment for each cluster of orthologous genes).
 - a. Concatenate the alignments into a single, long metagene.

Multiple alignment 1

A1 

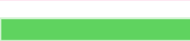
B1 

C1 

Cluster 1

Multiple alignment 2

A2 

B2 

C2 

Cluster 2



A, B, C - different species

- b. The end result should be a single fasta file with a sequence for each bacterial genome consisting of the aligned genes from each genome.
 - c. Perform tree reconstruction using Belvu. What is the tree like?
 - d. Perform sequence bootstrapping on the metagene. Can you say any-thing

on the quality of the reconstruction ?

Consensus reconstruction

9. Perform tree reconstruction using belvu for each individual alignment
 - a. You should get ten different trees (one for each ortholog cluster) in Newick format
 - b. How precise are those trees ?
 - c. Can you point a few specific genes or classes of genes that cause disagreement ?
 - d. How do you think that happens ?
10. Construct a consensus tree from the gene trees using Phylip example
 - a. For combining the tree, it is important that you use the species names as protein identifier in each tree (for example protein1 Mmusculus, where Mmusculus is species name)
 - b. Use phylip consense to construct a consensus tree, put the tree files together into a file containing a list of trees. Phylip should be pre-installed

phylip consense

- c. Name this file intree.
- d. Phylip reads the file named intree in the present directory, asks you a few questions, and then gives you the consensus tree.