Problem set 1: Valeriya_Kuznetsova

Valeriya Kuznetsova

11/12/2021

Problem 1

First, we download the tsv file using wget and unzip it with following commands (the actual chunck output is hidden by "include=FALSE command):

wget http://ftp.ebi.ac.uk/pub/databases/Pfam/current_release/proteomes/9606.tsv.gz

gunzip 9606.tsv.gz

How many unique domains (domain types) are present in human proteome?

```
cat 9606.tsv |cut -f 6,8 | grep Domain | cut -f 1 |sort -g |uniq | wc -l
```

```
## 2721
```

Returns 2721

Problem 2

Downloading files using wget (include = FALSE):

wget -r -A.faa ftp://ftp.ncbi.nlm.nih.gov/genomes/archive/old_refseq/Bacteria/Yersinia_pestis*

Checking the output:

```
   ls -F ftp.ncbi.nlm.nih.gov/genomes/archive/old_refseq/Bacteria/
```

```
## Yersinia_pestis_A1122_uid158119/
## Yersinia_pestis_Angola_uid58485/
## Yersinia_pestis_Antiqua_uid58607/
## Yersinia_pestis_biovar_Medievalis_Harbin_35_uid158537/
## Yersinia_pestis_biovar_Microtus_91001_uid58037/
## Yersinia_pestis_CO92_uid57621/
## Yersinia_pestis_D106004_uid158071/
## Yersinia_pestis_D182038_uid158073/
## Yersinia_pestis_KIM_10_uid57875/
## Yersinia_pestis_Nepal516_uid58609/
## Yersinia_pestis_Pestoides_F_uid58619/
## Yersinia_pestis_Z176003_uid47317/
```

Problem 3

```bash
find ftp.ncbi.nlm.nih.gov/ -name "*.faa" -exec cat {} \; |grep ">" | wc -l
```

```
## 48772
```

returns 48772

Problem 4

a. the average length of protein in a given E.coli strain = total number of amino acids / total number of proteins

The wget command has been silented with "include=FALSE"

```bash
echo `cat NC_000913.faa | grep -v ">" | tr -d "\n" | wc -c` / `cat NC_000913.faa | grep ">" | wc -l` | l
```

```
## 316
```

b. shell_script

To write a shell script, we create a new bash file called "shell_script_problem4" in RStudio and write a code:

'file=$1 no_prot=`cat $file | grep ">" | wc -l` aa_total=`cat $file | grep -v ">" | tr -d "\n" | wc -c` echo aa_total/no_prot | bc'

Then in the directory where the shell script and .faa file of interest are located, we run the following program:

'bash shell_script_problem4 NC_000913.faa'