Análise Bioinformática em R

Aplicações e exemplos reais

Jessica Magno
João Muzzi
Lana Peters

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

27.09.2024

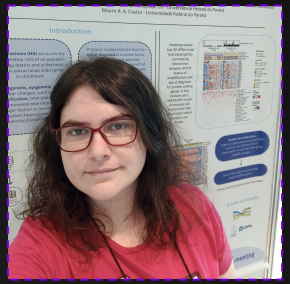
Quem somos nós?

Jessica Magno

Mestre em Bioinformática (UFPR, 2021) e doutoranda em Bioinformática (UFPR). Atualmente é pesquisadora técnica de Bioinformática no IPPPP.

João Muzzi

Doutor em Patologia, Microbiologia e Parasitologia (UFPR, 2022) e especialista em Statistics and Data Science (MIT, 2023). Atualmente é pós-doutorando no IPPPP e Senior Data Scientist na Grand Hill.



Lana Peters

Mestre em Engenharia de Bioprocessos e Biotecnologia (UFPR, 2022). Atualmente é doutoranda em Bioinformática (UFPR).

Cronograma do minicurso

01

Introdução

Revisão de conceitos importantes em R; pacotes que serão utilizados no decorrer do minicurso

02

Manipulação de dados

Exemplos práticos de análise de dados de expressão gênica

03

Análises estatísticas e Visualização

Exemplos práticos de análises estatísticas e plots que podem ser gerados

04

Encerramento

Dúvidas e finalização do minicurso

01

Introdução

Conceitos importantes em R e pacotes na área de Bioinformática

Tipos de dados

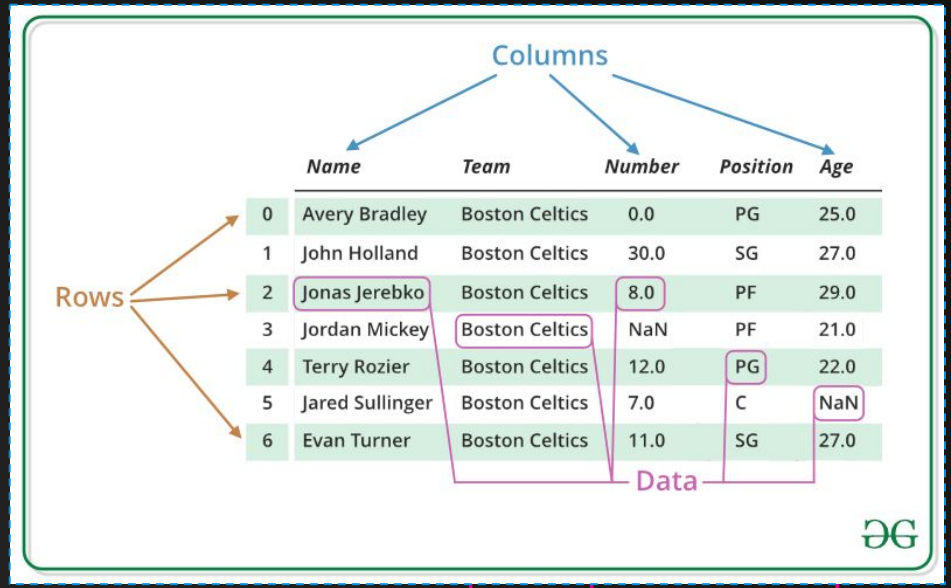
- **Data frames**
- **Matrizes**
- **Listas**

Exemplos de classes mais específicas:
SummarizedExperiment, **DESeqDataSet** (RNA-seq)

Data frames em R

Os data frames em R são objetos utilizados para **armazenar dados tabulares**.

Podem ser vistos como matrizes, **onde cada coluna pode ter tipos de dados diferentes**. Um data frame em R é composto por três componentes principais: **os dados, as linhas e as colunas**.



FONTE: https://www.geeksforgeeks.org/r-data-frames/

Data frames em R

```r
# R program to create dataframe

# creating a data frame
friend.data <- data.frame(
    friend_id = c(1:5),
    friend_name = c("Sachin", "Sourav",
                    "Dravid", "Sehwag",
                    "Dhoni"),
    stringsAsFactors = FALSE
)
# print the data frame
print(friend.data)
```

