UNIVERSIDADE FEDERAL DE SÃO PAULO

Instituto de Ciência e Tecnologia

Bacharelado em Ciência e Tecnologia

DANIEL PUGLISI - 140355

LILIANA MARIA DE SOUSA BUENO - 140403

REGINA MIYUKI ETO YOKOYAMA - 140071

TRABALHO FINAL SOBRE BIOPYTHON - GENBANK

São José dos Campos

2021

SUMÁRIO

INTRODUÇÃO

A linguagem Python é a principal linguagem utilizada na área de bioinformática, suas bibliotecas permitem maior organização, precisão e praticidade na hora de trabalhar com os temas voltados a esta área. No trabalho a seguir, a linguagem Python foi utilizada na análise e resolução de problemas envolvendo sequências de nucleotídeos, para isto foi usada a biblioteca Biopython que conta com um grande número de "facilitadores" que serão utilizados para responder às questões propostas. Além do Biopython, outras bibliotecas serão necessárias para realização deste trabalho tal qual a Matplotlib, uma biblioteca que permite a visualização e organização dos dados obtidos em gráficos, facilitando assim a visualização e comparação dos resultados, os apresentando de forma mais agradável.

O DNA é um composto de fita-dupla que armazena todas informações hereditárias de um ser vivo e é formado por longas cadeias de polímeros pareados não ramificados compostas por 4 monômeros: Adenina, Timina, Guanina e Citosina. Além disso, para ser expresso as características presentes no DNA é preciso que ocorra a transcrição - processo de leitura do DNA- para formar o RNA, onde a Timina será substituída pela Uracila, e, após esse processo, acontecerá a tradução, controlada pela molécula de RNA, sintetizando as proteínas. A proteína é semelhante ao DNA e RNA por apresentar cadeias poliméricas longas, contudo diferem nos monômeros, pois, enquanto os dois possuem 4 monômeros, a proteína possui 20 aminoácidos que são ligados de forma padronizada. [1]

Ademais, sabe-se que as proteínas possuem quatro níveis de organização: sequência de aminoácido é a estrutura primária, as que formam alfa-hélice, folha-beta e outros é a estrutura secundária, estrutura terciária é tridimensional e estrutura quaternária aquela que é formada por mais de uma cadeia polipeptídica. Apesar disso, para este trabalho será interessante apenas a estrutura secundária.[1]

METODOLOGIA

O trabalho presente foi dividido em duas partes, a primeira com análise dos nucleotídeos (letras A, B, C, D e E) e a segunda, análise das proteínas (letra F e G). Para isso foi preciso identificar os nucleotídeos de códigos MN908947.3, MT012098, MZ264787.1, NC_019843.3 na ferramenta computacional, Genbank, e foi baixado seus respectivos arquivos fasta. Para a letra A foram utilizadas as informações presentes no Geobank para descrevê-las como a origem e o organismo em que estão presentes. [2]

Além disso, para as letras B e C foi desenvolvido um código com o Biopython que leu os arquivos fasta baixados anteriormente, examinou as sequências dos nucleotídeos, descobriu seus respectivos tamanhos e "*plotou*" um gráfico de barra com a frequência dos nucleotídeos, com o auxílio do pacote matplotlib.pyplot, em ordem alfabética para facilitar a análise de cada um. Após isso, foi analisada a diferença de frequência e o porquê delas.

Ademais, a letra D teve um estudo realizado a partir do link disponibilizado [3], e descoberto uma função do biopython o qual forneceu o conteúdo GC e, a partir disso, foi desenvolvido um código que calculou a temperatura de melting, Tm, de cada GC e, por fim, discutiu-se sua importância para o PCR. Para fechar essa primeira parte, na letra E foi realizado o alinhamento global de 2 a 2 dos primeiros 800 nucleotídeos, resultando em 6 alinhamentos totais que foram imprimidos e anotados os scores máximo e sua similaridade de cada alinhamento.

