Parte 1/4: Introdução à Análise Comparativa de Sequências Genômicas

Gepoliano Chaves

Outubro, 2020

Contents

1) Introdução

1.1) Tabela de Conteudos do Mini-Curso

1.1.1) Resultados Esperados

Entre os objetivos deste curso, buscamos a familiarização dos alunos com análises computacionais e formação de grupo para desenvolvimento de projetos de pesquisa que visem estabelecer colaboração entre a Universidade de Chicago e a UFSB. Pretende-se que os estudantes entreguem ao final, um relatório com figuras geradas a partir da analise proposta do genoma viral. As figuras deverão explicitar a incidência de diferentes variantes genéticas de SARS-CoV-2 ao redor do globo.

1.1.2) Programação do Curso

Neste curso, pretende-se introduzir a noção de analise da informação contida na sequencia do RNA genomico viral, humano e outros, a partir da analise computacional. Para tal, iniciamos o ensino da analise de identificação de variantes em sequencias geneticas virais, disponíveis em plataformas publicas como GISAID e GEO. O codigo abaixo visa construir uma tabela de conteudo do mini-curso. Ele requer a instalação da livraria pander, como na linha comentada. Antes da instalação do pacote pander, o codigo nao pode ser executado como esperado.

A Tabela foi construida com base no seguinte endereco:

https://stackoverflow.com/questions/19997242/simple-manual-rmarkdown-tables-that-look-good-in-html-pdf-and-docx .

Table 1: Conteudo do Curso

Atividade	Descrição	Encontro
1	Apresentação de conteúdo e Atividades Computacionais	02/02/2021
2	Descrição das quatro Atividades Computacionais	04/02/2021
3	Virus, Procariotos e Eucariotos	09/02/2021
4	Tipos de Sequenciamento de DNA: Sanger, NGS, Nanopore	11/02/2021
5	Introdução a Programação Computacional	16/02/2021
6	R, Python e Linux: Comandos comumente usados	18/02/2021
7	Identificação de Variantes Geneticas ou Polimorfismos de DNA	23/02/2021
8	GATK, the Genome Analysis Tool-Kit	25/02/2021
9	Gene Expression Omnibus e Samtools	02/03/2021
10	Arvores filogenéticas e parentesco evolutivo entre estirpes	04/03/2021
11	Estudo Longitudinal de Saude do Adulto (Projeto ELSA/Brasil)	09/03/2021

1.2) Objetivos do Modulo

Neste módulo, introduzimos o conceito de programação nas Ciencias Genomicas. Utilizando o R, instalamos livrarias com pacotes para analise de sequencias de DNA (ou RNA). SARS-CoV-2 é um virus de RNA, porém ai analisar a sequencia do mesmo, a trataremos como DNA. Introduzimos também o primeiro de muitos tipos de arquivos usados em diferentes topicos de Biologia Computacional. As sequencias FASTA serao disponibilizadas pelo instrutor, mas podem também ser acessadas no site da base de dados GISAID, para os arquivos FASTA de SARS-CoV-2. Direcionamento para o site desta base de dados sera oferecido durante este curso.

O objetivo deste módulo é instruir o estudante no inicio da analise de sequencias de DNA. Este material baseia-se em tutorial original publicado por Thibaut Jombart, do Imperial College London. Aquele tutorial, em inglês, inicialmente apresenta a Genetica comparativa de sequencias de virus da gripe isoladas nos EUA. Por meio desta introdução, pretende-se que o aluno se familiarize com a ideia de sequencias FASTA e como plotar arvores filogeneticas comparando a identidade das mesmas. O arquivo PDF original de Thibaut Jombart chama-se "MRC-session2-tuto.1.3.pdf" e pode ser encontrado no seguinte link:

- http://adegenet.r-forge.r-project.org/files/MRC-session2-tuto.1.3.pdf

2) Programação

O principio da Programação de Computadores baseia-se na necessidade de analisar quantidades muito grandes de dados as quais nao podem ser analisadas senao com o auxilio de tais maquinas de calculos. Este pricipio tambem foi introduzido nas Ciencias Biologicas onde a necessidade de testes estatisticos e matematicos

conferiram a estas Ciencias, a acuracea e reprodutibilidade necessarias as sociedades industruais de producao em massa, notadamente as sociedades centrais do chamado Primeiro Mundo, principalmente os Estados Unidos da America.

Principalmente devido aos esforcos que levaram ao termino do Projeto Genoma Humano no comeco deste seculo, a quantidade de dados gerados a partir do sequenciamento dos genomas de plantas, animais e humanos tornou-se absurdamente grande com o uso de tecnicas chamadas de *high throughput sequencing* (HTS) ou *next generation sequencing* (NGS). Neste sentido, apenas com o auxilio de programas computacionais, a natureza e a diversidade das sequencias geneticas de plantas, animais e seres humanos podem ser conhecidas e apreciadas de modo a trazer o conhecimento necessario ao melhoramento da qualidade de vida das sociedades e mitigacao de modificaçoes ambientais desnecessarias.

