

R para Ciências da Vida (BCM13065) Aula 1

PPGBCM - UFRGS

**Diego Bonatto
2024/2**

Disciplina: R para Ciências da Vida

Período Letivo: 2024/2

Professor Responsável: DIEGO BONATTO

Sigla: BCM13065 **Créditos:** 4

Súmula da disciplina

Fundamentos de R: programas e interfaces; operadores, variáveis, funções e diretórios; pacotes para manipulação e criação de vetores, matrizes, tabelas e listas; tipos de dados biológicos e suas entradas; boas práticas de computação em R: ambiente, scripts e projetos; pacotes para análises de dados de larga escala (genômica, transcritômica e proteômica); pacotes para análises estatísticas; pacotes para a criação de gráficos e representação visual de dados.

Pré-requisitos: não há a necessidade de conhecimentos prévios em R ou qualquer outra linguagem de programação.

Objetivos

A disciplina tem como objetivo central proporcionar uma compreensão abrangente da programação em R, bem como as suas ferramentas e aplicações focadas para as Ciências da Vida. Para tanto, a disciplina busca familiarizar os discentes com a linguagem e a programação R para dados biológicos. Nesse sentido, o objetivo central da disciplina será alcançado: (I) pela compreensão dos diferentes componentes da programação em R, como operadores e funções, para análise de dados biológicos, (ii) pela escolha da forma de entrada de dados biológicos uni e multidimensionais, (iii) pelo processamento dos dados biológicos e uso de pacotes específicos para a descoberta de conhecimento biológico, (iv) pelo emprego específico de pacotes gráficos para a representação visual dos resultados gerados e (v) pelo emprego de boas práticas de programação.

Conteúdo programático e cronograma

Aula (Data)	Título	Conteúdo
1 (11/11/2024)	Introdução à disciplina; Histórico do R e a sua importância para a análise de dados biológicos	Programas para análises estatísticas de dados biológicos; o programa “S”. O desenvolvimento do R como plataforma para a resolução e análise de dados biológicos. Por que usar o R? Comparação com outras linguagens de programação (Python, C/C++).
2 (12/11/2024)	Procedimentos iniciais com o R: interfaces e atualizações	Escolha de interfaces para o R: IDEs e GUIs. Procedimentos para a instalação de pacotes e a sua manutenção. Resolvendo problemas de dependências de pacotes (terminal ou console do R).
3 (13/11/2024)	Boas práticas de programação em R	Uso de comentários e registro de linhas de comandos/scripts. Criação de arquivos do tipo R script, Notebook e Markdown.
4 (18/11/2024)	Pacotes do R para a análise de dados biológicos	O sistema de arquivos CRAN. Bioconductor e GitHub como fontes de pacotes para análises biológicas. Cuidados específicos para instalação de pacotes no R (dependências e atualizações). Pacotes usados para análises de dados de larga escala, estatísticos e taxonomia/filogenia.
5 (19/11/2024)	Objetos do R	O que são vetores, fatores, matrizes, arranjos, tabelas e listas. A importância de cada objeto para análises biológicas.
6 (25/11/2024)	Mineração e prospecção de bancos de dados biológicos	Tipos de pacotes para data mining e knowledge discovery. Integração de ferramentas e pipelines para busca de dados biológicos.
7 (26/11/2024)	Processamento de dados biológicos – partes 1 e 2	Ferramentas do R para a manipulação de tabelas e listas. Remoção de dados e filtragem. Fusão de dados multidimensionais e a sua representação. Uso do “tidyverse” para o processamento de dados. Dplyr, Plyr, Magrittr, data.table e tidyr. Operadores e “pipes” para otimização dos dados biológicos.
8 (27/11/2024)	Gráficos e visualização de dados – partes 1 e 2	Funções nativas do R para a representação de dados biológicos (gráficos de barras, histogramas, dispersão de pontos, boxplots). O pacote ggplot2 e as suas aplicações gráficas. Otimização de cores para representação gráfica de dados.
9 (02/12/2024)	Interfaces com outras linguagens e plataformas	Interfaces com Python. Uso do Latex para a geração de documentos.