A segunda parte do trabalho conta com a letra F, foi utilizado os arquivos fasta que contém somente as sequências de nucleotídeos codificadora de cada organismo, e utilizando a função *translate()* do biopython realizou-se a tradução das sequências possibilitando montar um gráfico de barras com a frequência dos aminoácidos em ordem alfabética. Para a última letra foi utilizado o arquivo fasta que continham as sequências de aminoácidos de cada proteína de cada organismo, e, assim, foi determinado suas respectivas estruturas secundárias de cada proteína utilizando a função *secondary_structure_fraction()* do pacote proteinanalysis, como a saída desta função é uma tupla com três números representando a fração correspondente de cada tipo de estrutura (*helix, turn, sheet*), e com o auxílio do matplotlib foram feitas 4 tabelas para representar a estrutura de cada proteína de cada organismo.

RESULTADO

O organismo e a origem de MN908947.3, MT012098, MZ264787.1 e NC_019843.3 estão ilustrados na tabela a seguir:

Tabela [1]: Origem e Organismo referente a MN908947.3, MT012098, MZ264787.1 e NC_019843.3.

	MN908947.3	MT012098	MZ264787.1	NC_019843.3
ORIGEM	China	India: Kerala State	Brazil: Amazonas, Manaus	Saudi Arabia
ORGANISMO	*Severe acute respiratory syndrome coronavirus 2*	*Severe acute respiratory syndrome coronavirus 2*	*Severe acute respiratory syndrome coronavirus 2*	*Middle East respiratory syndrome-related coronavirus*

O tamanho das sequências de DNA de cada organismos:

- MN908947.3 tem 29903 pares de base
- MT012098.1 tem 29854 pares de base
- MZ264787.1 tem 29866 pares de base
- NC_019843.3 tem 30119 pares de base

A frequência de nucleotídeos de cada uma das sequências está representada na Figura 1, analisando o gráfico é possível reparar que existe diferença entre as frequências de A com T e de G com C se dá pelo fato de ser um DNA fita simples.