Este curso apenas introduz conceitos basicos de Programação Computacional como o uso de codigo para execucao de um comando, a instalação de livrarias, o uso de linguagens como R e Python, e a conversão do arquivo texto original escrito no R Markdown, convertido em *pdf* nos notebooks usados neste curso. Voce devera receber os *pdf*s bem como o arquivo de texto original escrito usando o R Markdown e comparar um ao outro afim de familiarizar-se com a ideia de codigo e execucao do comando. Podemos pensar no arquivo de R Markdown como o codigo, e o *pdf*, como a execucao dos comandos contidos no R Markdown.

3) Instalação de Livrarias

3.1) Comando Geral

É comum a necessidade de instalação de livrarias ou pacotes contendo programas no R. Estas livrarias ou pacotes contêm os passos necessarios para fazer uma análise, disponibilizados por outra pessoa, comumente os autores intelectuais dos pacotes. A partir da utilizacao destes pacotes, diferentes tipos de análise podem ser feitos e replicados por grupos independentes utilizando diversos modelos e dados, biologicos ou nao. O tutorial abaixo apresenta a maneira mais comum de se instalar uma livraria no R. Esta maneira é usando o comando install.packages(), como ilustrado abaixo.

https://www.datacamp.com/community/tutorials/r-packages-guide

Como pode ser visto, no tutorial, o comando abaixo permite a instalacao da livraria de interesse.

```
## Comando geral para instatacao de pacotes no R
install.packages("package")
## Instalacao de livrarias especificas para atividade de hoje
install.packages("stats")
install.packages("ade4")
install.packages("ape")
install.packages("adegenet")
install.packages("phangorn")
```

3.2) Livrarias especificas para a atividade de hoje

Aqui, nosso escopo encontra-se delimitado pela area da Biologia Computacional, Bioinformatica ou Bioquimica. As livrarias especificas abaixo, sao requisitadas com base no tutorial de Jombart.

```
library(stats)
library(ade4)
library(ape)
library(adegenet)
library(phangorn)
```

4) Explorando o formato do arquivo FASTA

Para estudo, as sequencias FASTA utilizadas originalmente em analise de genoma de Influenza disponibilizadas por Jombart, necessitam ser baixaadas. Dois arquivos contêm as as informações necessárias para a avaliação: usaflu.fasta, contendo as sequencias FASTA do virus da gripe, e usflu.annot.csv, contendo o nome da sequencia e o descritor da localização geografica de isolamento da sequência. O comando "head" existe no R e permite a visualizacao de determinado numero de linhas do arquivo de interesse. O comando abaixo permite a visualizacao das primeiras 30 linhas do arquivo FASTA:

```
head -n 30 ~/Desktop/Gepoliano/UFSB/usaflu.fasta
```

```
## > CY013200
## atgaagactatcattgctttgagctacattttatgtctggttttcgctcaaaaacttccc
## ggaaatgacaacagcacagcaacgctgtgcctgggacaccatgcagtgccaaacggaacg
## ctagtgaaaacaatcacgaatgatcaaattgaagtgactaatgctactgagctggttcag
## agttcctcagcaggtagaatatgcgacagtcctcaccgaatccttgatggaaaaaactgc
## acactgatagatgctctattgggagaccctcattgtgatggcttccaaaataaggaatgg
## gacctttttgttgaacgcagcaaagcttacagcaactgttacccttatgatgtgccggat
## tatgcctccctaggtcactagttgcctcatcaggcaccctggagtttatcaatgaagac
## ttcaattggactggagtcgctcaggatgggaaaagctatgcttgcaaaaggggatctgtt
## aacagtttctttagtagattgaattggttgcacaaattagaatacaaatatccagcgctg
## aacgtgactatgccaaacaatggcaaatttgacaaattgtacatttgggggttcaccac
## ccgagcacggacagtgaccaaaccagcctatatgttcgagcatcagggagagtcacagtc
## tctaccaaaagaagccaacaaactgtaatcccgaatatcgggtctagaccctgggtaagg
## ggtctgtccagtagaataagcatctattggacaatagtaaaaccgggagacatacttttg
## attaatagcacagggaatctaattgctcctcggggttacttcaaaatacgaaatgggaaa
## agctcaataatgaggtcagatgcacccattggcaactgcagttctgaatgcatcactcca
## aatggaagcattcccaatgacaaaccttttcaaaatgtaaacaggatcacatatggggcc
## tgccccagatatgttaagcaaaacactctgaaattggcaacagggatgcggaatgtacca
## gagaaacaaactagaggcatattcggcgcaatcgcaggtttcatagaaaatggttgggag
## ggaatggtagacggttggtacggtttcaggcatcaaaattctgagggcacaggacaagca
## gcagatcttaaaagcactcaagcagcaaccgaccaaatcaacgggaaactgaataggtta
## atcgagaaaacgaacgagaaattccatcaaatcgaaaaagaattctcagaagtagaaggg
## agaattcaggacctcgagaaatatgttgaagacactaaaatagatctctggtcttacaac
## gcggagcttcttgttgccctggagaaccaacatacaattgatctaactgactcagaaatg
## aacaaactgtttgaaagaacaaggaagcaactgagggaaaatgctgaggacatgggcaat
## ggttgcttcaaaatataccacaaatgtgacaatgcctgcatagggtcaatcagaaatgga
## acttatgaccatgatgtatacagagacgaagcattaaacaaccggttccagatcaaaggt
## gttgagctgaagtcaggatacaaagattggatcctatggatttcctttgccatatcatgc
## tttttgctttgtgttgttttgctggggttcatcatgtgggcctgccaaaaaggcaacatt
## aggtgcaacatttgcatttga
```

