Finding "corresponding" genes

How to identify “corresponding” genes across species

Assumption 1: Genes in two species that originated from the same ancestral gene in the last common ancestor of the species are good candidates for being 'corresponding' genes (orthologs)

Assumption 2: Orthologs are genes in two species that are mutually most similar to each other

Analysis

In this course, we will follow the steps of a comparative genomics analysis of plant cell wall degrading enzymes (PCDs). Plant cell walls are made up of cellulose and hemi-cellulose, making them two of the most common organic molecules on Earth and key components of the carbon cycle.

It was thought for a long time that PCDs are only produced by fungi and certain bacteria, but in recent years, evidence has accumulated that some invertebrate animals may be able to degrade plant cell walls as well.

To find out how widespread the ability to degrade plant cell wall, really is we trace the distribution of 235 potential PCDs across all eukaryotic datasets available in the RefSeq database.

Task 1: Finding corresponding genes

The task at hand is to find out which genes “correspond” to each other in different species and should be displayed in the same row.

Have another look at the assumptions at the top of the page. Our best bet to find “corresponding” genes between species is to identify orthologs. In practice, this means finding genes in two species that are mutually most similar to each other.

Finding genes with significantly similar sequences is typically done with a BLAST search. To identify orthologs, we will perform a “reciprocal best hit search”. As an example, we will use the GH45 type cellulase of *Rhizoctonia solani* (XP_043186466.1).

1. Open NCBI BLAST

2. Start a search with the following parameters:
    1. Paste the sequence accession number into the search field: XP_043186466.1
    2. Select database: Reference proteins (refseq_protein)
    3. Select target species: *Bradysia coprophila*
3. Extract the accession number of the best hit
4. Perform reverse BLAST serach in *Rhizoctonia solani*

Task 2: Exploration

Option 1: Webviewer (if currently online)

1. Open the visualizaiton of the results using this interactive web-viewer

Route 2: Local PhyloProfile installation

1. Install Miniconda

```
mkdir -p ~/miniconda3
wget
https://repo.anaconda.com/miniconda/Miniconda3-latest-Linux-x86_64.sh -
O ~/miniconda3/miniconda.sh
bash ~/miniconda3/miniconda.sh -b -u -p ~/miniconda3
rm ~/miniconda3/miniconda.sh
```

2. Install Conplex (our local software management system)

```
echo 'CONPLEX_ENVS_FILE=/home/worker/data/conplex/envs.tsv' >>
~/.bashrc
echo 'alias conplex="source /home/worker/bin/conplex.sh"' >> ~/.bashrc
source ~/.bashrc
```

3. Download and unpack the input data

```
wget
https://applbio.biologie.uni-frankfurt.de/download/fDOG_pPCD/pPCD-Phylo
geneticProfiles.tar.gz
tar -zxvf pPCD-PhylogeneticProfiles.tar.gz
```

4. Start Phyloprofile

```
conplex load phyloprofile
```

5. Upload the file you downloaded to PhyloProfile
6. Wait a moment for the data to load. Once you were redirected to the “Main profile” page, select a taxonomic rank and click the red **PLOT** button
7. Explore the “Main plot” on different taxonomic levels
    1. You can draw selection boxes on the overview plot and generate a more detailed plot of the selected section
    2. Which information is displayed in the rows, which in the columns?

3. What patterns can you observe in the plot and how do you interpret them?

Summary and Disussion

- Record your observations and questions
- Discuss with the group

Back to main