The Journal of Open Source Software

restez: Create and Query a Local Copy of GenBank in R

Dominic J. Bennett[1, 2], **Hannes Hettling**[3], **Daniele Silvestro**[1, 2], **Rutger Vos**[3], **and Alexandre Antonelli**[1, 2, 4]

1 Gothenburg Global Biodiversity Centre, Box 461, SE-405 30 Gothenburg, Sweden **2** Department of Biological and Environmental Sciences, University of Gothenburg, Box 461, SE-405 30 Gothenburg, Sweden **3** Naturalis Biodiversity Center, P.O. Box 9517, 2300 RA Leiden, The Netherlands **4** Gothenburg Botanical Garden, SE 41319 Gothenburg, Sweden

DOI: 10.21105/joss.01102

Software

- Review
- Repository
- Archive

Submitted: 26 November 2018
Published: 27 November 2018

License
Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License (CC-BY).

Summary

Downloading sequences and sequence information from GenBank (Benson et al., 2013) and related NCBI databases is often performed via the NCBI API, Entrez (J. Ostell, 2002). Entrez, however, has a limit on the number of requests, thus downloading large amounts of sequence data in this way can be inefficient. For situations where a large number of Entrez calls is made, downloading may take days, weeks or even months and could result in a user's IP address being blacklisted from the NCBI services due to server overload. Additionally, Entrez limits the number of entries that can be retrieved at once, requiring a user to develop code for querying in batches.

The `restez` package (D. J. Bennett, 2018a) aims to make sequence retrieval more efficient by allowing a user to download the GenBank database, either in its entirety or in subsets, to their local machine and query this local database instead. This process is more time efficient as GenBank downloads are made via NCBI's FTP server using compressed sequence files. With a good internet connection and a computer with currently standard capabilities, a database comprising 7 GB of sequence information (i.e. the total sequence data available for Rodentia as of 27 June 2018) can be generated in less than 10 minutes. (For an outline of the functions and structure of `restez`, see Figure 1.)

Rentrez integration

`rentrez` (Winter, 2017) is a popular R package for querying NCBI's databases via Entrez in R. To maximize the compatibility of `restez`, we implemented wrapper functions with the same names and arguments as the `rentrez` equivalents. Whenever a wrapper function is called the local database copy is searched first. If IDs are missing in the local database a secondary call to Entrez is made via the internet. This allows for easy employment of `restez` in scripts and packages that are already using `rentrez`. At a minimum, a user currently using `rentrez` will only need to create a local subset of the GenBank database, call `restez` instead of `rentrez` and ensure the `restez` database is connected.

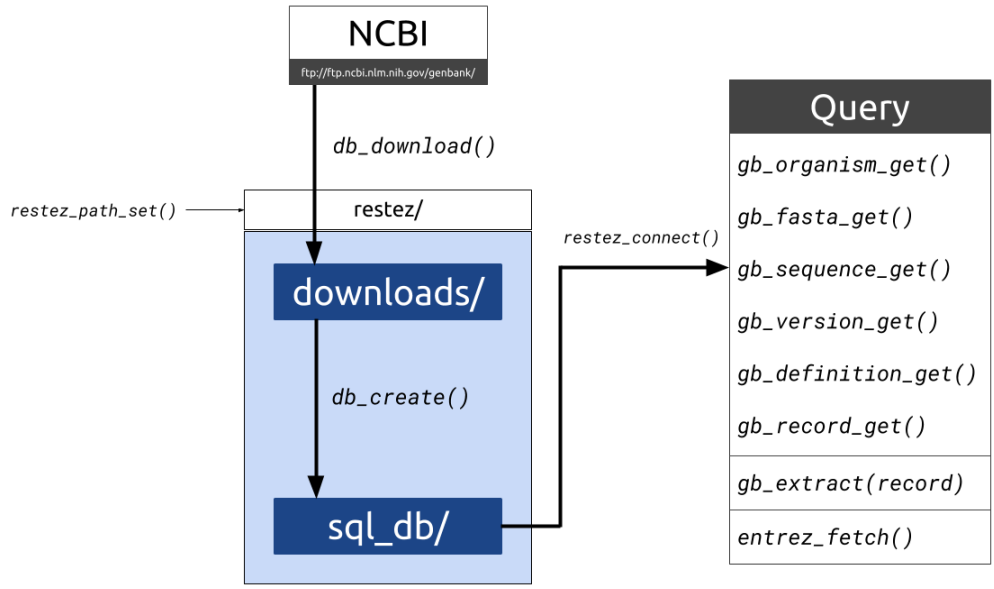

Figure 1: The functions and file structure for downloading, setting up and querying a local copy of GenBank