GitHub

The screenshot shows the GitHub interface for the repository 'R-para-Ciencias-da-Vida' by user 'bonattod'. The repository is public and has 2 branches and 1 tag. The main branch is selected. The repository description is 'Repositório de scripts e exercícios para a disciplina "R para Ciências da Vida"'. The repository has 0 stars, 1 watching, and 0 forks. The repository is licensed under CC0-1.0. The repository contains 12 files and folders, including 'Exercicios', 'RScripts', 'AllGene_ChromosomalFeatures_with_descri...', 'Conteudo_programatico_e_cronograma_BC...', 'LICENSE', 'ORF_Genes_names_processed_update_201...', 'ORF_dd_text_9423_all_partial.txt', 'README.md', 'base-r-cheat-sheet.pdf', 'data_table_cheat_sheet.pdf', 'dplyr_cheat_sheet.pdf', and 'ggplot2_cheat_sheet.pdf'. The repository also has a 'Releases' section with 1 release, 'RCV_1.0', which is the latest version. The repository also has a 'Packages' section with no packages published. The repository also has a 'Languages' section showing R at 100.0%.

bonattod / R-para-Ciencias-da-Vida

<> Code Issues Pull requests Discussions Actions Projects Wiki Security Insights Settings

R-para-Ciencias-da-Vida Public

Unpin Unwatch 1 Fork Star 0

main 2 Branches 1 Tag Go to file Add file <> Code About

bonattod Update README.md fdf5ed2 · 5 days ago 96 Commits

Exercicios	Rename Exercicio_9_R.pdf to Exercício_9_R.pdf	5 days ago
RScripts	Add files via upload	5 days ago
AllGene_ChromosomalFeatures_with_descri...	Add files via upload	2 weeks ago
Conteudo_programatico_e_cronograma_BC...	Add files via upload	5 days ago
LICENSE	Initial commit	3 weeks ago
ORF_Genes_names_processed_update_201...	Add files via upload	2 weeks ago
ORF_dd_text_9423_all_partial.txt	Add files via upload	2 weeks ago
README.md	Update README.md	5 days ago
base-r-cheat-sheet.pdf	Add files via upload	2 weeks ago
data_table_cheat_sheet.pdf	Add files via upload	5 days ago
dplyr_cheat_sheet.pdf	Add files via upload	5 days ago
ggplot2_cheat_sheet.pdf	Add files via upload	5 days ago

README CC0-1.0 license

Súmula da disciplina

About

Repositório de scripts e exercícios para a disciplina "R para Ciências da Vida"

Readme

CC0-1.0 license

Activity

0 stars

1 watching

0 forks

Releases 1

RCV_1.0 Latest 5 days ago

Packages

No packages published

[Publish your first package](#)

Languages

R 100.0%

<https://github.com/bonattod/R-para-Ciencias-da-Vida>

Google Drive

Q Search in Drive

✓ ? ⚙ ⋮

My Drive > R_Ciências_Vida_Discipli...