Para contarmos quantas vezes o **nome** de uma sequencia FASTA aparece no nosso arquivo FASTA, usamos o comando grep, seguido de um padrao presente na primeira linha do arquivo FASTA. Cada sequencia FASTA tem o nome que aparece depois do sinal de "maior que" (>). Usar o sinal "maior que" (>) com grep, pode ser problematico pois este simbolo tambem e o simbolo para saida, usado pelo R. Acima, vimos que CY, aparece na primeira linha do arquivo FASTA). Assim, podemos procurar tanto pelos simbolos "CY", quanto pelos simbolos "E", e tambem, ">". A ressalva e que ">" precisa ser procurado utilisando-se uma sintaxe especial, como visto abaixo.

```
grep CY ~/Desktop/Gepoliano/UFSB/usaflu.fasta | wc -l
grep E ~/Desktop/Gepoliano/UFSB/usaflu.fasta | wc -l
grep -Ri -- '>' ~/Desktop/Gepoliano/UFSB/usaflu.fasta | wc -l
```

```
##       70
```

```
##        9
##       80
```

```
dna <- read.dna(file = "~/Desktop/Gepoliano/UFSB/usaflu.fasta", format = "fasta")
object.size(as.character(dna))/object.size(dna)
```

```
## 7.7 bytes
```

```
as.character(dna)[1:5, 1:10]
```

```
##          [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
## CY013200 "a"  "t"  "g"  "a"  "a"  "g"  "a"  "c"  "t"  "a"
## CY013781 "a"  "t"  "g"  "a"  "a"  "g"  "a"  "c"  "t"  "a"
## CY012128 "a"  "t"  "g"  "a"  "a"  "g"  "a"  "c"  "t"  "a"
## CY013613 "a"  "t"  "g"  "a"  "a"  "g"  "a"  "c"  "t"  "a"
## CY012160 "a"  "t"  "g"  "a"  "a"  "g"  "a"  "c"  "t"  "a"
```

```
annot <- read.csv("~/Desktop/Gepoliano/UFSB/usflu.annot.csv", header = TRUE, row.names = 1)
## Visualizar annot
annot
```

```
##    accession year                              misc
## 1   CY013200 1993       (A/New York/783/1993(H3N2))
## 2   CY013781 1993       (A/New York/802/1993(H3N2))
## 3   CY012128 1993       (A/New York/758/1993(H3N2))
## 4   CY013613 1993       (A/New York/766/1993(H3N2))
## 5   CY012160 1993       (A/New York/762/1993(H3N2))
## 6   CY012272 1994       (A/New York/729/1994(H3N2))
## 7   CY010988 1994       (A/New York/733/1994(H3N2))
## 8   CY012288 1994       (A/New York/734/1994(H3N2))
## 9   CY012568 1994       (A/New York/746/1994(H3N2))
## 10  CY013016 1994       (A/New York/750/1994(H3N2))
## 11  CY012480 1995       (A/New York/666/1995(H3N2))
## 12  CY010748 1995       (A/New York/648/1995(H3N2))
## 13  CY011528 1995       (A/New York/669/1995(H3N2))
## 14  CY017291 1995       (A/New York/681/1995(H3N2))
## 15  CY012504 1995       (A/New York/678/1995(H3N2))
## 16  CY009476 1996       (A/New York/565/1996(H3N2))
## 17  CY010028 1996       (A/New York/591/1996(H3N2))
## 18  CY011128 1996       (A/New York/599/1996(H3N2))
## 19  CY010036 1996       (A/New York/592/1996(H3N2))
## 20  CY011424 1996       (A/New York/577/1996(H3N2))
## 21  CY006259 1997       (A/New York/511/1997(H3N2))
## 22  CY006243 1997       (A/New York/508/1997(H3N2))
## 23  CY006267 1997       (A/New York/513/1997(H3N2))
## 24  CY006235 1997       (A/New York/505/1997(H3N2))
## 25  CY006627 1997       (A/New York/547/1997(H3N2))
## 26  CY006787 1998       (A/New York/506/1998(H3N2))
## 27  CY006563 1998       (A/New York/533/1998(H3N2))
## 28  CY002384 1998       (A/New York/330/1998(H3N2))
## 29  CY008964 1998       (A/New York/540/1998(H3N2))
## 30  CY006595 1998       (A/New York/542/1998(H3N2))
## 31  CY001453 1999       (A/New York/184/1999(H3N2))
## 32  CY001413 1999       (A/New York/263/1999(H3N2))
## 33  CY001704 1999       (A/New York/257/1999(H3N2))
## 34  CY001616 1999       (A/New York/265/1999(H3N2))
```

