Download Viral Data and preprocess it

LJM

2019-11-20

Contents

Chapter 1

Acquiring viral EggNOG data

In order to utilise EggNOGs viral data for classification of fOTUs, certain pieces of data should be downloaded from the database v.5.0.0 (Huerta-Cepas et al. 2019). It seemed quite difficult to find a complete list of all data regarding viruses in the database apart from possibly just looking at the last 28 entries in EggNOG 5.0.0 downloads page. When gathering that information following list of various virus data with different taxonomic ids could be obtained:

```
library(tidyverse)
```

```
## -- Attaching packages ----------------------------------------------------------
```

```
## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   0.8.3     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------------------------------- tidyve
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(kableExtra)
```

```
##
## Attaching package: 'kableExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```
taxids <- read_csv("../analyses/HMM_scan_using_eggNOG_HMMs/virus_taxids.csv")
```

```
## Parsed with column specification:
## cols(
##   `taxonomic name` = col_character(),
##   taxid = col_double()
## )
```

```
kable(taxids) %>%
  kable_styling(bootstrap_options = "striped",
                full_width = F,
                position = "center",
                fixed_thead = T)
```

taxonomic name	taxid
Viruses	10239
Myoviridae	10662
Siphoviridae	10699
Podoviridae	10744
Caudovirales	28883
dsRNA viruses	35325
Herpesvirales	548681
Hepadnaviridae	10404
Fuselloviridae	10474
Lipothrixviridae	10477
Tectiviridae	10656
Microviridae	10841
Inoviridae	10860
Mononegavirales	11157
Retroviridae	11632
Leviviridae	11989
ssDNA viruses	29258
Retro-transcribing viruses	35268
ssRNA positive-strand	35278
ssRNA negative-strand	35301
Nidovirales	76804
Rudiviridae	157897
Bicaudaviridae	423358
ssRNA viruses	439488
Picornavirales	464095
Tymovirales	675063
Ligamenvirales	1511857

It seems difficult to figure out what nognames each taxid has so using the

RESTFul (web) API provided by EggNOG seems out of question. However, browsing EggNOG database v. 5.0 data on per taxa level, it seems that the data is stored in systematic way where it is essential to know the taxids. Therefore, let's now download the data using the txids. The following returns some strange output and will be discarded:

```
INPUT="../analyses/HMM_scan_using_eggNOG_HMMs/virus_taxids.csv"
URL_ROOT="http://eggnog5.embl.de/download/eggnog_5.0/per_tax_level/"
OUTPUT_DIR="../analyses/HMM_scan_using_eggNOG_HMMs/eggNOG_data"
OLDIFS=$IFS
IFS=','
[ ! -f $INPUT ] && { echo "$INPUT file not found"; exit 99; }
while read taxon_name taxid; do
  # Skip the header row with taxonomic name as the first item
  if [ $taxon_name != "taxonomic name" ]; then
    #echo "Others"
    TAXON=$(echo $taxon_name | sed -e "s/[[:space:]]/_/g")
    #echo $TAXON
    # Download annotations files

    #wget -c -O "$OUTPUT_DIR""$TAXON""_""$taxid""_annotations.tsv.gz" "$URL_ROOT""$taxid""/
    #echo $OUT   """_""$taxid""_annotations.tsv.gz"
    IFS=$OLDIFS
    echo "${OUTPUT_DIR}${TAXON}_${taxid}_annotations.tsv.gz"
    #; ""$URL_ROOT""$taxid""/""$taxid""_annotations.tsv.gz"
    OLDIFS=$IFS
    IFS=','
    # Download hmms
    #wget -c -O "$OUTPUT_DIR""$TAXON""_""$taxid""_hmms.tar" "$URL_ROOT""$taxid""/""$taxid""_
    #echo "$OUTPUT_DIR""$TAXON""_""$taxid""_hmms.tar ; " "$URL_ROOT""$taxid""/""$taxid""_hm
  fi
done < $INPUT
IFS=$OLDIFS
```

Maybe another script might do the trick instead:

```
INPUT="../analyses/HMM_scan_using_eggNOG_HMMs/virus_taxids.csv"
URL_ROOT="http://eggnog5.embl.de/download/eggnog_5.0/per_tax_level/"
OUTPUT_DIR="../analyses/HMM_scan_using_eggNOG_HMMs/eggNOG_data/"
SCRIPT="../scripts/eggNOGdataDownloader.awk"

# Skip the header line
tail -n +2 $INPUT | awk -F, -v url="$URL_ROOT" -v output="$OUTPUT_DIR" -f "$SCRIPT"
```

Yes. The download succeeded!