✓ ≡



⋮

i

Type ▾

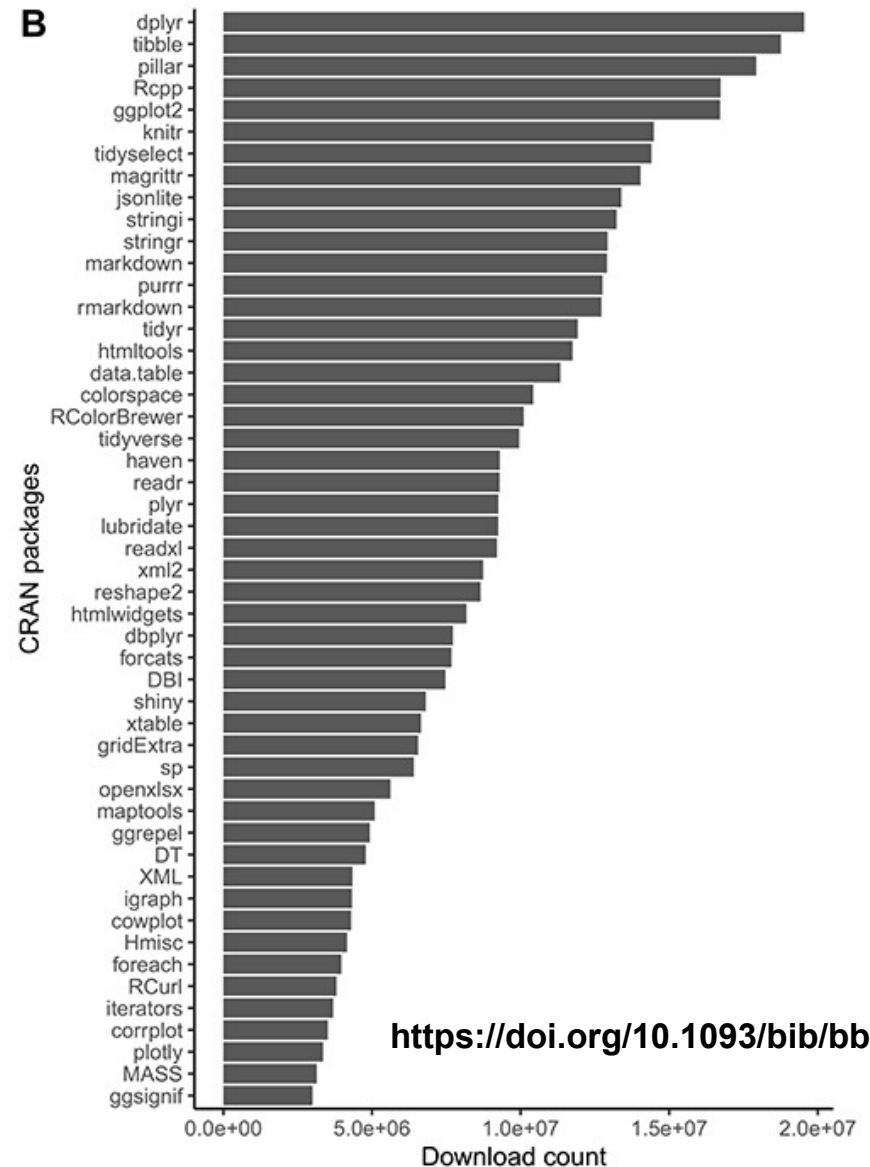
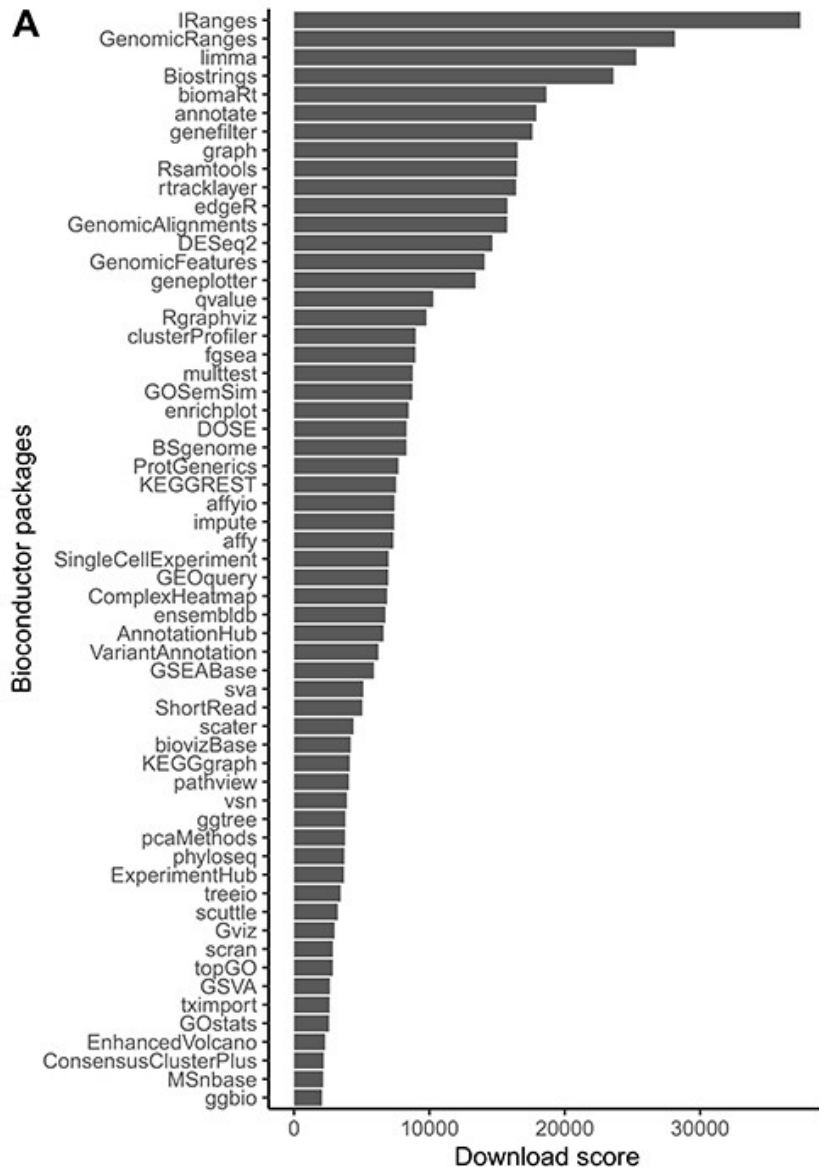
People ▾