```
## 35  CY003785 1999       (A/New York/422/1999(H3N2))
## 36  CY000737 2000       (A/New York/180/2000(H3N2))
## 37  CY001365 2000       (A/New York/187/2000(H3N2))
## 38  CY003272 2000       (A/New York/437/2000(H3N2))
## 39  CY000705 2000       (A/New York/175/2000(H3N2))
## 40  CY000657 2000       (A/New York/169/2000(H3N2))
## 41  CY002816 2001       (A/New York/301/2001(H3N2))
## 42  CY000584 2001       (A/New York/127/2001(H3N2))
## 43  CY001720 2001       (A/New York/273/2001(H3N2))
## 44  CY000185 2001        (A/New York/83/2001(H3N2))
## 45  CY002328 2001        (A/New York/77/2001(H3N2))
## 46  CY000297 2002        (A/New York/96/2002(H3N2))
## 47  CY003096 2002       (A/New York/403/2002(H3N2))
## 48  CY000545 2002       (A/New York/115/2002(H3N2))
## 49  CY000289 2002        (A/New York/92/2002(H3N2))
## 50  CY001152 2002        (A/New York/74/2002(H3N2))
## 51  CY000105 2003       (A/New York/60A/2003(H3N2))
## 52  CY002104 2003           (A/Memphis/31/03(H3N2))
## 53  CY001648 2003       (A/New York/270/2003(H3N2))
## 54  CY000353 2003        (A/New York/21/2003(H3N2))
## 55  CY001552 2003       (A/New York/215/2003(H3N2))
## 56  CY019245 2004       (A/New York/908/2004(H3N2))
## 57  CY021989 2004       (A/New York/908/2004(H3N2))
## 58  CY003336 2004       (A/New York/354/2004(H3N2))
## 59  CY003664 2004       (A/New York/471/2004(H3N2))
## 60  CY002432 2004       (A/New York/362/2004(H3N2))
## 61  CY003640 2005       (A/New York/463/2005(H3N2))
## 62  CY019301 2005       (A/New York/918/2005(H3N2))
## 63  CY019285 2005       (A/New York/913/2005(H3N2))
## 64  CY006155 2005       (A/New York/258/2005(H3N2))
## 65  CY034116 2005       (A/Wisconsin/67/2005(H3N2))
## 66  EF554795 2006               (A/Ohio/2006(H3N2))
## 67  CY019859 2006       (A/New York/938/2006(H3N2))
## 68  EU100713 2006        (A/Maryland/09/2006(H3N2))
## 69  CY019843 2006       (A/New York/933/2006(H3N2))
## 70  CY014159 2006         (A/New York/7/2006(H3N2))
## 71  EU199369 2007       (A/Minnesota/08/2007(H3N2))
## 72  EU199254 2007           (A/Idaho/01/2007(H3N2))
## 73  CY031555 2007 (A/Kentucky/UR06-0571/2007(H3N2))
## 74  EU516036 2007         (A/Georgia/07/2007(H3N2))
## 75  EU516212 2007      (A/California/33/2007(H3N2))
## 76  FJ549055 2008        (A/Illinois/14/2008(H3N2))
## 77  EU779498 2008     (A/Mississippi/01/2008(H3N2))
## 78  EU779500 2008         (A/Indiana/02/2008(H3N2))
## 79  CY035190 2008 (A/Pennsylvania/PIT43/2008(H3N2))
## 80  EU852005 2008           (A/Texas/06/2008(H3N2))
```

```
table(annot$year)
```

```
##
## 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008
##    5    5    5    5    5    5    5    5    5    5    5    5    5    5    5    5
```

