

LibRillo de R y Biología

Cristian Román Palacios

2022-06-19

Contents

1	Una breve introduccion a R	11
1.1	R y RStudio	11
1.1.1	Instalando R y RStudio	11
1.2	Primeros pinitos en R	12
1.2.1	Exportar datos	13
1.3	R un poco mas avanzado...	13
1.3.1	for loops	13
1.3.2	Vectorizando	13
1.3.3	Funciones de la familia <code>*apply</code>	13
1.3.4	<code>data.table</code>	13
1.3.5	El tidyverso	13
1.3.6	Personalizando R	13
2	Paginas web con GitHub, RStudio, y R	15
2.1	Jekyll	15
2.2	Hugo	15
3	Usando GitHub, RStudio, y R	17
3.1	Git	17
3.2	GitHub	17
3.3	RStudio e integracion con GitHub	17
4	Regresiones lineales	19
5	Regresiones mas avanzadas	21
5.1	Regresiones Bayesianas	21

6 Ecologia basica	23
6.1 Indices de diversidad	23
6.2 Curvas de acumulacion	23
6.3 Rarefaccion	23
6.4 Usando datos del gbif	23
6.5 Modelos de distribucion	23
7 Filogenetica basica en R	25
7.1 Descargar secuencias de GenBank	25
7.2 Usar secuencias existentes	25
7.3 Alinear secuencias	25
7.4 Visualizar alineamientos	25
7.5 Correr arboles filogeneticos	25
7.6 IPS	25
7.7 Exportar datos desde R	25
8 Macroeolucion basica en R	27
8.1 Regresiones filogeneticas	27
8.2 Estimaciones de señal filogenetica	27
8.3 Modelos SSE	27
9 Web scraping con R	29
10 Machine learning con R	31
10.1 Keras en R	31
10.2 Tensorflow en R	31
10.3 Spark R	31
11 Data Mining con R	33
12 Tests estadisticos	35
13 Estadistica descriptiva en R	37
14 Libros, paquetes, apps en R	39
14.1 Libros en R	39
14.2 Paquetes en R	39
14.3 Shiny apps en R	39
15 Introduccion a RMarkdown	41

<i>CONTENTS</i>	5
16 Autocalificar tareas en R	43
16.1 SHA	43
16.2 Objetos invisibles en el espacio de trabajo	43
17 Graficos en R	45
17.1 Graficando con R base	45
17.2 Graficando con ggplot2 en R	45
18 Macroeolucion basica en R	47
18.1 Regresiones filogeneticas	47
18.2 Estimaciones de señal filogenetica	47
18.3 Modelos SSE	47
19 Manipulando datos en R	49
20 Bases de datos en R	51
20.1 SQL	51
21 Organizando proyectos reproducibles	53
21.1 RStudio projects	53
21.2 GitHub repo	53
22 Mejores Practicas R	55
23 Como pedir ayuda de forma eficiente?	57
23.1 Ejemplos reproducibles	57
23.2 Fuentes potenciales	57
24 Como hacer mapas en R	59
24.1 Base R	59
24.2 ggplot	59
25 Simulando datos	61
26 Reportes reproducibles	63

Prólogo

Este libro es una compilación para consulta rápida sobre diferentes aspectos que combinan el uso de R con diferentes temas de biología. El **LibRillo de R y Biología** pretende servir de referencia sobre temas particulares, enfocando capítulos sobre objetivos específicos y desarrollando brevemente el contenido principalmente alrededor de un paquete en R. Este librito cubre temas que engloban desde lo más fundamental en R así como aplicaciones que combinan R, `git`, `GitHub`, entre otros.

Estructura del libro

Este **LibRillo de R y Biología** pretende avanzar a quien lo consulte en temas específicos que relacionan diferentes ámbitos y aspectos biológicos con R. Sin embargo, el libro no se enfoca solamente en revisar la aplicación de conceptos usando técnicas. En el capítulo 1 XXX...

Convenciones generales del libro

Este libro fue generado usando principalmente `R knitr` y `bookdown`. En general, los bloques de código que se incluyen en el libro permiten una ejecución sencilla e intuitiva de los temas. En diferentes secciones del libro, se pretende generar salidas gráficas y numéricas directamente usando códigos en R. Al inicio de cada capítulo incluyo información sobre los paquetes que son relevantes para llevar a cabo los diferentes objetivos del mismo.