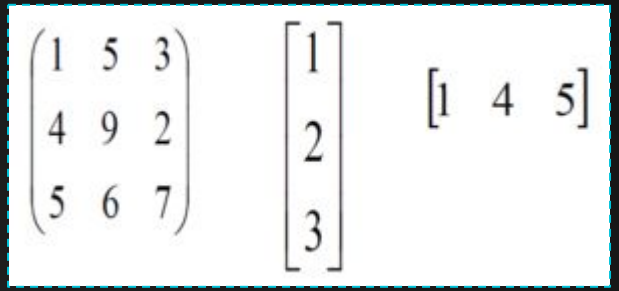
```
  friend_id friend_name
1         1      Sachin
2         2      Sourav
3         3      Dravid
4         4      Sehwag
5         5       Dhoni
```

FONTE: https://www.geeksforgeeks.org/r-data-frames/

Matrizes em R

A matriz em R é um arranjo bidimensional de **dados em linhas e colunas**.

Em uma matriz, as linhas são as que correm horizontalmente e as colunas são as que correm verticalmente. Na programação em R, **as matrizes são estruturas de dados bidimensionais e homogêneas**.



$$\begin{pmatrix} 1 & 5 & 3 \\ 4 & 9 & 2 \\ 5 & 6 & 7 \end{pmatrix} \qquad \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \qquad [1 \quad 4 \quad 5]$$

matrix(data, nrow, ncol, byrow, dimnames)

FONTE: https://www.geeksforgeeks.org/r-matrices/

Matrizes em R

```r
# R program to create a matrix

A = matrix(

  # Taking sequence of elements
  c(1, 2, 3, 4, 5, 6, 7, 8, 9),

  # No of rows
  nrow = 3,

  # No of columns
  ncol = 3,

  # By default matrices are in column-wise order
  # So this parameter decides how to arrange the matrix
  byrow = TRUE
)

# Naming rows
rownames(A) = c("a", "b", "c")

# Naming columns
colnames(A) = c("c", "d", "e")

cat("The 3x3 matrix:\n")
print(A)
```

```
The 3x3 matrix:
  c d e
a 1 2 3
b 4 5 6
c 7 8 9
```

FONTE: https://www.geeksforgeeks.org/r-matrices/

Listas em R

Uma lista na programação em R é um objeto genérico que consiste em uma **coleção ordenada de objetos**. As listas são estruturas de dados **unidimensionais e heterogêneas**.

A lista pode ser uma lista de vetores, uma lista de matrizes, uma lista de caracteres, uma lista de funções, e assim por diante.

Uma lista é um vetor, mas com elementos de **dados heterogêneos**. Uma lista em R é criada com o uso da função **list()**.

O R permite acessar elementos de uma lista usando o valor do índice. **Em R, a indexação de uma lista começa em 1 em vez de 0**.

FONTE: https://www.geeksforgeeks.org/r-lists/

Listas em R

```r
# R program to create a List

# The first attributes is a numeric vector
# containing the employee IDs which is created
# using the command here
empId = c(1, 2, 3, 4)

# The second attribute is the employee name
# which is created using this line of code here
# which is the character vector
empName = c("Debi", "Sandeep", "Subham", "Shiba")

# The third attribute is the number of employees
# which is a single numeric variable.
numberOfEmp = 4

# We can combine all these three different
# data types into a list
# containing the details of employees
# which can be done using a list command
empList = list(empId, empName, numberOfEmp)

print(empList)
```

```
[[1]]
[1] 1 2 3 4

[[2]]
[1] "Debi"    "Sandeep" "Subham"  "Shiba"

[[3]]
[1] 4
```

FONTE: https://www.geeksforgeeks.org/r-lists/

Pacotes R - Bioinformática

Coleção de pacotes R projetados para análise e compreensão de dados genômicos

Gramática de manipulação de dados

Sistema para criação de gráficos, baseado em The Grammar of Graphics

Facilita a análise diferencial de expressão gênica com base em dados de contagem, comumente usados para análise de RNA-Seq

Pacote popular para análise de expressão diferencial de RNA-Seq e outros dados de contagem

E muitos outros...

Outras recomendações

- SummarizedExperiment
- GenomicRanges
- Biostrings
- limma
- clusterProfiler
- ComplexHeatmap
- Gviz
- GenomicFeatures
- ...

02

Manipulação de dados

Exemplos práticos de análise de dados de expressão gênica

https://github.com/jmuzzi/sbib2024

Materiais do minicurso

Obrigado!

Nossos contatos:

Jessica Magno (jessicamagno@outlok.com)

João Muzzi (joao.muzzi@gmail.com)

Lana Peters (lanabpq@gmail.com)

CREDITS: This presentation template was created by **Slidesgo,** and includes icons by **Flaticon**, and infographics & images by **Freepik**