BACHARELADO EM CIÊNCIA E TECNOLOGIA - BCT

Trabalho Final

João Victor de Carvalho Rangel 163833

Renan Bueno Angioletto 142754

UC: Algoritmos em Bioinformática

Professor: : Drª Thiago Martini

São José dos Campos - SP

Março de 2023

Sumário

1.Resumo

O trabalho consiste em baixar os arquivos fasta de uma sequência de nucleotídeos do Genbank, um site com proteínas, e fazer sua análise para obter informações, determinando o tamanho das sequências,além de criar gráficos para ver visualmente as informações desejadas e além disso queremos ver quais as propriedades físicas como ponto de fusão e obter a síntese proteica e o alinhamento global, além de saber as estruturas secundárias desses nucléotideos.No geral queremos conhecer sua estrutura sua propriedades e quais proteínas ela expressa.

2.Introdução

Deseja-se a partir de uma de uma sequência de quatro nucleotídeos obter informações da sequência do DNA então, primeiros achamos a frequência de cada base e depois de saber seu tamanho e as frequências de cada base então desejamos conhecer as propriedades físicas desse nucleotídeo como o ponto de fusão dos nucleotídeos.

Posteriormente a isso, faremos o alinhamento global dois a dois para então achar o score resultado para ser analisado, “o score do alinhamento global é uma medida que permite quantificar a similaridade entre sequências e fornecer informações valiosas sobre a conservação evolutiva e funcionalidade das moléculas biológicas.”["Bioinformatics: Sequence and Genome Analysis" de David W. Mount.”]

Após isso, faremos a síntese dessas sequências de nucleotídeos em proteínas com uma sequência de aminoácidos para determinar a proteína resultante dessa síntese.E por último determinaremos a estrutura secundária do aminoácido analisado.

3.Metodologia

3.1 Descrição das Informações

Para realizar a descrição das informações de cada registro, utilizamos a biblioteca Biopython e suas funções apropriadas para extrair os dados relevantes de cada sequência de nucleotídeos, como o organismo de origem e a descrição da fonte da sequência. As informações foram obtidas a partir dos metadados presentes nos arquivos FASTA baixados do GenBank.

3.2 Leitura dos Arquivos Fasta e Tamanho das Sequências

Efetuamos a leitura dos arquivos FASTA que contêm as sequências de nucleotídeos utilizando a biblioteca Biopython com as funções adequadas para acessar as sequências presentes nos arquivos e armazenamos os dados em variáveis apropriadas. Em seguida, determinamos o tamanho de cada sequência, ou seja, o número de nucleotídeos contidos em cada uma delas.

3.3 Gráfico de barras com a frequência dos nucleotídeos de cada registro

Por meio da biblioteca Biopython, realizamos a contagem da frequência de cada nucleotídeo (A, C, G, T) em cada sequência de nucleotídeos. Utilizamos essas contagens para criar gráficos de barras, os quais possibilitam visualizar as frequências relativas de cada nucleotídeo em cada sequência. Para facilitar a análise comparativa, organizamos os nucleotídeos em ordem alfabética.

3.4 Cálculo da temperatura de melting (Tm) de cada sequência

Identificamos, por meio da biblioteca Biopython, a função específica que fornece o conteúdo de GC (porcentagem de nucleotídeos guanina e citosina) em uma sequência de nucleotídeos. Em seguida, desenvolvemos um código em Python para calcular a temperatura de melting (Tm) de cada sequência. A Tm é um parâmetro relevante na biologia molecular, especialmente na técnica de PCR, pois indica a temperatura na qual ocorre a desestabilização da dupla hélice de DNA e a separação das cadeias.

3.5 Alinhamento global de pares das primeiras 1200 nucleotídeos das sequências

Por meio da biblioteca Biopython, realizamos o alinhamento global entre todas as combinações possíveis das primeiras 1200 bases nucleotídicas das sequências. Essa abordagem nos permitiu comparar as sequências entre si e identificar regiões conservadas e variantes.