Modified ▾

Name ▾	Owner	Last mo... ▾	File size	⋮
<div>📁 Exercícios</div>	<div> me</div>	Oct 20, 2024 me	—	⋮
<div>📁 Discentes</div>	<div> me</div>	Oct 20, 2024 me	—	⋮

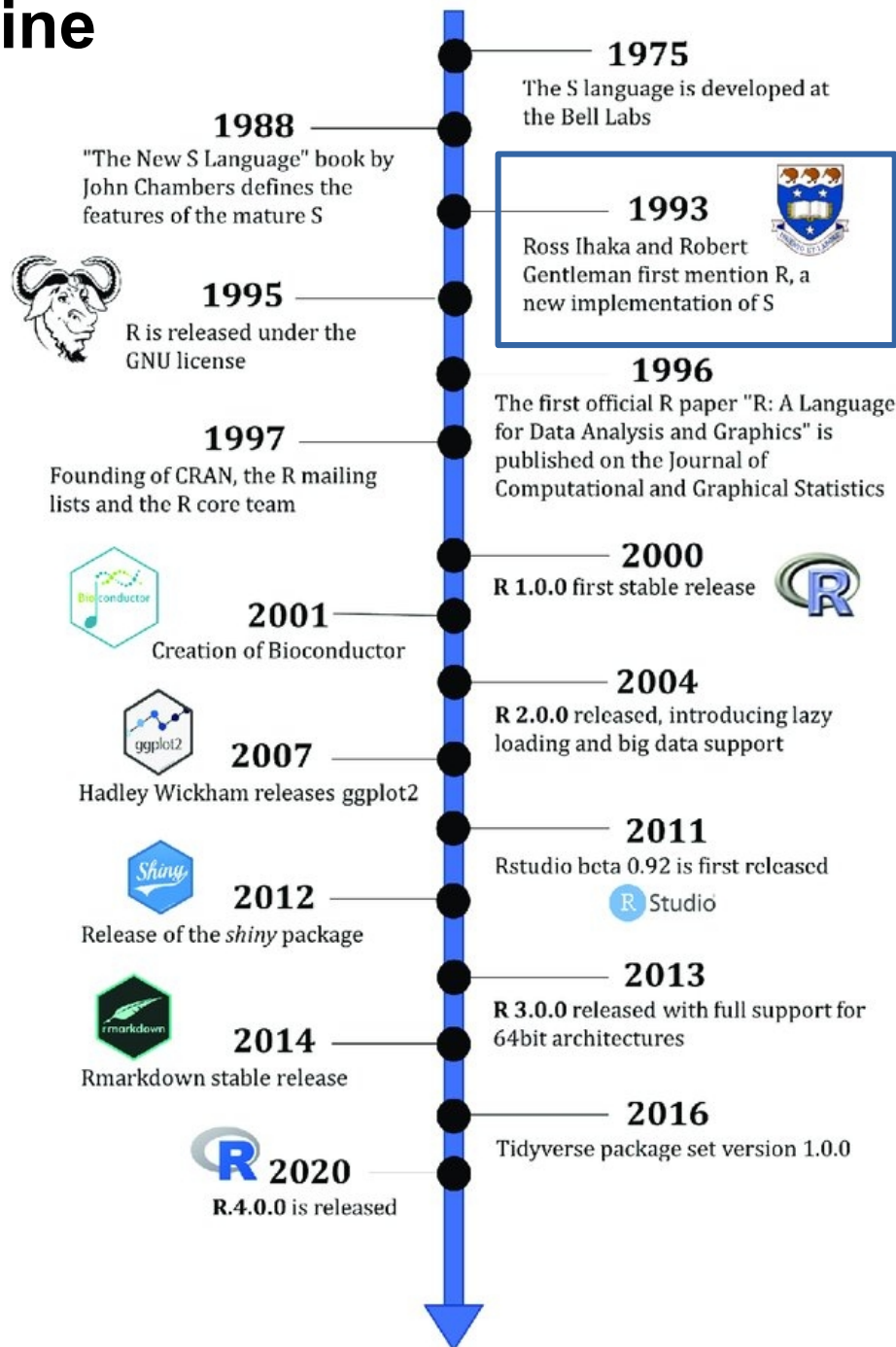
<https://drive.google.com/drive/folders/19eYJ6qq459pR3G500TfAq42v0T4sMDNQ?usp=sharing>

Introdução ao R e a sua importância para análise de dados biológicos



<https://doi.org/10.1093/bib/bbab415>

R timeline



Histórico: Programa 'S' e Desenvolvimento do R

We're continuing to fight for universal access to quality information—and you can help as we continue to make improvements. Will you chip in?