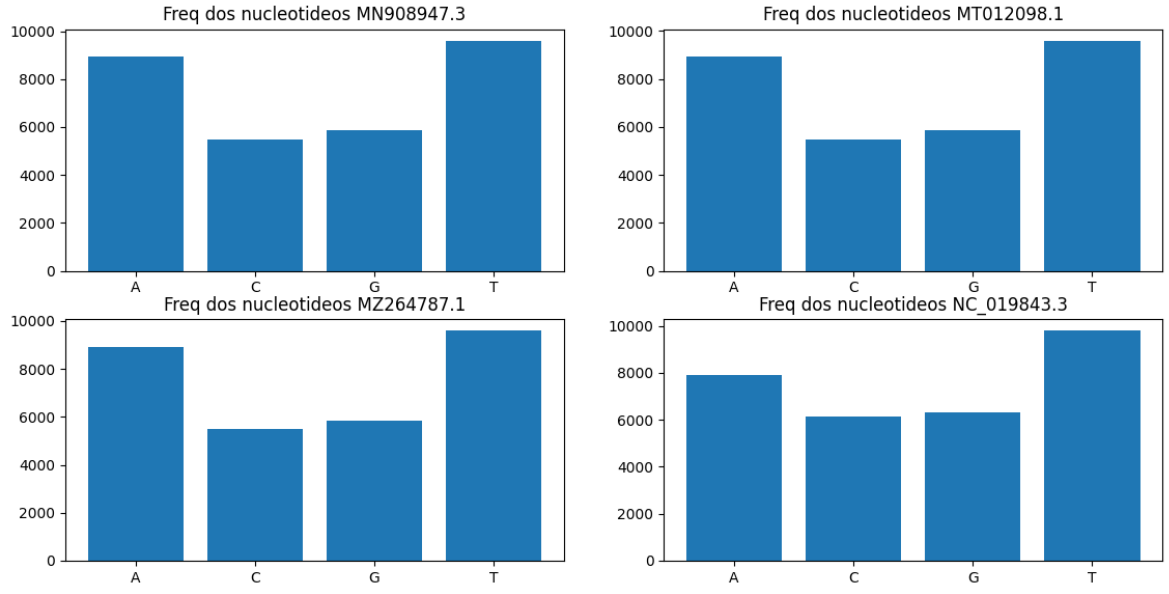

Figura 1: Gráfico de barras representando a frequência de nucleotídeos

A temperatura de melting de cada sequência de DNA foi encontrada através da função GC dentro do Biopython, esta função fornece a porcentagem de GC na sequência e possibilita a realização da fórmula, tm=64.9+0,41(%GC)- (500/length of DNA). Após a aplicação da fórmula em todas as sequências temos as seguintes temperaturas em Celsius:

Tabela [2]: Temperatura de Melting das sequências.

	MN908947.3	**MT012098.1**	**MZ264787.1**	**NC_019843.3**
Temperatura de Melting (tm)	80.45211851653681	80.47077778522142	80.44667514899888	81.79033500448222

A temperatura de melting é muito importante para a técnica de PCR pois se trata da temperatura média, que garante que enquanto metade dos primers estiverem nas cadeias-mãe, a outra metade estará inteiramente disponível para o próximo ciclo. A temperatura pode interferir na reação de PCR e, consequentemente, o número 2n de sequências de DNA resultantes pode não ser obtido após a execução de n ciclos de PCR. [5][6]

O Alinhamento de sequências é uma forma de organizar estruturas primárias de DNA, RNA ou proteína para identificar regiões similares que possam ser consequência de relações funcionais, estruturais ou evolucionárias entre elas. Os scores e as porcentagens de similaridade dos 800 primeiros nucleotídeos das sequências analisadas são:

- MN908947.3 + MT012098.1 = Score: 787, Similaridade: 98,375%
- MN908947.3 + MZ264787.1 = Score 786; Similaridade: 98,25%
- MN908947.3 + NC_019843.3 = Score 530; Similaridade: 66,25%
- MT012098.1 + MZ264787.1 = Score 792; Similaridade: 99%
- MT012098.1 + NC_019843.3 = Score 531; Similaridade: 66,375%
- MZ264787.1 + NC_019843.3 = Score: 530; similaridade: 66,25%

É possível observar com base nos resultados dos scores de alinhamento a similaridade entre as 3 primeiras sequências e a notável discrepância de NC_019843.3 perante as outras, o que corrobora com a diferença de organismos que já vimos no primeiro resultado deste trabalho.(todas sequências são organismos *Severe acute respiratory syndrome coronavirus 2,* exceto o NC_019843.3 que é *Middle East respiratory syndrome-related coronavirus)*

A frequência dos aminoácidos presentes em cada organismo está representada na Figura 2.