3.6 Síntese proteica das sequências e frequência de aminoácidos

Com auxílio da biblioteca Biopython, realizamos a tradução das sequências de nucleotídeos para sequências de aminoácidos, ou seja, a síntese proteica. Utilizando a função apropriada, contabilizamos a frequência de cada aminoácido nas sequências e criamos gráficos de barras para visualizar as frequências relativas. Ordenamos os aminoácidos alfabeticamente, facilitando a análise comparativa.

3.7 Determinação da estrutura secundária das proteínas das sequências

Utilizando a função "*proteinanalysis*" específica da biblioteca Biopython, foi possível determinar a estrutura secundária das proteínas resultantes da tradução das sequências de nucleotídeos. A estrutura secundária é um aspecto importante para compreender a função e as características das proteínas. Analisamos os resultados obtidos e discutimos as principais características da estrutura secundária das proteínas nas sequências em questão.

4.Resultados

4.resultados

1. As descrições foram registradas em uma seção específica do relatório final, fornecendo uma visão geral das características de cada sequência analisada.

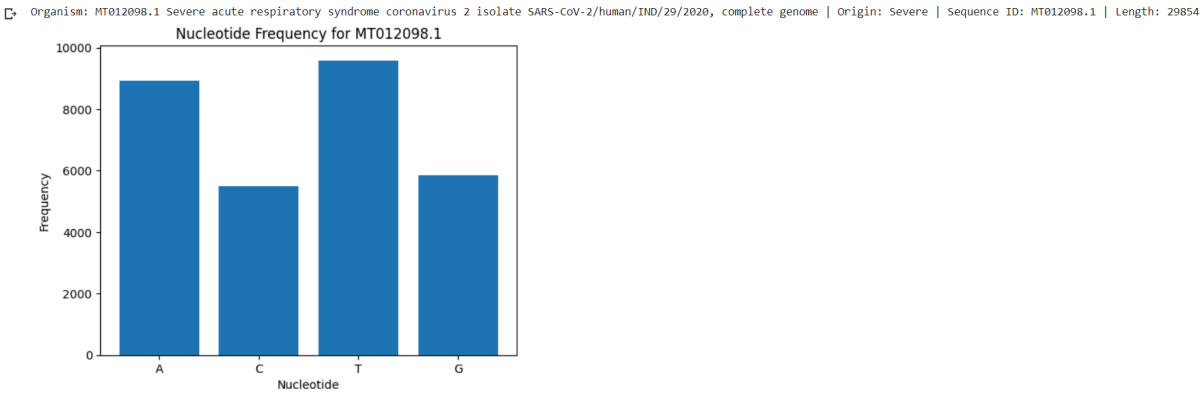

Imagem 1: mostra a classificação do organismo e a frequência dos nucleotídeos

Organism: MN908947.3 Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome | Origin: Severe | Sequence ID: MN908947.3 | Length: 29903

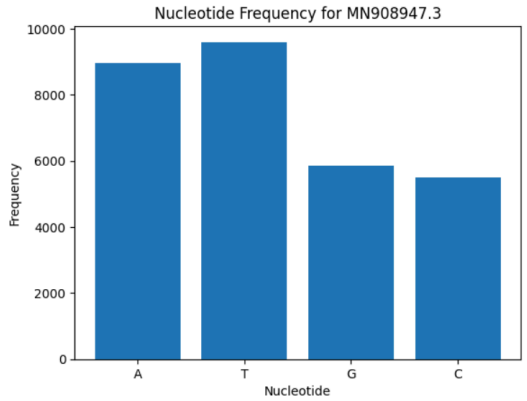

Imagem 2: mostra a classificação do organismo e a frequência dos nucleotídeos

Organism: MT324062.1 Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/ZAF/R03006/2020, complete genome | Origin: Severe | Sequence ID: MT324062.1 | Length: 299

