Maria Luisa Santos Moreno Sanches - 111859

Algoritmos em Bioinformática: Contagem de nucleotídeos de uma sequência FASTA do Genoma do vírus da SARS-Covid-2

São José dos Campos - Brasil

Abril de 2021

Maria Luisa Santos Moreno Sanches - 111859

Algoritmos em Bioinformática: Contagem de nucleotídeos de uma sequência FASTA do Genoma do vírus da SARS-Covid-2

Relatório apresentado à Universidade Federal de São Paulo como parte dos requisitos para aprovação na disciplina de Algoritmos em Bioinformática.

Docente: Prof. Dr. Claudio Saburo Shida

Universidade Federal de São Paulo - UNIFESP

Instituto de Ciência e Tecnologia - Campus São José dos Campos

São José dos Campos - Brasil

Abril de 2021

Sumário

1 Assunto e Objetivos

O vírus da SARS-Covid-19 é uma fita de RNA dentro de uma capa lipo-proteica. Na superfície da capa há sinalizadores moleculares que auxiliam o vírus a injetar a fita de RNA para dentro da célula. Uma vez dentro, a fita de RNA utiliza a maquinaria de síntese de proteínas para se replicar, produzir novos vírus e lançá-los para fora da célula.

O Genoma da SARS-Covid-19 é uma sequência de DNA obtida através da fita de RNA extraída de um paciente infectado, que pode ser interpretada como uma *string*.

O objetivo desta atividade é se familiarizar com a biblioteca *biopython* do *Python* (1) e contar os nucleotídeos A, T, C, G de 3 sequências SARS-Covid-2, sendo elas da China, Brasil e Chile.

2 Metodologia

Os 3 arquivos disponibilizados estão no formato FASTA, logo é necessário importar a biblioteca *biopython* do *Python*. Para deixar o código genérico, foi adotado a leitura dos arquivos de um diretório como dados de entrada. Abaixo mostra o código correspondente a essa leitura.

```
# All files in entrada/ directory will be used
from os import listdir
from os.path import isfile, join
path = 'entrada/'
files = [f for f in listdir(path) if isfile(join(path, f))]
```

A leitura dos dados de cada arquivo foi realizada pelo comando *SeqIO*. O código desenvolvido está comentado abaixo, e vale ressaltar que seu desenvolvimento foi baseado num *post* do fórum *Bioinformatics Explained* (2).

```
# Import parts of Biopython
from Bio import SeqIO
from Bio.Seq import Seq

# For each file in the directory
for file in files:
    # File path to your FASTA file
    path_to_file = 'entrada/' + str(file)
    # Open file with "with" statement to avoid problems with access
    # to original file (in case computer hangs or there will be any other problem)
    with open(path_to_file, mode='r') as handle:
        # Use Biopython's parse function to process individual
        # FASTA records (thus reducing memory footprint)
        for record in SeqIO.parse(handle, 'fasta'):
            # Extract individual parts of the FASTA record
            identifier = record.id
            description = record.description
            sequence = record.seq

            print('\n-----------------------------------------------------------')
            print('File: ' + file + '\n')
            print('Processing the record {}:'.format(identifier))
            print('Its description is: {}'.format(description))
            amount_of_nucleotides = len(sequence)
            print('Its sequence contains {} nucleotides, which:'.format(
                amount_of_nucleotides))
            print('A: {}'.format(sequence.count("A")))
            print('C: {}'.format(sequence.count("C")))
            print('G: {}'.format(sequence.count("G")))
            print('T: {}'.format(sequence.count("T")))
```

O código na íntegra está disponível no GitHub do autor.

3 Resultados

A quantidade de cada nucleotídeo das sequências SARS-Covid-2 da China, Brasil e Chile se encontra na Tabela 1.

Tabela 1 – Quantidade de nucleotídeos de cada sequência SARS-Covid-2

Sequência SARS-Covid-2	# A	# C	# G	# T	# Total
Brasil	8927	5492	5861	9596	29876
China	8954	5492	5863	9594	29903
Chile	8876	5461	5836	9551	29724

Fonte: O Autor.

Somente com esses dados é possível dizer que a sequência SARS-Covid-2 da China, entre as analisadas, é a que possui um número maior de nucleotídeos. Não há uma grande diferença entre as quantidades de cada nucleotídeo entre as sequências.

4 Conclusão

A biblioteca *biopython* é muito útil para análises de grandes arquivos, não sendo necessário um grande conhecimento em programação para poder realizar algumas análises, o que tornou a atividade de contar nucleotídeos fácil e rápida. Os resultados apresentados, além de acurados, foram bem próximos entre si, com a China apresentando a maior quantidade de nucleotídeos no total.

Referências

1 BIOPYTHON. *Biopython*. Disponível em: <https://biopython.org/>. Acesso em: 21 abr. 2021. Citado na página 3.

2 BIOINFORMATICS EXPLAINED. *Best way to processing huge fasta files using python*. Disponível em: <https://www.biostars.org/p/250123/>. Acesso em: 21 abr. 2021. Citado na página 4.