Final project - "Weiteführende Softwareentwicklung", Bio Data Science, WS 2019/20

[Max. achievable points: 100 + 7 bonus points. Submission deadline: 6.12., 23:59]

Project description

In any organism group (like Eukaryota, Fungi or Mammalia), a number of genes are shared among multiple lineages, making it probable that they were inherited from the last common ancestor of the group. Such genes, that descend from a common ancestor gene, can be clustered in what is called orthologous gene groups, OGs. Many genes are lost in specific lineages, and only few OGs are mostly universal to the taxonomic group. These patterns of gene conservation and gene loss are useful for functional and evolutionary studies.

One interesting object of study are highly conserved single-copy genes. They can be used e.g. for taxonomic classification, because they are only rarely exchanged between organisms via HGT. To study such genes, we would have to compare the genetic content of all Bacteria, Archaea and Eukaryota to identify groups of orthologous genes and organisms where they are present or absent. Luckily, we don't have to do this ourselves: It has already been done, and the results are provided in databases like the eggNOG database (*evolutionary genealogy of genes: Non-supervised Orthologous Groups*). We just need to use this data to answer our questions.

Background information

The review "Orthologs, Paralogs, and Evolutionary Genomics" (Eugene V. Koonin, Annual Review of Genetics 2005) provides a good introduction to evolutionary genomics. An early version of a database of orthologous genes was published in 2000 as "The COG database: a tool for genome-scale analysis of protein functions and evolution" (Tatusov et al., Nucleic Acids Res. 2000). This COG ("Clusters of Orthologous Groups of proteins") database was generated by comparing proteins between all sequenced microbial genomes via sequence similarity searches to infer sets of orthologs. Each COG consists of a group of proteins found to be orthologous across at least three lineages and likely corresponds to an ancient conserved domain. The eggNOG database ("eggNOG: automated construction and annotation of orthologous groups of genes", Jensen et al., NAR 2008), currently at release 5.0 ("eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses", Jaime Huerta-Cepas et al., NAR 2018), is built upon the COG database and provides information about orthologous groups on many evolutionary levels:

The data can be accessed from the eggNOG website or downloaded as tab-separated text files. E.g., the file 2157_members.tsv.gz lists archaeal OGs, pro-

teins that comprise the OGs and the taxonomy ids of the organisms. (The protein sequence identifers have the eggNOG-specific format `taxonomy_id.unique_id`.) This gives information about which genes in a particular species correspond by an evolutionary relationship to which genes in other species. The *_annotations.tsv* files contain the functional characterization of OGs (if available), as human-readable functional description and/or as single-letter functional codes (broad functional category). You might want to make a paper sketch to get a better overview over the information contained in different eggNOG files.

Questions

Let's use the information provided by eggNOG 5.0 to explore highly conserved genes (OGs) in the three domains of life, Bacteria, Archaea and Eukaryota.

Your task is to answer the following questions:

1. Which genes are universally required for an organism to survive? Being more precise: Which genes (OGs) occur in at least 99% of all genomes in the eggNOG5 database in each domain of life, respectively? (The results should be around 100-300.)
    1. How many such genes did you identify in each domain? [**10 points**]
    2. Provide the results as three files (one for each domain) listing the OGs in sorted order. [**15 points**]
2. Which common bacterial genes occur almost exclusively as single-copy genes? Being more precise: Which OGs occur in at least 50% of all bacterial genomes, and in at least 99% thereof as single-copy?
    1. Provide the results as a sorted text file. [**15 points**]
    2. How many of these OGs were also identified as universal bacterial OGs (previous question)? [**10 points**] Compare it to the number of OGs selected for the mOTUs2 profiler that we used in the metagenomics course. [**1 bonus point**]
3. Identify all OGs that occur as single-copy in at least 97% of all Archaea.
    1. How many such OGs did you identify? Provide the result as a sorted text file. [**10 points**]
    2. It would be interesting to know if there are archaeal genomes which substantially deviate from this "default" archaeal gene set. Are there Archaea which lack 4 or more of these universal OGs? Which organism (scientific name) lacks most? [**15 points**] What is its preferred growing temperature/environment? [**1 bonus point**]
4. Compile an overview of the functional categories of these 121 archaeal OGs.
    1. Provide the result as a text file sorted by the number of the functional categories. [**25 points**] The result should look something like this:

```
...
2    Replication, recombination and repair
1    Cell cycle control, cell division, chromosome partitioning
1    Signal transduction mechanisms
```

...

5. A clear and transparent project structure (folders: *bin*, *data*, *doc*, *results*, ...) together with your notes and comments is favorable. **[5 bonus points]**

Required files and scripts

Relevant data files are located on the server in these folders:

- /mirror/eggnog/eggnog_5.0/per_tax_level/2/
- /mirror/eggnog/eggnog_5.0/per_tax_level/2157/
- /mirror/eggnog/eggnog_5.0/per_tax_level/2759/

or can be downloaded from the eggNOG website. The *_members* and the *_annotations* files are required. An additional required file (*functional_categories.txt*) is provided, as well as a Bash runscript and the templates for two Python scripts, *read_members_file.py* and *annotate_cogs.py*. The scripts were tested on the server in the conda py3 environment. Please let me know if you find any bugs.

Note that the eggNOG naming scheme is somewhat inconsistent for historical reasons: bacterial OGs have names like "2Z7HP" or "COG0001", archaeal OGs have names like "2N551" or "arCOG00001" and eukaryotic OGs have names like "2QPHS" or "KOG0001". (OGs of the root taxonomic level or LUCA also have names like "COG0001", but were calculated independently from the bacterial level.)

Project requirements

1. Modify the provided Python scripts to solve the presented tasks. To obtain the max. number of points, runnable Python (and Bash) scripts need to be provided together with the results.
2. You can work in the same groups as for the exercises. The groups shouldn't exchange code with one another.
3. Structure your work as seems appropriate to you. A transparent and reproducible organization is an immanent evaluation criterion. (If I can't run the scripts and reproduce the results, I can't give the full number of points.) Test your code before submitting.
4. The scripts can be provided as a *.tar.gz* archive or as a Github repository, containing the folder structure and everything necessary to run the program, apart from the eggNOG data files.
5. Grading is based not only on the correct answer, but also on the clarity of the solution and the readability of the code.

General tips

- Develop the code in stages. Once a part works, commit the changes before continuing.

- Define all required constants (like file names) in one place, at the beginning of your script, or in a configuration file. Use no more constants than necessary.
- As the last stage of testing, run your complete workflow in a clean environment/folder.
- The code doesn't need to be highly optimized, but make sure that your program runs in reasonable time. Running each script should take no more than a few seconds.
- Designing/understanding the program logic is usually more difficult than writing the code.
- Don't forget about programming style (comments, docstrings, naming conventions etc.).