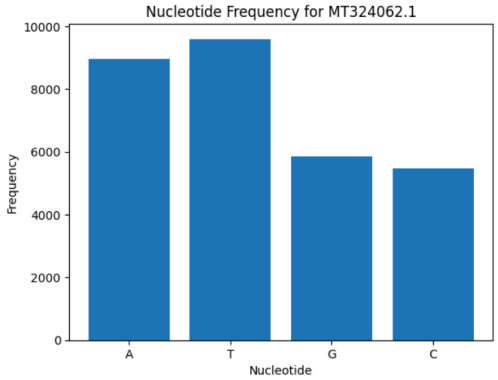

Imagem 3:mostra a classificação do organismo e a frequência dos nucleotídeos

Organism: MZ264787.1 Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/BRA/CD1739-P4/2020, complete genome | Origin: Severe | Sequence ID: MZ264787.1 | Length:

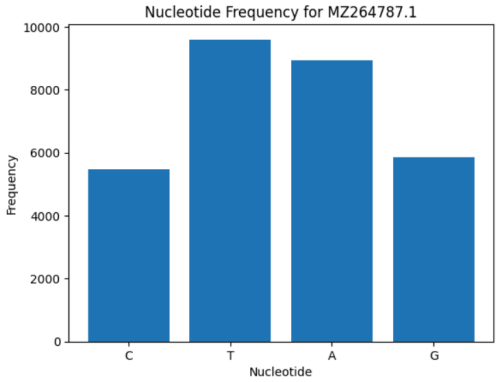

Imagem 4: mostra a classificação do organismo e a frequência dos nucleotídeos

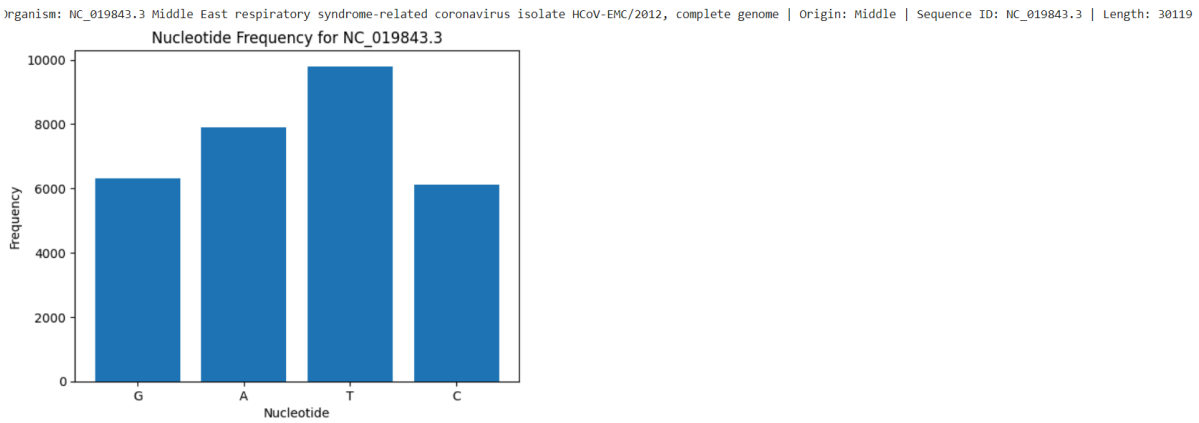

Imagem 5: mostra a classificação do organismo além de sua frequência de nucleotídeos

2. A temperatura de melting é a temperatura na qual mais da metade dos indicadores estão anelados a fita de DNA e a outra solução livre, como pcr é uma técnica que consiste em ampliar em milhares ou bilhares de vezes uma amostra coletada, que será colocada na máquina então essa temperatura de melting é um valor desejado já que ela consegue fazer essa separação da fita de DNA e permitir ele se duplicar.

```
75.47033022230951
```

Figura 1: Temperatura de melting da sequência MT012098.1(sequência 2)

```
75.4516764424424
```