5) Filogenia Baseada em Matriz de Distanciamento de pares de Influenza

5.1) Matriz ou Heatmap de Influenza

dist.dna é a funcao usada para calcular a distancia entre as amostras. Note que o parametro "model", abaixo, pode assumir diferentes valores, de acordo com a preferencia do usuario. D é a matriz de distancia, um objeto de classe "dist". Funcoes podem ser acessadas em R, digitando-se ?funcao, ou, no caso de dist.dna, ?dist.dna, para mais ajuda oferecida pelo R. Essas digitacoes sao feitas na parte do R chamada de Console.

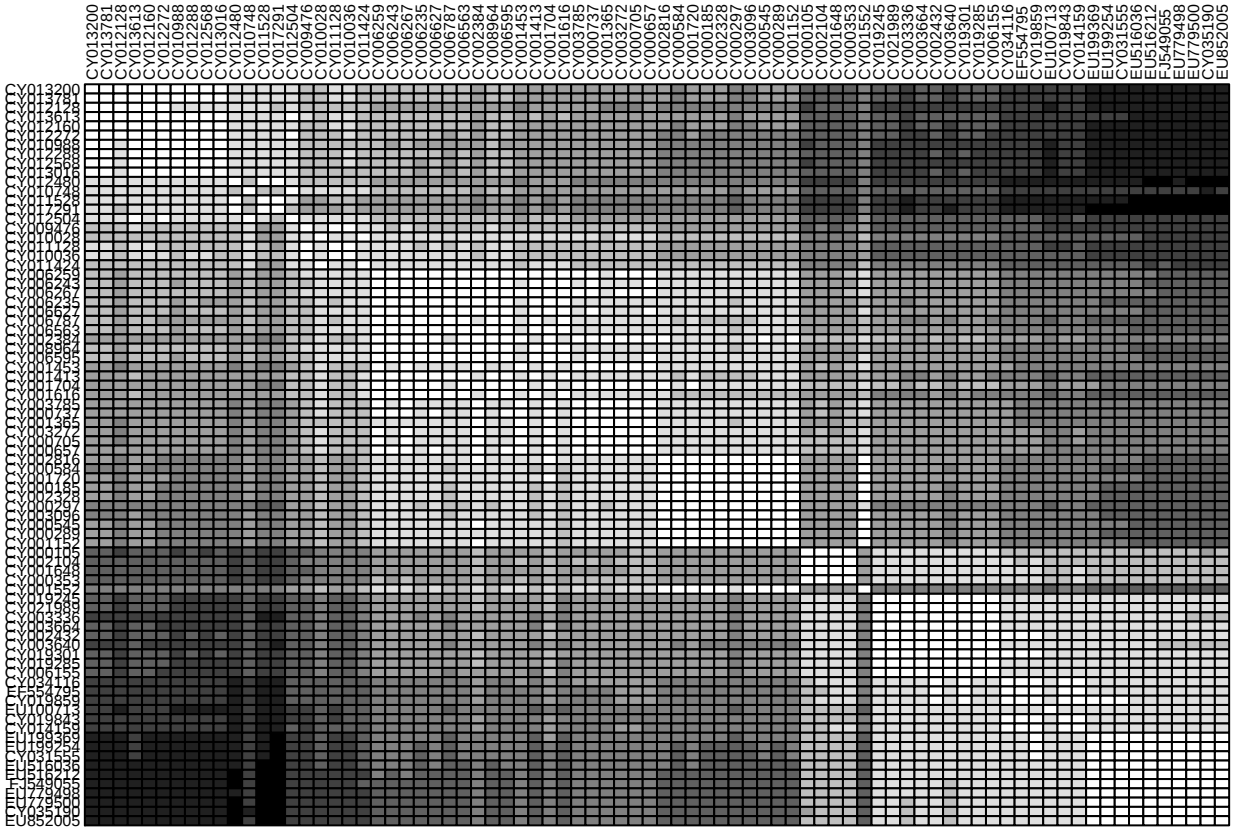
```
D <- dist.dna(dna, model = "TN93")
class(D)
```

```
## [1] "dist"
```

```
length(D)
```

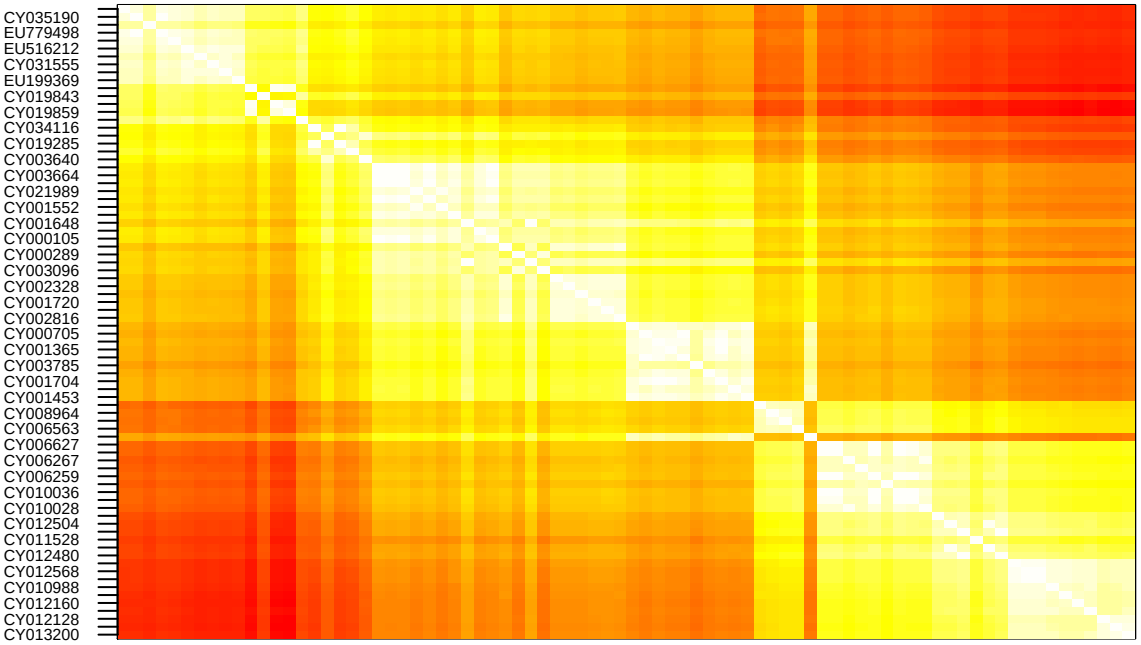
```
## [1] 3160
```

```
temp <- as.data.frame(as.matrix(D))
table.paint(temp, cleg = 0, clabel.row = 0.5, clabel.col = 0.5)
```



```
temp <- t(as.matrix(D))
temp <- temp[, ncol(temp):1]
par(mar = c(1, 5, 5, 1))
image(x = 1:80, y = 1:80, temp,
      col = rev(heat.colors(100)),
      xaxt = "n", yaxt = "n", xlab = "", ylab = "")
```

```
axis(side = 2, at = 1:80, lab = rownames(dna), las = 2, cex.axis = 0.5)
```



