Introduction to Scientific Computing for Biologists

ISCB20.09 - Introduction to R

Md. Jubayer Hossain
https://jhossain.me/
jubayer@hdrobd.org

Founder
Health Data Research Organization
Lead Organizer
Scientific Computing for Biologists

01 February 2021

Section-1: Introduction to R

What is R

- R is a dialect of S(R is an implementation of the S programming language).

What is R

- R is a dialect of S(R is an implementation of the S programming language).
- R was created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand, and is developed by the R Development Core Team.

What is R

- R is a dialect of S(R is an implementation of the S programming language).
- R was created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand, and is developed by the R Development Core Team.
- R is a programming language and environment commonly used in statistical computing, data analytics and scientific research.

What is R

- R is a dialect of S(R is an implementation of the S programming language).
- R was created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand, and is developed by the R Development Core Team.
- R is a programming language and environment commonly used in statistical computing, data analytics and scientific research.
- R is a programming language and free software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing.

What is R

- R is a dialect of S(R is an implementation of the S programming language).
- R was created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand, and is developed by the R Development Core Team.
- R is a programming language and environment commonly used in statistical computing, data analytics and scientific research.
- R is a programming language and free software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing.
- The R language is widely used among statisticians and data miners for developing statistical software and data analysis.

Why R?

- R is open source and free!

Why R?

- R is open source and free!
    - R is free to download as it is licensed under the terms of the GNU General Public License.
    - You can look at the source to see what's happening under the hood.
    - There's more, most R packages are available under the same license so you can use them, even in commercial applications without having to call your lawyer.
- R is popular – and increasing in popularity.

Why R?

- R is open source and free!
  - R is free to download as it is licensed under the terms of the GNU General Public License.
  - You can look at the source to see what's happening under the hood.
  - There's more, most R packages are available under the same license so you can use them, even in commercial applications without having to call your lawyer.
- R is popular – and increasing in popularity.
- R runs on all platforms.(Windows, Linux and Mac)

Why R?

- R is open source and free!
  - R is free to download as it is licensed under the terms of the GNU General Public License.
  - You can look at the source to see what's happening under the hood.
  - There's more, most R packages are available under the same license so you can use them, even in commercial applications without having to call your lawyer.
- R is popular – and increasing in popularity.
- R runs on all platforms.(Windows, Linux and Mac)
- R is being used by the biggest tech giants(google, facebook, microsoft, twitter)

Applications of R

- Data Science
- Data Analysis
- Genomic Data Science
- Biological Data Analysis
- Mutational Signature Analysis
- Genomic Analysis
- Statistical Computing
- Machine Learning

R Packages for Data Analysis/Data Science

- `dplyr`
  - dplyr is a grammar of data manipulation, providing a consistent set of verbs that help you solve the most common data manipulation challenges
  - Documentation- https://dplyr.tidyverse.org/
- `ggplot2`
  - for static data visualization
  - https://ggplot2.tidyverse.org/
- Plotly
  - for interactive data visualization
  - https://plotly.com/r/
- tidyverse
  - combination of `dplyr`, `ggplot2`
  - https://www.tidyverse.org/

R Packages for Bioinformatics/Genomic Data Science

- Bioconductor
  - for genomic data analysis
  - https://www.bioconductor.org/
- seqinr
  - DNA or Protein sequence analysis
  - https://cran.r-project.org/web/packages/seqinr/index.html
- MutatioanlPattern
  - mutational signature analysis
  - https://bioconductor.org/packages/release/bioc/html/MutationalPatterns.html

Resources: Books

- R for Data Science by Roger D.Peng
- Introduction to Data Science by Rafael Irizarry
- Data Analysis for the Life Sciences by Rafael Irizarry
- Statistics using R
- R for Biologists
- R for Public Health
- Rmarkdown

Resources: Blogs

- https://www.datamentor.io/r-programming/
- https://online.stat.psu.edu/stat484/
- https://online.stat.psu.edu/stat485/
- https://www.statmethods.net/index.html
- https://simplystatistics.org/
- https://www.tutorialspoint.com/r/index.htm
- https://www.rforbiologists.org/
- https://compgenomr.github.io/book/
- https://statsandr.com/
- https://rafalab.github.io/pages/harvardx.html
- https://bolt.mph.ufl.edu/software/r-phc-6055/

Alternatives to R Programming

- Python
  - Python is a very powerful high-level, object-oriented programming language with an easy-to-use and simple syntax.
  - Python is extremely popular among data scientists and researchers. Most of the packages in R have equivalent libraries in Python as well.
- SAS (Statistical Analysis System)
  - SAS is a powerful software that has been the first choice of private enterprise for their analytics needs for a long time.
- SPSS – Software Package for Statistical Analysis
  - SPSS is another popular statistical tool. It is used most commonly in the social sciences and is considered the easiest to learn among enterprise statistical tools.