La información contenida en este libro fue compilada usando la siguiente sesión de R:

```
sessionInfo()

## R version 4.2.0 (2022-04-22)
## Platform: x86_64-apple-darwin17.0 (64-bit)
## Running under: macOS Catalina 10.15.7
##
## Matrix products: default
## BLAS: /Library/Frameworks/R.framework/Versions/4.2/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.2/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats      graphics  grDevices  utils      datasets  methods   base
##
## loaded via a namespace (and not attached):
```

```
## [1] compiler_4.2.0  magrittr_2.0.3  fastmap_1.1.0  bookdown_0.27
## [5] cli_3.3.0        htmltools_0.5.2  tools_4.2.0   rstudioapi_0.13
## [9] yaml_2.3.5       stringi_1.7.6    rmarkdown_2.14 knitr_1.39
## [13] stringr_1.4.0    digest_0.6.29    xfun_0.31     rlang_1.0.2
## [17] evaluate_0.15
```

Pretendo no agregar simbolos como `>` y `+` en el codigo de R dentro de este libro. En general, los codigos aparecen destacados en cajas grises:

```
a <- c(2, 4, 5)
```

Las salidas de codigo tambien aparecen en un bloque gris. Sin embargo, las lineas de salida incluyen los caracteres `##` antes del contenido de cada linea. Por ejemplo, multipliquemos el vector `a` creado anteriormente por 20:

```
## [1] 40 80 100
```

La salida de esta operacion necesariamente tiene los `##` en el bloque. Por lo tanto, unicamente el codigo que **no** tiene `##` debe ser copiado (en caso de ser necesario). Las salidas de codigo realmente no son tan relevantes para reproducir el codigo.

Por ultimo, el nombre de paquetes se indica en negrilla dentro del texto (e.g. **bookdown**), las funciones se presentan seguidas de parentesis (e.g. `render_book()`), los dobles dos puntos (`::`) significan que un determinado objeto pertenece a un paquete determinado (e.g. `bookdown::render_book()`), y codigos en el texto aparecen en formato de maquina de escribir (e.g. R).

Estas son las convenciones mas generales del libro. Otras convenciones mas particulares y relevantes a otros temas seran discutidas en los capitulos siguientes.

Dedicacion

Este libro esta escrito principalmente para estudiantes de universidades publicas o pertenecientes a grupos historicamente excluidos de la academia. Escribo el libro con mucho cariño, pensando en que ayudara a equilibrar la balanza entre quienes tienen acceso directo a recursos “avanzados” en ingles y aquellos que por diferentes razones, solamente acceden a informacion en español. El libro es de acceso abierto y no pretendo cobrar por su uso.

Agradecimientos

Mis agradecimientos mas fuertes van para mi padre y madre, Cesar Roman y Yocasta Palacios. Al parecer, de mi padre, Ictiologo y profesor titular en la Universidad del Quindio, termine heredando la profesion. De mi madre, profesora de preescolar en diferentes instituciones del Quindio (Colombia), la pasion y paciencia por enseñar y ayudar. Grandes agradecimientos van tambien a mi hermano, Carlos, ingeniero en sistemas, gran programador y conversador, quien por su cercania me mostro que escribir codigo (al menos al nivel basico que yo hago), no es imposible. Este libro lo empiezo a escribir estado en Tucson, AZ. Gracias a mi pareja, Heidi E. Steiner, quien me apoya con su actitud, tiempo, y cariño, mientras me dedico por ratos a escribir apartados de este libro. A estos cuatro personajes, motor de mi vida, todo mi cariño y admiracion.

Sobre el autor

Cristian Román Palacios es Profesor Asistente en la Universidad de Arizona desde el 2021. Cristian es Biologo de la Universidad del Valle (2015), Magister y Ph.D. en Ecología y Evolucion (2019, 2020) de la Universidad de Arizona, con cargos postdoctorales en Temple University (2021) y University of California, Los Angeles (2021). Los intereses investigativos de Cristian contemplan temas de evolucion a grandes escalas taxonomicas, asi como la relacion entre cambio climatico y diversidad. Sin embargo, el aprendizaje de areas relacionadas con programacion resulto ser una prioridad durante su formacion academica. Ahora pretende divulgar de forma sencilla, contenidos de R y biologia usando diferentes recursos, uno de estos son los libros. Actualmente, Cristian enseña cursos de Machine Learning y Data Mining en la Escuela de La Informacion, en la Universidad de Arizona. Mas informacion sobre el autor esta disponible en su pagina web.

Chapter 1

Una breve introduccion a R

R será el principal lenguaje que guiara el libro. En diferentes disciplinas biologicas, el uso de R se ha extendido significativamente durante la ultima decada. Sin embargo, R, por ser un lenguaje principalmente enfocado en aspectos estadisticos, tiene ciertas limitaciones en cuanto a velocidad y flexibilidad. Por otro lado, debido a la compartimentalizacion de implementaciones metodologicas en codigo, aprender y usar R es probablemente central en un campo biologico determinado pero no en otros. Otros recursos pueden ser usados para seguir la estructura del presente libro en otros lenguajes (e.g. `python`). Por ahora, nos enfocaremos en como instalar y usar R. Por simplicidad, vamos a procurar trabajar desde `RStudio`.

1.1 R y RStudio

Aunque no son equivalentes, en diferentes contextos, R y `RStudio` se usan erroneamente de forma intercambiable. R es explicitamente un lenguaje de programacion, desarrollado por decadas como parte del `R Core Team` y la `R Foundation for Statistical Computing`. En pocas palabras, R es un lenguaje en si mismo, que esta enfocado en computaciones estadisticas. `RStudio` requiere de R para funcionar. Especificamente, `RStudio` permite manejar (potencialmente) mas eficientemente y esteticamente el espacio de trabajo en R.