Figura 2: Temperatura de melting da sequência MN908947.3(sequência 1)

```
75.44619204288382
```

Figura 3: Temperatura de melting da sequência MT324062.1(sequência 3)

```
75.44622893195246
```

Figura 4: Temperatura de melting da sequência MZ264787.1 (sequência 4)

```
76.78991691310809
```

Figura 5: Temperatura de melting da sequência NC_019843.3(sequência 5)

3. Os métodos de alinhamentos globais envolvendo sequências grandes é um problema para o qual inventaram diversos métodos inclusive de programação inteira e metaheurística para realizá-los sem muito custos computacionais como o blast. No projeto foi usado um código de alinhamento global de sequências 2 a 2 da biblioteca biopython.

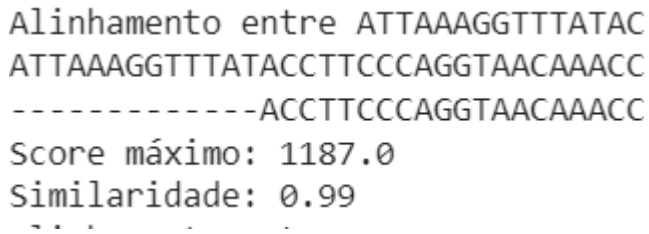

Figura 6: Alinhamento entre a sequência 1 e a 2 no qual percebe-se que a similaridade é bem próxima por se tratar de cepas próximas, ou seja, são organismos praticamente idênticos.

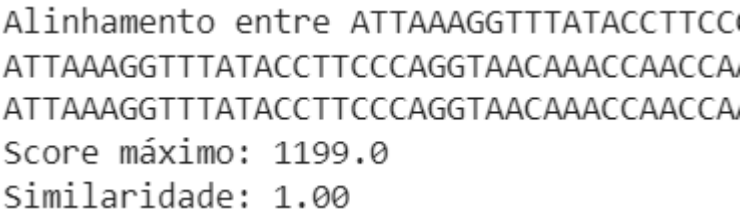

Figura 7: Alinhamento da sequência 1 com a sequência 3 e de novo mostra a similaridade muito alta novamente pois se trata do mesmo vírus o que mostra a o score máximo muito alto.

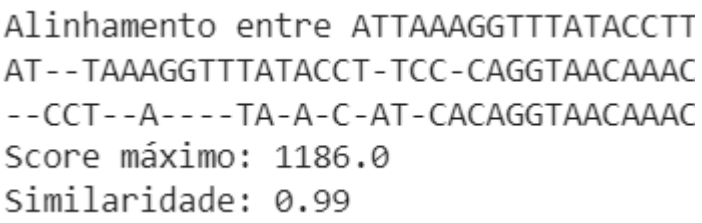

Figura 8: Denovo um alinhamento com similaridade muito alta por de novo ser da mesma espécie de vírus que é o Covid o alinhamento entre a sequência 1 e 4 tem alta similaridade.

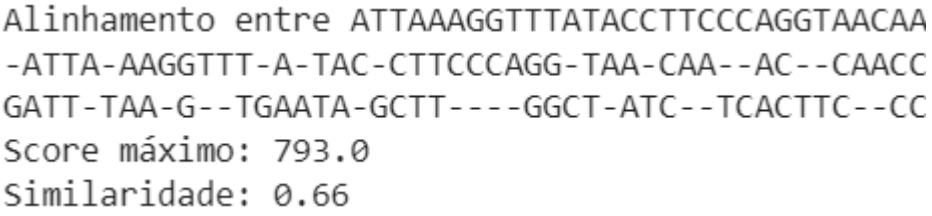

Figura 9: Agora temos uma similaridade bem menor com a sequência 1 e 5 que o apresentado, isso deve-se ao fato de ser uma variante que já sofreu uma série de mutações e por isso apresenta resultado mais baixo que as outras, mas ainda sim é alto.