5.2) Arvore Filogenetica de Influenza

A funcao nj( ) executa a estimacao da arvore juntada pelos vizinhos, de Saitou e Nei (1987).

```
tre <- nj(D)
class(tre)
```

```
## [1] "phylo"
```

```
plot(tre, cex = 0.6)
title("Filogenia de Influenza (A simple NJ tree)")
```

6) Filogenia Baseada em Matriz de Distanciamento pares de SARS-CoV-2

6.1) Matriz ou Heatmap de SARS-CoV-2

Preparei os arquivos FASTA de SARS-CoV-2 de modo que tivessem o mesmo tamanho de sequencia. Inicialmente, devido a finalizacao de identificacao de genes em Neuroblastoma, consequi incluir apenas 4 sequencias de SARS-CoV-2 isoladas de pangolins chineses. Necessito ainda correr o alinhamento das outras sequencias e ver como posso transformar os arquivos FASTA de modo a terem o mesmo tamanho.

6.2) Carregar dados no workspace R

```
dna_SARS_CoV_2 <- read.dna(
  file = "~/Desktop/Gepoliano/UFSB/WorldSARS-CoV-2_Exact_LengthPangolin.fasta",
  format = "fasta", as.matrix = TRUE)
object.size(as.character(dna_SARS_CoV_2))/object.size(dna_SARS_CoV_2)
```

```
## 7.9 bytes
```

```
## Anotacao com metadata das sequencias FASTA
annot_SARS_CoV_2 <- read.csv(
  "~/Desktop/Gepoliano/UFSB/SARS_CoV-2_World.annot.csv",
  header = TRUE, row.names = 1)

annot_SARS_CoV_2
```

```
##                                                          accession year
## 1         hCoV-19/Brazil/BA-312/2020|EPI_ISL_415105|2020-03-04|30140bp 2020
## 2         hCoV-19/Brazil/BA-510/2020|EPI_ISL_427293|2020-03-06|30195bp 2020
## 3      hCoV-19/Brazil/L17_CD358/2020|EPI_ISL_476304|2020-03-31|30276bp 2020
## 4      hCoV-19/Brazil/L17_CD359/2020|EPI_ISL_476305|2020-03-31|30276bp 2020
## 5     hCoV-19/bat/Yunnan/RaTG13/2013|EPI_ISL_402131|2013-07-24|30229bp 2013
## 6     hCoV-19/bat/Yunnan/RmYN01/2019|EPI_ISL_412976|2019-06-25|28024bp 2019
## 7     hCoV-19/bat/Yunnan/RmYN02/2019|EPI_ISL_412977|2019-06-25|30041bp 2019
## 8        hCoV-19/pangolin/Guangxi/P4L/2017|EPI_ISL_410538|2017|30178bp 2017
## 9        hCoV-19/pangolin/Guangxi/P1E/2017|EPI_ISL_410539|2017|30174bp 2017
## 10       hCoV-19/pangolin/Guangxi/P5L/2017|EPI_ISL_410540|2017|30179bp 2017
## 11       hCoV-19/pangolin/Guangxi/P5E/2017|EPI_ISL_410541|2017|30178bp 2017
## 12       hCoV-19/pangolin/Guangxi/P2V/2017|EPI_ISL_410542|2017|30168bp 2017
## 13       hCoV-19/pangolin/Guangxi/P3B/2017|EPI_ISL_410543|2017|30173bp 2017
## 14     hCoV-19/pangolin/Guangdong/P2S/2019|EPI_ISL_410544|2019|30142bp 2019
## 15       hCoV-19/pangolin/Guangdong/1/2019|EPI_ISL_410721|2019|30198bp 2019
## 16 hCoV-19/pangolin/China/MP789/2019|EPI_ISL_412860|2019-03-19|27554bp 2019
## 17  hCoV-19/pangolin/Guangdong/cDNA8-S/2019|EPI_ISL_471461|2019|3845bp 2019
## 18  hCoV-19/pangolin/Guangdong/cDNA9-S/2019|EPI_ISL_471462|2019|3845bp 2019
## 19 hCoV-19/pangolin/Guangdong/cDNA16-S/2019|EPI_ISL_471463|2019|3845bp 2019
## 20 hCoV-19/pangolin/Guangdong/cDNA18-S/2019|EPI_ISL_471464|2019|3845bp 2019
## 21 hCoV-19/pangolin/Guangdong/cDNA20-S/2019|EPI_ISL_471465|2019|3845bp 2019
## 22   hCoV-19/pangolin/Guangdong/cDNA31-S/2019|EPI_ISL_471466|2019|3845 2019
## 23     hCoV-19/pangolin/Guangdong/A22-2/2019|EPI_ISL_471467|2019|30197 2019
## 24    hCoV-19/pangolin/Guangdong/FM45-9/2019|EPI_ISL_471468|2019|30198 2019
## 25    hCoV-19/pangolin/Guangdong/SM44-9/2019|EPI_ISL_471469|2019|30197 2019
## 26    hCoV-19/pangolin/Guangdong/SM79-9/2019|EPI_ISL_471470|2019|30197 2019
```

```
##                                             misc
## 1                          (A/Bahia/783/1993(H3N2))
## 2                          (A/Bahia/802/1993(H3N2))
## 3              (A/Brazil_L17_CD358/758/1993(H3N2))
## 4              (A/Brazil_L17_CD359/766/1993(H3N2))
## 5               (A/Yunnan_Bat_2013/762/1993(H3N2))
## 6               (A/Yunnan_Bat_2019/729/1994(H3N2))
## 7               (A/Yunnan_Bat_2019/733/1994(H3N2))
## 8              (A/Guangxi_Pangolin/734/1994(H3N2))
## 9          (A/Guangxi_Pangolin_P1E/746/1994(H3N2))
## 10         (A/Guangxi_Pangolin_P5L/750/1994(H3N2))
## 11         (A/Guangxi_Pangolin_P5E/666/1995(H3N2))
## 12         (A/Guangxi_Pangolin_P2V/648/1995(H3N2))
## 13         (A/Guangxi_Pangolin_P3B/669/1995(H3N2))
## 14       (A/Guangdong_Pangolin_P2S/681/1995(H3N2))
## 15         (A/Guangdong_Pangolin_1/678/1995(H3N2))
## 16         (A/China_Pangolin_MP789/565/1996(H3N2))
## 17  (A/Guangdong_Pangolin_cDNA8-S/591/1996(H3N2))
## 18  (A/Guangdong_Pangolin_cDNA9-S/599/1996(H3N2))
## 19 (A/Guangdong_Pangolin_cDNA16-S/591/1996(H3N2))
## 20 (A/Guangdong_Pangolin_cDNA18-S/577/1996(H3N2))
## 21 (A/Guangdong_Pangolin_cDNA20-S/511/1997(H3N2))
## 22 (A/Guangdong_Pangolin_cDNA31-S/508/1997(H3N2))
## 23    (A/Guangdong_Pangolin_A22-2/513/1997(H3N2))
## 24   (A/Guangdong_Pangolin_FM45-9/505/1997(H3N2))
## 25   (A/Guangdong_Pangolin_FM44-9/547/1997(H3N2))
## 26   (A/Guangdong_Pangolin_SM79-9/547/1997(H3N2))
```

```
## O seguinte comando permite visualizar o ano de isolamento da sequencia genetica.
table(annot_SARS_CoV_2$year)
```

```
##
## 2013 2017 2019 2020
##    1    6   15    4
```

