TP n°2 Unix

samuel.ortion@universite-paris-saclay.fr

2026-09

Get the data files

1. Download the `data.tar.gz` archive from `https://s.ortion.fr/w/course/mssb-computer-refresher/unix/tp/tp2/data.tar.gz`.
2. Extract the content of the archive into a folder `data`.
3. List the content of this folder.

The `csv_fields` bash function

1. Create a bash function named `csv_fields` that takes as argument the path to a CSV file and outputs in two columns: (1) the column index and (2) the column name.

   Your function should be called as follows:

   ```
   csv_fields table.csv
   ```

2. Improve this function by adding an optional argument to allow the function user to specify the field separator character in function of the format of the file (comma, tabulation, etc.). By default the field separator should be the comma ','.
3. You can add your function `csv_fields` to your `~/.bashrc` file in your home folder. Then, when you will launch a new bash session, or by typing `source ~/.bashrc` you will load the function in the current shell environment.

Hints

For bash functions as well as shell scripts, to get the value of an argument, we use "$1" for the first argument, "$2" for the second and so on.

To define a function in bash, the syntax is the following:

```
my_function() {
    first_argument="$1"
    # enter the code of the function here
}
```

To add an argument with a default value if the argument is not given, we can use the following syntax:

```
MY_VARIABLE=${VARIABLE_VALUE:-$DEFAULT_ARGUMENT_VARIABLE}
```

With this syntax, the variable `MY_VARIABLE` will take the value of the variable `DEFAULT_VARIABLE_VALUE`, when the variable `VARIABLE_VALUE` is not defined.

Sort and join

In the archive `data.tar.gz`, there are two files that contains information on the number of chromosomes that exists in some example model organisms.

The file `model_names.dat` contains the name of the organism model species. The file `model_chromosomes.dat` contains the number of chromosomes each species contains.

Both files are formatted as tab-separated values file.

An identifier is set in the first column of both files.

Merge the files with sort and join.

Check the content of the merged file.

Source: `https://en.wikipedia.org/wiki/List_of_organisms_by_chromosome_count`.

AWK on an EMBL file

In the `data.tar.gz` archive you can find the file `./data/CP116886.1.embl` which is an EMBL file containing the genome of the *Thermococcus kodakarensis*, a thermophile archeae.

AWK is a programming language dedicated to advanced file filtering on Unix.

The EMBL format is a fixed-width columnar format. To deal with such a file with AWK, we can use the special variable `FIELDWIDTH` to define the width of the columns of the file (ref).

1. Study the content of the EMBL file to get the proper column width and how to use the `FIELDWIDTH` variable in AWK.
2. Create a `awk` script to count the number of `gene` feature of an EMBL file.
3. Create a `awk` script to extract a table mapping the gene identifier (`/locus_tag`) to the length of the gene sequence from an EMBL file.
4. Create an awk script to get the size of the genome recorded in the an EMBL.

Hints

To use a script with `awk`, we use the option `-f`. A valid shebang for `awk` is `#!/usr/bin/env -S awk -f`.