1.1.1 Instalando R y RStudio

Dependiendo del sistema operativo en consideracion, existen multiples maneras de instalar R. En MacOS, vamos a usar `homebrew` para instalar paquetes y demas software que necesitemos. Para instalar `homebrew`, pueden correr la siguiente linea en la linea de comandos:

```
/bin/bash -c "$(curl -fsSL https://raw.githubusercontent.com/Homebrew/install/HEAD/install.sh)"
```

Si estan usando linux, intentaremos hacer las instalaciones usando `apt install` para instalar programas desde la linea de comandos. En Windows, los ejecutables estan generalente disponibles para los programas que estaremos usando.

1.1.1.1 Instalando R

En linux, la instalacion de R normalente requiere de un par de comandos:

```
sudo apt update
sudo apt install r-base
```

En MacOS y desde `homebrew`, la instalacion de R involucra la siguiente linea:

```
brew install r
```

En Linux y MacOS, la instalacion debe crear una nueva aplicacion. Sin embargo, desde la misma linea de comandos, correr

```
R
```

debe permitir ingresar a R directamente. Por ultimo, es potencialmente preferible descargar e instalar R usando el instalador en la pagina oficial de R project usando este link.

1.1.1.2 Instalando RStudio

Instalar RStudio no es totalmente necesario para seguir las instrucciones en todos los capitulos de este librito. Sin embargo, es util para muchas de ellas. Por otro lado, RStudio no es completamente necesario para correr R. Para descargarlo e instalarlo pueden seguir las instrucciones en la pagina oficial de RStudio.

RStudio tambien puede ser instalado desde la linea de comando. En Ubuntu Linux, instalar RStudio toma las siguientes lineas:

```
sudo apt update
sudo apt install gdebi-core
wget https://cdn.rstudio.com/package-manager/ubuntu/amd64/rstudio-pm_2022.04.0-7_amd64.deb
sudo gdebi rstudio-pm_2022.04.0-7_amd64.deb
```

En MacOS, `homebrew` puede usarse para instalar RStudio muy facilmente:

```
brew install --cask rstudio
```

En Windows, la recomendacion es descargar la aplicacion directamente desde la pagina de RStudio en el siguiente enlace.

1.2 Primeros pinitos en R

Ahora que tenemos R instalado vamos a empezar a conocer algunos aspectos basicos del lenguaje!

1.2.1 Exportar datos

1.3 R un poco mas avanzado...

1.3.1 for loops

1.3.2 Vectorizando

1.3.3 Funciones de la familia *apply

1.3.4 data.table

1.3.5 El tidyverso

1.3.6 Personalizando R

Mensaje de inicio

Chapter 2

Paginas web con GitHub, RStudio, y R

2.1 Jekyll

2.2 Hugo

Chapter 3

Usando GitHub, RStudio, y R

3.1 Git

3.2 GitHub

3.3 RStudio e integracion con GitHub

Chapter 4

Regresiones lineales

Chapter 5

Regresiones mas avanzadas

5.1 Regresiones Bayesianas

Chapter 6

Ecologia basica

6.1 Indices de diversidad

6.2 Curvas de acumulacion

6.3 Rarefaccion

6.4 Usando datos del gbif

6.5 Modelos de distribucion

Chapter 7

Filogenetica basica en R

7.1 Descargar secuencias de GenBank

7.2 Usar secuencias existentes

7.3 Alinear secuencias

7.4 Visualizar alineamientos

7.5 Correr arboles filogeneticos

7.6 IPS

7.7 Exportar datos desde R

Chapter 8

Macroevolucion basica en R

8.1 Regresiones filogeneticas

8.2 Estimaciones de señal filogenetica

8.3 Modelos SSE

Chapter 9

Web scraping con R

Chapter 10

Machine learning con R

10.1 Keras en R

10.2 Tensorflow en R

10.3 Spark R

Chapter 11

Data Mining con R

Chapter 12

Tests estadísticos

Chapter 13

Estadística descriptiva en R

Chapter 14

Libros, paquetes, apps en R

14.1 Libros en R

14.2 Paquetes en R

14.3 Shiny apps en R

Chapter 15

Introduccion a RMarkdown

Chapter 16

Autocalificar tareas en R

16.1 SHA

16.2 Objetos invisibles en el espacio de trabajo

Chapter 17

Graficos en R

17.1 Graficando con R base

17.2 Graficando con ggplot2 en R

Chapter 18

Macroevolucion basica en R

18.1 Regresiones filogeneticas

18.2 Estimaciones de señal filogenetica

18.3 Modelos SSE

Chapter 19

Manipulando datos en R

Chapter 20

Bases de datos en R

20.1 SQL

Chapter 21

Organizando proyectos reproducibles

21.1 RStudio projects

21.2 GitHub repo

Chapter 22

Mejores Practicas R

Chapter 23

Como pedir ayuda de forma eficiente?

23.1 Ejemplos reproducibles

23.2 Fuentes potenciales

Chapter 24

Como hacer mapas en R

24.1 Base R

24.2 ggplot

Chapter 25

Simulando datos

Chapter 26

Reportes reproducibles