Agora, desejo usar a interface linux do R Markdown para visualizar o arquivo fasta acima. Ao executar o comando abaixo, poderemos visualizar a sequencia FASTA contendo todos os genomas de SARS-CoV-2.

```
head ~/Desktop/Gepoliano/UFSB/WorldSARS-CoV-2_Exact_LengthPangolin.fasta
# head ~/Desktop/Gepoliano/UFSB/SARS_CoV-2_World.annot.csv
```

```
## >hCoV-19/pangolin/Guangdong/A22-2/2019|EPI_ISL_471467|2019
## NNNNNNNNNNNNNNNNNTCCCAGGTAACAAACCAACCAACTCTCGATCTCTTGTAGATCTGTTCTCTAAACGAACTTTAA
## AATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAATAACTAATTACTGTCGTTGACAGG
## ACACGAGTAACTCGTCTATCTTCTGCAGGCTGCTTACGGTTTCGTCCGTGTTGCAGCCGATCATCAGCATACCTAGGTTT
## CGTCCGGGTGTGACCGAAAGGTAAGATGGAGAGCCTTGTCCCTGGTTTCAACGAGAAAACACACGTCCAACTCAGTTTGC
## CTGTTTTACAGGTTCGCGACGTGCTCGTACGTGGCTTTGGAGACTCCGTGGAGGAGGCTATCTCAGAGGCACGTCAACAT
## CTCAAGGATGGCACTTGTGGCTTAGTAGAGGTTGAAAAAGGCGTCTTGCCTCAACTTGAACAGCCCTATGTGTTCATCAA
## ACGTTCTGATGCCCGAACTGCACCGCATGGCCATGTAATGGTTGAATTGGTTGCAGAACTCAATGGTGTTCAGTACGGTC
## GTAGTGGTGAGACACTTGGTGTTCTCGTACCCCATGTGGGTGAAACACCTGTTGCTTACCGCAAAGTTCTTCTTCGCAAG
## AACGGTAATAAAGGAGCTGGTGGTCACAGCTATGGCGCCGATCTAAAGTCCTATGACTTAGGTGACGAGCTGGGCACTGA
```

6.3) Matriz de Distanciamento de pares de SARS-CoV-2

Esta parte produz o plot de uma matriz de distancia, mostrando o quanto cada amostra se assemelha as demais amostras da analise. Esta clusterizacao e semelhante a clusterizacao produzida entre as amostras usadas em Analise de Expressao Genica feita no DESeq2.