Then we need to untar some of the downloads and …

```
DATA="../analyses/HMM_scan_using_eggNOG_HMMs/eggNOG_data/"
OUTPUT="../analyses/HMM_scan_using_eggNOG_HMMs/eggNOG_data/HMMs/"

read -r -a TARs <<< $( find $DATA -name "*.tar" -and -type f -print0 | xargs -0 echo )
for TAR in "${TARs[@]}"; do
  tar -xvf $TAR -C "$OUTPUT"
done
```

unzip some others:

```
DATA="../analyses/HMM_scan_using_eggNOG_HMMs/eggNOG_data/"
gunzip "$DATA"*.tsv.gz
```

And lastly let's rearrange the files to own directories:

```
# Move tar files
DATA="../analyses/HMM_scan_using_eggNOG_HMMs/eggNOG_data/"
mv "$DATA"*.tar "$DATA""tars/"
# Move annotation files
mv "$DATA"*.tsv "$DATA""annotations/"
```

Chapter 2

Session info

```
sessionInfo()
```

```
## R version 3.6.1 (2019-07-05)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Debian GNU/Linux 9 (stretch)
##
## Matrix products: default
## BLAS/LAPACK: /usr/lib/libopenblasp-r0.2.19.so
##
## locale:
##  [1] LC_CTYPE=C.UTF-8       LC_NUMERIC=C           LC_TIME=C.UTF-8
##  [4] LC_COLLATE=C.UTF-8     LC_MONETARY=C.UTF-8    LC_MESSAGES=C
##  [7] LC_PAPER=C.UTF-8       LC_NAME=C              LC_ADDRESS=C
## [10] LC_TELEPHONE=C         LC_MEASUREMENT=C.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] kableExtra_1.1.0 forcats_0.4.0    stringr_1.4.0    dplyr_0.8.3
##  [5] purrr_0.3.2      readr_1.3.1      tidyr_0.8.3      tibble_2.1.3
##  [9] ggplot2_3.2.1    tidyverse_1.2.1
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.2        cellranger_1.1.0  pillar_1.4.2
##  [4] compiler_3.6.1    tools_3.6.1       zeallot_0.1.0
##  [7] digest_0.6.20     viridisLite_0.3.0 lubridate_1.7.4
## [10] jsonlite_1.6      evaluate_0.14     nlme_3.1-140
## [13] gtable_0.3.0      lattice_0.20-38   pkgconfig_2.0.2
```

```
## [16] rlang_0.4.0       cli_1.1.0         rstudioapi_0.10
## [19] yaml_2.2.0        haven_2.1.1       xfun_0.9
## [22] withr_2.1.2       xml2_1.2.2        httr_1.4.1
## [25] knitr_1.24        vctrs_0.2.0       generics_0.0.2
## [28] hms_0.5.1         webshot_0.5.1     grid_3.6.1
## [31] tidyselect_0.2.5  glue_1.3.1        R6_2.4.0
## [34] readxl_1.3.1      rmarkdown_1.15    bookdown_0.13
## [37] modelr_0.1.5      magrittr_1.5      backports_1.1.4
## [40] scales_1.0.0      htmltools_0.3.6   rvest_0.3.4
## [43] assertthat_0.2.1  colorspace_1.4-1  stringi_1.4.3
## [46] lazyeval_0.2.2    munsell_0.5.0     broom_0.5.2
## [49] crayon_1.3.4
```

References

Huerta-Cepas, Jaime, Damian Szklarczyk, Davide Heller, Ana Hernández-Plaza, Sofia K Forslund, Helen Cook, Daniel R Mende, et al. 2019. "eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses." *Nucleic Acids Research* 47 (D1): D309–D314. https://doi.org/10.1093/nar/gky1085.