Genome Sequence Data in R

Author: Robert Furrow, University of California, Davis, refurrow@ucdavis.edu.

Course Information

Department: **Biological Sciences (general lower division bio)**
Level: **Lower Undergraduate**
Course type: **Lab (computational)**
Students: **Non-majors**
Number of Students:**17**

Focus: Students will learn how to manipulate large strings in R using the Biostrings package, which will allow them to perform analyses on genomic data including calculating GC content and exploring genome assembly quality.

Overview: This lesson aims to familiarize students with DNAStringSets from the Biostrings package in the programming language R. This will build student skills to manipulate and analyze genomic sequence data. The lesson was designed as part of a course-based research experience in which students perform exploratory data analysis on genome assemblies of novel microbes. These assemblies can be loaded into R using tools from the Biostrings package, and the genomes can be analyzed easily both with custom functions and with many built-in tools in the package. Although we only introduce the basics of counting nucleotide frequencies and exploring contig lengths, many students went on to use additional tools in the Biostrings package for their individual projects.

Learning objectives: By the end of this lesson, students will be able to:

- manipulate large strings of genomic data in R as DNAStringSets, and
- perform calculations on these strings, including GC content and length.

Lesson sequence:

1. Working with Biostrings

Pre-lesson activities: Up to this point, students had already devoted about 3 hours of lab time to coding in R. They had learned about types of objects in R, including vectors and functions, and had manipulated some short DNA sequences stored as

strings. The swirl lesson is their first experience with the Biostrings package. However, they had already installed the Biostrings package. Students also already knew what a FASTA file looked like, as they had opened those files in a notepad to see the general structure of the information. We include an Rmd and pdf file of Lab 2 for our class, which took place before this homework (and includes the homework prompt related to the swirl lesson). The Rmd will not knit as-is, as there are other files used in the lab. But it demonstrates the material used to introduce this swirl lesson. Please contact Robert Furrow at refurrow@ucdavis.edu if you would like additional materials, and refer to our online course lab book for the full arc of the class.
https://bookdown.org/joelrome88/bis23b/

Post-lesson activities: In a brief post-lesson activity, we asked students to load the FASTA file of “their strain” and look at a few features of the assembly using width(). They build on this in Lab 3 of the class, in which they calculate GC content and genome assembly quality metrics like N50 and L50. Lab 3 is included on QUBES as an Rmd and pdf file.The Rmd will not knit as-is, as there are other files used in the lab. But it demonstrates the material used to introduce this swirl lesson. Please contact Robert Furrow at refurrow@ucdavis.edu if you would like additional materials, and refer to our online course lab book for the full arc of the class.
https://bookdown.org/joelrome88/bis23b/

Implementation notes: Biostrings is a package within Bioconductor, which means that it is slightly trickier to install than a typical R package, and the help files are not as well-written. This swirl lesson should automatically install Biostrings and load the data appropriately. ***However, test users have reported that the R console might print some warnings during this process. Despite these warnings the lesson will still work.*** The installation may also ask about updating other R packages; when prompted “Update all/some/none? [a/s/n]:” you should enter “a” by typing the letter into the console. This lesson draws upon two data files: an .RData file and a FASTA file. Installing the course by using install_course() and navigating to the .swc file you’ve downloaded should also properly save those two files into the correct folder in the swirl/Courses folder on your computer. In case that doesn’t work, those two files are also posted to QUBES, and can be downloaded directly to the lesson folder for the “Working_with_Biostrings” lesson within the “Genome_Sequence_Data_in_R” course.

To simplify the installation of the lesson, and of the swirl package, we used RStudio Cloud. This web-based R interface allows instructors to create an R environment in which key packages or data have already been installed and loaded. It is not necessary, but doing this allowed us to avoid troubleshooting installation issues with swirl or the lesson, as this course was taught online during the Spring quarter of 2020 due to COVID-19 shelter-in-place restrictions. Using a cloud-based R implementation also avoids issues with students who have chromebooks or are otherwise unable to install R locally.