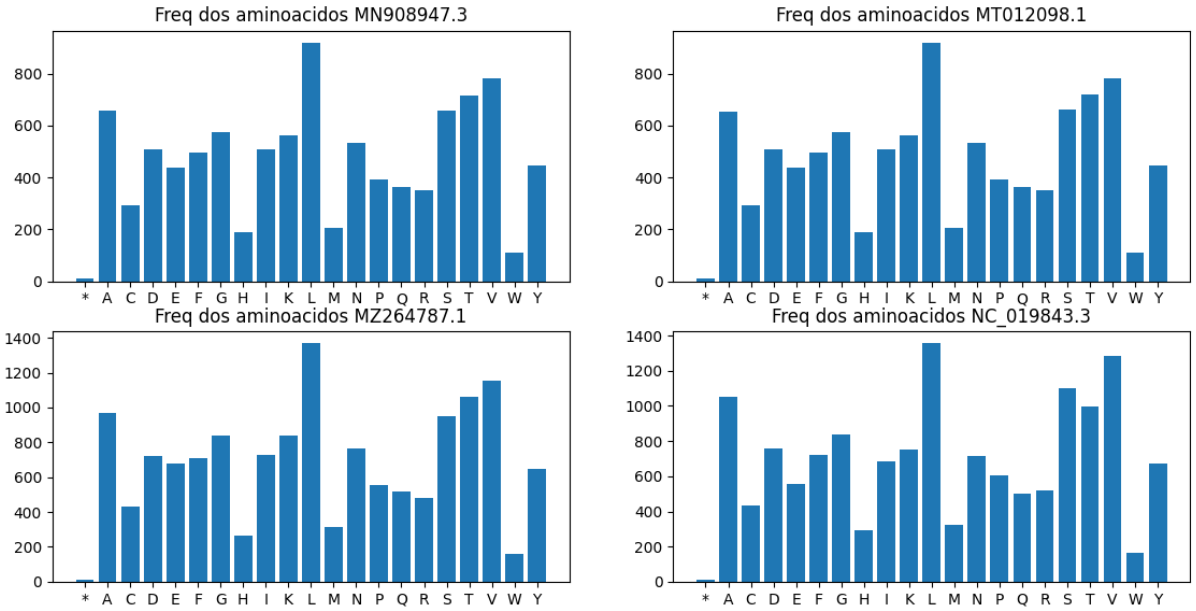

Figura 2: Frequência dos aminoácidos

Com o código desenvolvido foi possível determinar a estrutura secundária. É possível observar que a estrutura helix está presente em maior proporção na maioria das proteínas quando comparado com outras, e isso ocorre pelo simples fato de ser a mais simples e apresentar abundância na ligação de hidrogênio entre o **H** do amino e o **O** da carbonila, formando, então, uma alfa-hélice mais estável.[4] As tabelas a seguir mostram as estruturas das proteínas de cada organimo.

Tabela [3]: Estruturas secundárias das proteínas de MN908947.3

HELIX	TURN	SHEET
0.3341319052987599	0.21505073280721532	0.2343573844419391
0.3330714846818539	0.25687352710133543	0.19560094265514533
0.41090909090909095	0.20363636363636362	0.2109090909090909
0.52	0.21333333333333335	0.28
0.42342342342342343	0.20270270270270271	0.2927927927927928
0.4426229508196722	0.14754098360655737	0.2786885245901639
0.3801652892561984	0.15702479338842976	0.2727272727272727
0.396694214876033	0.19008264462809918	0.18181818181818182
0.18615751789976134	0.31026252983293556	0.19809069212410502
0.47368421052631576	0.23684210526315788	0.21052631578947367

Tabela [4]: Estruturas secundárias das proteínas de MT012098.1

HELIX	TURN	SHEET
0.33413190529875986	0.21505073280721532	0.2342164599774521
0.3333333333333333	0.25707547169811323	0.19575471698113206
0.41090909090909095	0.20363636363636362	0.2109090909090909
0.52	0.21333333333333335	0.28
0.42342342342342343	0.20270270270270271	0.2927927927927928
0.4426229508196722	0.14754098360655737	0.2786885245901639
0.3801652892561984	0.15702479338842976	0.2727272727272727
0.396694214876033	0.19008264462809918	0.18181818181818182
0.18615751789976134	0.31026252983293556	0.19809069212410502
0.47368421052631576	0.23684210526315788	0.21052631578947367

Tabela [5]: Estruturas secundárias das proteínas de MZ264787.1

HELIX	TURN	SHEET
0.334554678692221	0.21434611048478014	0.23463923337091316
0.33416572077185014	0.21589103291713962	0.2474460839954597
0.3362136684996072	0.2584446190102121	0.19402985074626866
0.41090909090909095	0.20363636363636362	0.2109090909090909
0.52	0.21333333333333335	0.28
0.42342342342342343	0.20270270270270271	0.2927927927927928
0.4426229508196722	0.14754098360655737	0.2786885245901639
0.3801652892561984	0.15702479338842976	0.2727272727272727
0.5813953488372092	0.06976744186046512	0.4186046511627907
0.396694214876033	0.19008264462809918	0.17355371900826447
0.18615751789976134	0.3054892601431981	0.19809069212410502
0.47368421052631576	0.23684210526315788	0.21052631578947367

Tabela [6]: Estruturas secundárias das proteínas de NC_019843.