INTERNET ARCHIVE
Wayback Machine

http://ect.bell-labs.com/sl/S/

70 captures

26 Jun 2015 - 14 Oct 2018

Go

SEP

OCT

NOV

14

2017

2018

2019

About this capture



The S System

[Overview](#)

[The S project](#)

[History](#)

[S and S-Plus](#)

The S System

S is a language and system for organizing, visualizing, and analyzing data. It has been a [project](#) of statistics research at Bell Labs since 1976, [evolving continually](#) through that time. In 1998, S became the first statistical system to receive the [Software System Award](#), the top software award from the ACM.

This page is a brief author's-eye view of the system, with pointers to other sources of information.

S has from the start been aimed at *programming with data*; that is, at describing to the computer some graphical view, numerical summary, statistical model, or other information you want to produce. It occupies a middle ground between packages that emphasize standard operations and research projects in language design that start from a more abstract goal. S has always been designed to be used in practice, but with an emphasis on users who wanted to turn new ideas into software.

Although S was invented at Bell Labs, and we continue to be involved very much in its evolution, the implementations actually available, S-Plus and R, are distinct from the S language itself.

S-Plus products are distributed by the MathSoft Corporation. In particular, the S-Plus language is based on the S software from Bell Labs; MathSoft has an exclusive license with Lucent Technologies to distribute software based on S from Bell Labs. For more background on S and S-Plus, [click here](#).

The [R language](#), is an open-source system distributed under the GPL license, which is sometimes described as a "free clone" of S. More accurately, it is a separate project, based on the S language, but with a number of additional software directions.

Finally, we should mention the [Omegahat](#) software. Like R, it is a joint, open-source project for statistical computing. In part, it is concerned with next-generation software. However, getting there from the current generation software is also part of Omegahat. In particular, there are a number of inter-system interfaces from the S language (S-Plus or R), and some tools for programming in the S language.

More Information

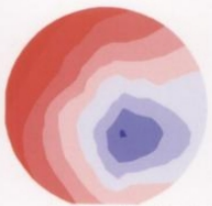
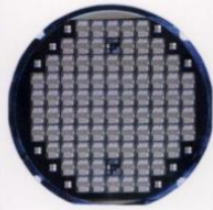
There exist many books and articles on S (plus countless articles that include S-based graphics), plus information from the web or e-mail. Here are a few pointers that may be helpful.


A linguagem S

John M. Chambers

PROGRAMMING WITH DATA

A Guide to the S Language



 Springer

Copyrighted material

Preface

S is a programming language and environment for all kinds of computing involving data. It has a simple goal:

To turn ideas into software, quickly and faithfully

Any application involving data is suitable for S, particularly if some of your ideas involve the structure and meaning of the data.

You use S interactively, giving it tasks, looking at data, and creating objects that describe your projects. S can be, and is, used in a “non-programming” style, exploiting quick interaction and graphics to look at data. This use often leads to a desire to customize what you are doing, and S encourages you to slide into programming, perhaps without noticing.

Por que usar o R?

- 1. Forte Suporte para Estatística e Análise de Dados**
- 2. Visualizações Poderosas e Customizáveis**
- 3. Ambiente Aberto e de Código Livre**
- 4. Grande Quantidade de Pacotes**
- 5. Comunidade Ativa e Colaborativa**
- 6. Integração com Outros Softwares e Linguagens**
- 7. Reprodutibilidade e Compartilhamento de Análises**
- 8. Facilidade de Automação e Manipulação de Dados**
- 9. Usado em Diversos Setores e por Grandes Empresas**

Comparação: R vs Python

Parameter	R	Python
Objective	Data Analysis and Statistical Modeling	Data Science, Web Development, Embedded Systems
Workability	Consists of many easy to use packages	Can easily perform matrix computation as well as optimization
Integration	Locally Run Programs	Programs integrated with web-app for easy deployment
Database Handling Capacity	Poses problem for handling large dataset	Can handle large data easily without any fault
IDE	Rstudio, R GUI	Spyder, IPython, Jupyter Notebook
Essential Packages and library	ggplot2, tidyverse, caret	Numpy, pandas, scipy, scikit-learn, TensorFlow

Comparison between R Programming and Python



<https://data-flair.training/blogs/r-vs-python/>

Outros programas/linguagens

[Top](#) [Packages](#) [Community](#)



JuliaStats

Statistics and **Machine Learning** made easy in Julia.

- Easy to use tools for statistics and machine learning.
- Extensible and reusable models and algorithms
- Efficient and scalable implementation
- Community driven, and open source

[Learn more](#)