Examples

A small example

After a restez database has been set-up, we can retrieve all the sequences from an `rentrez::entrez_search()` with a single command.

```
# Use rentrez to search for accession IDs of interest
# Sequences in fasta format can then be retrieved with entrez_fetch
res <- rentrez::entrez_fetch(db = 'nuccore', id = ids, rettype = 'fasta')
# ^ likely to raise an error if too many IDs
res <- restez::entrez_fetch(db = 'nuccore', id = ids, rettype = 'fasta')
# ^ not likely to raise an error
```

A large example

`phylotaR` is an R package for retrieving and identifying orthologous sequence clusters from GenBank as a first step in a phylogenetic analysis (D. Bennett et al., 2018). Because the package runs an automated pipeline, multiple queries to GenBank via Entrez are made using the `rentrez` package. As a result, for large taxonomic groups containing well-sequenced organisms the pipeline can take a long time to complete.

```
library(phylotaR)
# run phylotaR pipeline for New World Monkeys
txid <- 9479  # taxonomic ID
setup(wd = 'nw_monkeys', txid = txid)
run(wd = wd)
# ^ takes around 40 minutes
```

We can download and create a local copy of the primates GenBank locally and re-run the above code with a library call to `restez` for speed-up gains and increased code reliability.

```
# setup database
library(restez)
# Specify path to a local directory in which database will be stored
# Make sure you have sufficient disk space!
restez_path_set(filepath = 'restez_db')
db_download(db = 'nucleotide') # Interactively download GenBank data
db_create(db = 'nucleotide')
```

Now when re-running the first `phylotaR` code block with the inclusion of the `restez` package, the procedure completes approximately eight times faster.

```
# run phylotaR again
library(phylotaR)
library(restez)
restez_path_set(filepath = 'restez_db')
txid <- 9479
setup(wd = 'nw_monkeys', txid = txid)
run(wd = wd)
# ^ takes around 5 minutes
```

For more detailed and up-to-date examples and tutorials, see the `restez` GitHub page (D. J. Bennett, 2018b).

Availability

`restez` is open source software made available under the MIT license. It can be installed through CRAN (D. J. Bennett, 2018c), `install.package("restez")`, or from its GitHub source code repository using the `devtools` package, e.g. as follows: `devtools::install_github("ropensci/restez")`

Funding

This package has been developed as part of the supersmartR project (D. J. Bennett, 2018d) which has received funding through A.A. (from the Swedish Research Council [B0569601], the Swedish Foundation for Strategic Research, a Wallenberg Academy Fellowship, the Faculty of Sciences at the University of Gothenburg, the Wenner-Gren Foundations, and the David Rockefeller Center for Latin American Studies at Harvard University) and through D.S. (from the Swedish Research Council [2015-04748]).

References

Bennett, D. J. (2018a). Restez: Create and query a local copy of genbank in r. Retrieved June 27, 2018, from https://doi.org/10.5281/zenodo.1299236

Bennett, D. J. (2018b). Restez: Create and query a local copy of genbank in r. Retrieved June 19, 2018, from https://github.com/ropensci/restez

Bennett, D. J. (2018c). Restez: Create and query a local copy of genbank in r. Retrieved November 27, 2018, from https://CRAN.R-project.org/package=restez

Bennett, D. J. (2018d). SupersmartR. Retrieved June 27, 2018, from https://github.com/AntonelliLab/supersmartR

Bennett, D., Hettling, H., Silvestro, D., Zizka, A., Bacon, C., Faurby, S., Vos, R., et al. (2018). phylotaR: An Automated Pipeline for Retrieving Orthologous DNA Sequences from GenBank in R. *Life*, *8*(2), 20. doi:10.3390/life8020020

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2013). GenBank. *Nucleic Acids Research*, *41*(D1). doi:10.1093/nar/gks1195

Ostell, J. (2002). The Entrez search and retrieval system. In *The ncbi handbook* (pp. 1–6). Retrieved from http://www.ncbi.nlm.nih.gov/books/NBK21081/

Winter, D. J. (2017). rentrez: An R package for the NCBI eUtils API. *The R journal*, *9*(2), 520–526. doi:10.7287/peerj.preprints.3179v2