Manual Orthology analysis

Author: Kaiden R. Sewradj
Date of last update: 28/03/2025

Supervised by: Quentin Helleu, Ellen Carbo & Jean-Paul Concordet

0. Table of Contents

1. Requirements

This pipeline was designed to work with the Slurm workload manager and relies on this system to dispatch jobs. Other requirements are:

- SonicParanoid v1.3.8 (Cosentino *et al*., 2024)
- R v4.4.1 (R Core Team, 2023)
    - ggplot2 v3.5.1 (Wickham, 2016)
    - tidyr v1.2.1 (Wickham *et al*., 2024)

Except for R packages, all software is loaded with Modules v4.6.1 as follows: `module <tool>\<version number>`. All R packages are expected to be pre-installed.

2. Usage

2.1. Clone repo

```
git clone https://github.com/Kaiden-exe/orthology_analysis.git
```

2.1.1. What is in the repo?

- `scripts` [directory]
  - All scripts are here
- `config.sh`
  - Template configuration file.
- `orthology.sh`
  - Main script file that dispatches the Slurm jobs
- `orthology_summary.sh`
  - Run with: `bash orthology_summary.sh ortholog_groups.tsv`
  - Will work with any tsv file that has the same structure and column names as SonicParanoid output.
  - Run AFTER pipeline is done.
- Manual Orthology Analysis.pdf
  - You are here.

2.2. Set up config file

- `transdecoderOut`
  - Change to the directory from the transcriptomics analysis pipeline or leave it empty if all proteome files are already in the sonicIn directory.
- `sonicIn=sonic_input`
  - Put all proteome files here. Make sure the titles are ID.fasta
  - Keep file names simple as the entire filename will become column names in the results.
- `sonicOut=results`

- SonicParanoid output will be here

2.4. Run pipeline

2.4.1. Running for the first time

After following previous steps, run the pipeline with:
`bash orthology.sh -c config.sh`

2.4.2. Running with new data

To add data, add proteome files to the sonicIn directory and rerun. Be aware that cluster IDs are not conserved and will not correspond to those in previous runs.

3. Output

3.1. logfiles

All stdout and stderror logfiles are in here. Filenames include batch number and array number if applicable.

3.2. sonicOut

See here. Runs are suffixed by a date and time stamp.

4. Get a summary

The pipeline comes with a companion script that generates a summary file with the following information: amount of species analysed, number of proteins in the input, number of proteins in the ortholog_groups.tsv table, number of clusters and some

statistics on cluster sizes. In addition, two graphs are generated. The first shows the distribution of cluster sizes. Be aware that the x-axis is cut off dynamically, but may cut off the tail too early. Check this in the summary text file. The second graph is a bar graph showing how many proteins of each file are in the clustering.

To get this summary run:

```
bash orthology_summary.sh ortholog_groups.tsv
```

5. References

Cosentino, S., Sriswasdi, S. & Iwasaki, W. SonicParanoid2: fast, accurate, and comprehensive orthology inference with machine learning and language models. *Genome Biol,* **25**, 195 (2024). DOI: 10.1186/s13059-024-03298-4

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. *Springer-Verlag New York*. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org

Wickham H, Vaughan D, Girlich M (2024). tidyr: Tidy Messy Data. R package version 1.3.1, https://github.com/tidyverse/tidyr, https://tidyr.tidyverse.org