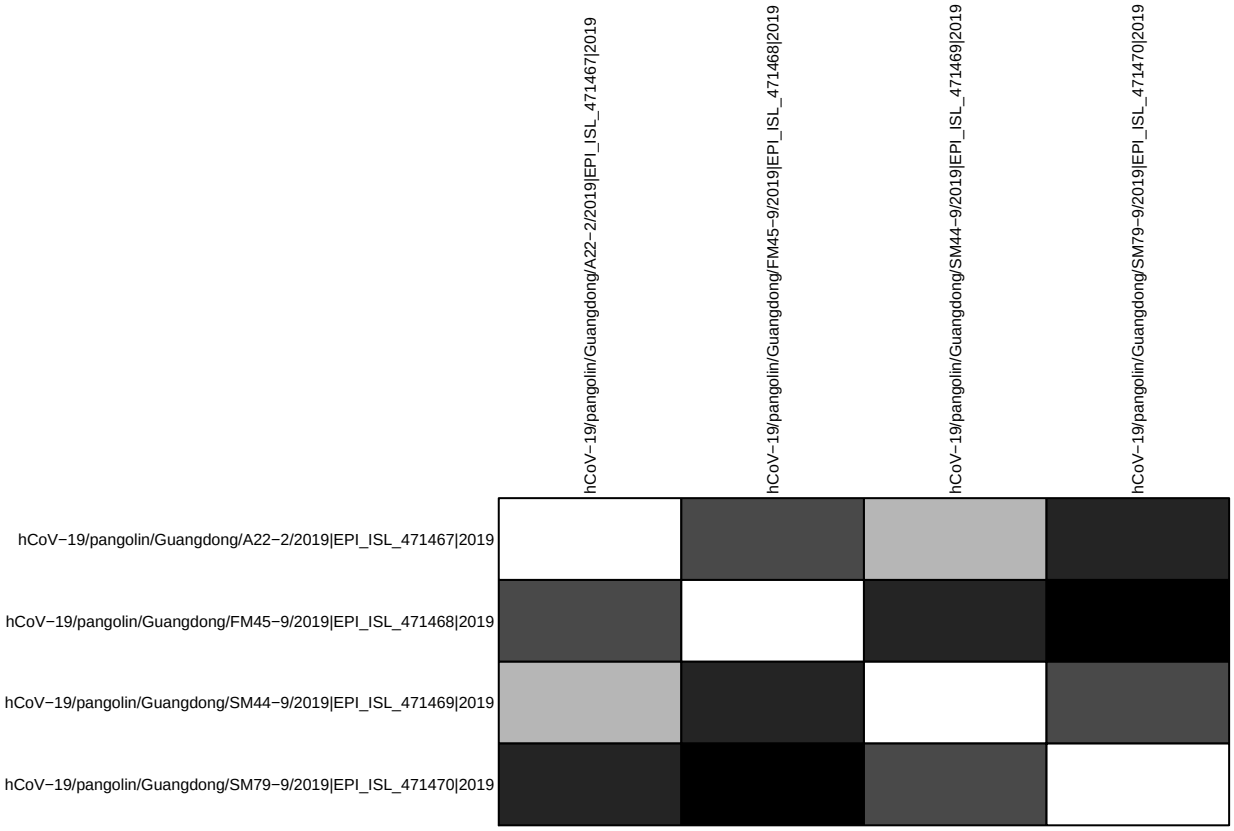
```
D_SARS_CoV_2 <- dist.dna(dna_SARS_CoV_2, model = "TN93")
class(D_SARS_CoV_2)
```

```
## [1] "dist"
```

```
length(D_SARS_CoV_2)
```

```
## [1] 6
```

```
temp <- as.data.frame(as.matrix(D_SARS_CoV_2))
table.paint(temp, cleg = 0, clabel.row = 0.5, clabel.col = 0.5)
```



Note que para usar image para produzir plots similares, os dados precisam antes ser transformados; por exemplo:

```
temp <- t(as.matrix(D_SARS_CoV_2))
temp <- temp[, ncol(temp):1]
par(mar = c(1, 5, 5, 1))
image(
  x = 1:80, y = 1:80,
  temp, col = rev(heat.colors(100)),
  xaxt = "n", yaxt = "n", xlab = "", ylab = "")
axis(side = 2, at = 1:80, lab = rownames(dna_SARS_CoV_2), las = 2, cex.axis = 0.5)
axis(side = 3, at = 1:80, lab = rownames(dna_SARS_CoV_2), las = 3, cex.axis = 0.5)
```

6.4) Construcao Arvore Filogenetica de SARS-CoV-2 com base na Matriz de Distanciamento de pares

Finalmente plotamos a arvore filogenetica de SARS-CoV-2 e poderemos observar como as diferentes sequencias distribuidas ao redor do globo, se relacionam geneticamente, ou filogeneticamente.

```
tre <- nj(D_SARS_CoV_2)
class(tre)
```

```
## [1] "phylo"
```

```
plot(tre, cex = 0.6, main = "Primeira Figura Filogenia de SARS-CoV-2")
```