The Journal of Open Source Software

Ripeline and Rmanual speed up biological research and reporting

Alexey Shipunov[1]

1 Minot State University, Minot ND

DOI: 10.21105/joss.02070

Software

- Review
- Repository
- Archive

Editor: Pending Editor

Submitted: 04 February 2020
Published: 06 February 2020

Summary

Ripeline (R-based sequence analysis pipeline)

Ripeline is an R-based pipeline application that allows for the wide variety of sequence analyses. In the core, it is a set of R and UNIX shell script templates (i.e., simple text files), which are extensively commented and easy to adapt for the needs of a particular project. I made the Ripeline when I realized that I must run four independent phylogenetic projects, each with plenty of marker data, and with only the restricted help from undergraduate students. It saved a significant amount of my time, which otherwise would have been spent on repetitive tasks.

Ripeline includes both R-based and standalone methods of analysis. While many of these are available separately, the actual workflows are diverse, not standardized, and therefore not fully reproducible. Ripeline allows for the fully standardized, reproducible sequence analyses. Installation of Ripeline is also completely portable, as it is equal to downloading the single directory. Each Ripeline script is numbered (like in the UNIX "init" startup system) and works independently (so if it fails for any reason, others will continue to work). It also outputs a full report of what was done, thanks to the "Rresults" utility from "shipunov" R package.

DNA-related data combined into "sets" (one per marker), aligned, trimmed, gap coded, and then concatenated in super-matrices. Two types of concatenation are employed: Strict concatenation Based on sequence IDs of locally obtained sequences. Semi-strict concatenation The next step. It uses "strict" dataset and adds sequences of any origin in order to fill all gaps. Ripeline uses some external tools like MrBayes, but can also work without them. Some proficiency in R is required to adopt these scripts for the needs of a particular project. However, nothing beyond small modifications (e.g., changing the size of the output image file) is necessary. Only rudimentary shell scripting knowledge is required, on the level of commenting or uncommenting particular lines. All in all, I believe that Ripeline is suitable for users without excessive programming skills.

As all its components are cross-platform, Ripeline works on all major platforms. Ripeline is fast. Template scripts are built around an artificial example which involves two DNA markers from 12 species, multiple alignment, trimming flanks, gap coding, concatenation, creation of "technical" k-mer and NJ trees, a maximum parsimony example, maximum likelihood (ML) hypotheses testing, two examples of ML, and Bayesian tree estimation. With parameters set to minimum (like 100 bootstrap replicates), the whole pipeline takes about 2 minutes on an average laptop.

Rmanual (R-based natural history manual)

Software that assists in producing the natural history manual should be able to extract data from databases, use images, and output typographically formatted text. To my knowledge, R and TeX are the best candidates for making this kind of software. Rmanual requires working TEX and R installations, plus one additional R package. All of these are available on all major platforms. Rmanual includes only one R script and also one shell script whose job is to use the former, and then run the TeX engine. R code is short, fully commented, and is easy to modify. There are only a few TeX definitions, which are easy to understand and modify.

The basic idea of the Rmanual is that after R outputs a text table (made from inter-combined input tables), TeX uses it as a "body" (main, taxonomic part) of the book. In the end, each page inside the PDF book consists of multiple table pieces.

The header and footer of the book (the title page, first and last pages) should be prepared manually using the supplied template. Apart from a PDF book, Rmanual also outputs diagnostic data (e.g., which species do not correspond with images).

The main strength of Rmanual is, therefore, the production of semi-automatic, typographically ready output from the set of "flat" text tables and images. Besides, Rmanual allows for the constant update of the output and therefore produces books that are not only semi-automatic but also "living." Rmanual, similarly to Ripeline, requires some knowledge in R and TeX but does not require extensive skills. If the researcher can modify R and TeX files (note again that they are simple text files), it should be enough to run Rmanual with their data and make simple PDF books.

Rmanual is very fast. When preparing this paper, I made a new illustrated checklist, which included 122 plant species and their images within just two hours. The actual book production (creation of a typographically ready PDF file) takes only a few seconds on an average laptop. Even if an inexperienced person will be ten times slower, 20 hours for a small book is a good result. Therefore, I believe that when all data is ready, with a help from Rmanual or with a Rmanual-like approach, the production of typographic results might

The illustrated checklist of North Dakota plants (http://ashipunov.info/shipunov/fnddb/) is one of the most advanced use cases of Rmanual.

Rmanual main example, two other working examples (which use real data) and Ripeline are presented on the Github: https://github.com/ashipunov/Rmanual , https://github.com/ashipunov/Rmanual_Jatun_Sacha , https://github.com/ashipunov/Rmanual_El_Yunque , https://github.com/ashipunov/Ripeline

Both Rmanual and Ripeline are licensed under GPL 3.0. They are parts of the bigger system which I develop to help field biologists in their work.

Acknowledgements

The author is grateful to the Vitalij Arnold, Sergej Sukhov, Vera Shipunova, and authors of R and TeX software.