HELIX	TURN	SHEET
0.3424696241876236	0.21701045493077142	0.22901949703306018
0.3457071282168071	0.21817353677977683	0.23707583693919382
0.3296378418329638	0.2697708795269771	0.2032520325203252
0.30097087378640774	0.27184466019417475	0.22330097087378642
0.3211009174311927	0.25688073394495414	0.26605504587155965
0.35772357723577236	0.22357723577235772	0.21544715447154472
0.45535714285714285	0.22321428571428573	0.20535714285714285
0.45121951219512196	0.17073170731707316	0.24390243902439024
0.3926940639269406	0.2328767123287671	0.2557077625570776
0.20823244552058112	0.3365617433414044	0.19370460048426152
0.3125	0.2767857142857143	0.375

Além disso foi possível determinar, também, a quantidade de proteínas que cada organismo possui:

- MN908947.3 tem 10 proteínas
- MT012098.1 tem 10 proteínas
- MZ264787.1 tem 12 proteínas
- NC_019843.3 tem 11 proteínas

CONCLUSÃO

Ao longo do desenvolvimento deste trabalho foi possível aplicar os conhecimentos obtidos em aula, desde a manipulação de arquivos, necessárias para análise dos arquivos fasta durante todo o percurso, até a utilização de bibliotecas essenciais para a linguagem Python em bioinformática. A análise das sequências de nucleotídeos revelaram dados importantes sobre os mesmos, sua identificação, origem, tradução entre outros, possibilitam uma maior compreensão sobre os DNAs observados. A utilização de gráficos é também um destaque, onde, sem o auxílio deles, todos esses resultados seriam mais complexos de serem feitos e compreendidos, logo esta visualização provida pelo pacote matplotlib.pyplot reitera a utilização de Python neste projeto.

Os resultados obtidos possibilitaram uma clara comparação entre as sequências de nucleotídeos escolhidas, as similaridades e diferenças encontradas são reforçadas ao longo do trabalho possibilitando uma conversação das diferentes análises entre si, como por exemplo os scores de alinhamentos sendo menores na comparação de dois organismos diferentes já identificados anteriormente ou a temperatura de melting de NC_019843.3 sendo a maior de todas, corroborando com o gráfico de frequência de nucleotídeos que o apresenta com maior número de GC dentre todas as sequências.

REFERÊNCIA

[1] Alberts, B. Biologia Molecular da Célula. [Digite o Local da Editora]: Grupo A, 2017. 9788582714232. Disponível em: https://integrada.minhabiblioteca.com.br/#/books/9788582714232/. Acesso em: 2021 ago. 09.

[2] Site: https://www.ncbi.nlm.nih.gov/genbank/

[3] Link: http://www.biology.arizona.edu/biomath/tutorials/Linear/LinearFunctionApplication/DNAmelt.html

[4] Experimentos de bioquímica. Fcfar UNESP. Dsponível em: https://www.fcfar.unesp.br/alimentos/bioquimica/introducao_proteinas/introducao_proteinas_dois.htm. Acesso em: 09 de Agosto de 2021.

[5] Técnicas de PCR: Aplicações e Padronização de Reações . Disponível em: https://www.imt.usp.br/wp-content/uploads/proto/protocolos/aula2.pdf. Acesso em 09 de Agosto de 2021

[6] Temperatura de Melting: um estudo comparativo . Disponível em: http://www.facom.ufms.br/wp-content/uploads/2015/11/Temperatura-de-Melting.pdf. Acesso em 09 de Agosto de 2021

APÊNDICE

Link com os arquivos fasta utilizados para a realização dos códigos: https://drive.google.com/drive/folders/1zCMceSF2A6U9WbjtkL5oFhOFnvUC-dwL?usp=sharing