Searching NCBI databases in R with rentrez (+ basic iteration)

Raphael Eisenhofer

2022_07_07

Outline for today:

- 1. NCBI entrez
- 2. Basic rentrez functions
- 3. Easy iteration with the glue and map functions

1. The NCBI entrez

NCBI entrez

- More about the NCBI entrez: https://en.wikipedia.org/wiki/Entrez

- We typically interact with the NCBI (including PubMed) using the graphical user interface (GUI) of our internet browsers

- It's also possible to access the NCBI databases from an R session using the 'rentrez' package (https://github.com/ropensci/rentrez)
  - **Benefits include:**
    - Greater reproducibility
    - Slick integration and piping into other R functions (e.g. ggplot2 for visualization)
    - More powerful and sophisticated search options

2. Basic rentrez functions

See recorded video

- **Link:**

entrez_search()

entrez_search(

db = "*database of choice*",

term = "*search term*",

retmax = *number*

)

- Note we need to use **""** for **db** and **term**
- Term will default to [all fields], but can specify from fields in entrez_db_searchable() – e.g. **term** = "Eisenhofer R[AUTH]"

- We can combine multiple search terms together for refined searches!
  - E.g. **term** = "wombat **AND** Eisenhofer R[AUTH]"
  - E.g. **term** = "wombat **OR** Eisenhofer R[AUTH]"
  - E.g. **term** = "wombat **NOT** Eisenhofer R[AUTH]"
  - E.g. **term** = "Eisenhofer R[AUTH] **AND (**wombat **OR** echidna**)**"

entrez_search() output

- **entrez_search()** will create a *list* object of length five:
  - **ids**: vector of unique NCBI ID – e.g. '35785225'
  - **count**: the count of entries the search returned
  - **retmax**: the maximum number of returned results (can be increased)
  - **QueryTranslation**: search terms used – e.g. ""wombat[All Fields] AND Eisenhofer R[Author]"
  - **file**: search in XML format

entrez_fetch()

- **entrez_fetch(** will download data from the NCBI for us!
  db = *database*,
  id = *NCBI IDs to fetch* – *e.g.* '35785225'
  rettype = format for which to get data – e.g. *"fasta"* or "*abstract*"
)

  - Note that **id** can be a vector of multiple entries

  - See here for a full list of **rettypes**: https://www.ncbi.nlm.nih.gov/books/NBK25499/table/chapter4.T._valid_values_of__retmode_and/

3. Easy iteration with the glue and map functions

The glue package

- Lightweight and super easy way of basic iteration
- Say you want to create multiple search terms for a year range – e.g. 1922-2021
- First, create vector of the range: **year** <- 1952:2021
- Then, use glue to iterate: **year_searches** <- glue("wombat AND {year}[PDAT]")
- Glue will automatically place the values from the year vector into the **{ }.** We could then use the vector created by glue in our **entrez_search() term** argument.

The purrr package

- Another way of doing iteration with functions (analogous to 'apply')
- We have a vector of our year query range that we can input into entrez_search()
- We can use the **map_dbl()** function in purrr to iterate entrez_search()
- **map_dbl(**
  .x = *list or vector of things you want to iterate,* (e.g. **year_searches**)
  .f = *function or formula* (e.g. ~entrez_search(db = "pubmed", term = .x)
)
- The above will run the entrez_search() function for each entry in **year_searches**
- **N.B.** a tilde '~' is needed before the function

Exercise time!

- See the exercises in “exercises.R”

Summary

- rentrez is a fantastic package for dealing with the NCBI database

- glue is an easy-to-use function for iteration!

Acknowledgements to Pat Schloss for inspiration for the lesson (https://www.youtube.com/watch?v=QX5aIzG8SQk)