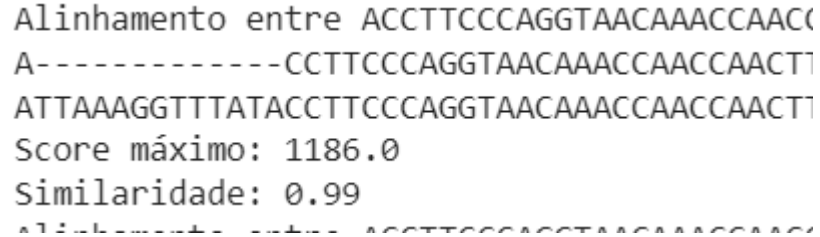

Figura 10: O alinhamento de 2 com 3 é bem alto também, mostrando que os vírus embora tenham sofrido mutação continuam praticamente iguais entre si.

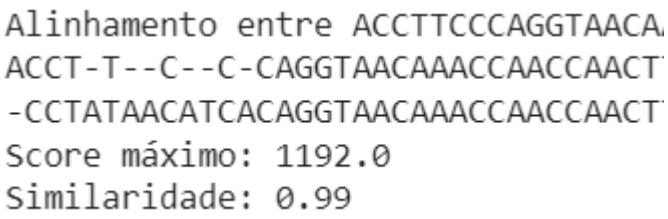

Figura 11: Novamente os resultados do alinhamento entre sequência 2 e 4 são altos mostrando a grande similaridade das sequências.

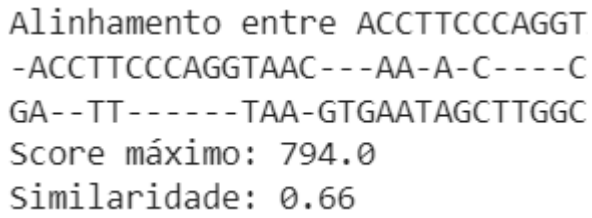

Figura 12: Agora com o alinhamento entre as sequências 2 e 5 apresenta algo parecido com o que aconteceu com a 1 e 5, ou seja por ser uma variante mais nova apresenta um grau de mutação maior o que reflete nessa similaridade baixa em relação às outras.

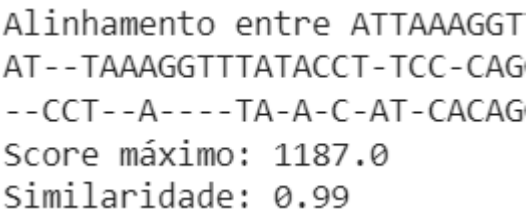

Figura 13: Novamente o alinhamento entre a sequência 3 e 4 apresenta alta similaridade como aconteceu com a 1,2,3,4 em si.

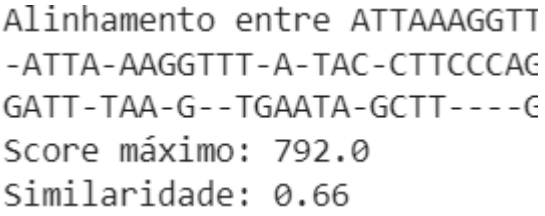

Figura 14: Novamente como esperado entre as sequências 3 e 5 apresenta baixa similaridade como dito anteriormente.

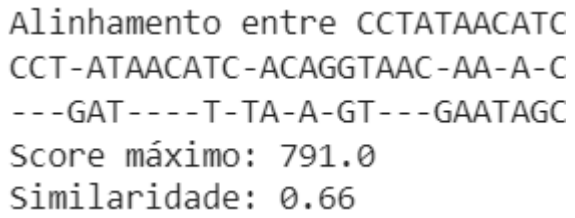

Figura 15: E para comprovar de vez que o alinhamento de 4 e 5 apresenta o mesmo resultado ou seja realmente a mutação alterou as sequências, mas mesmo assim ainda continua alta a similaridade.

4. As proteínas traduzidas foram usadas novamente os métodos da biblioteca do python.

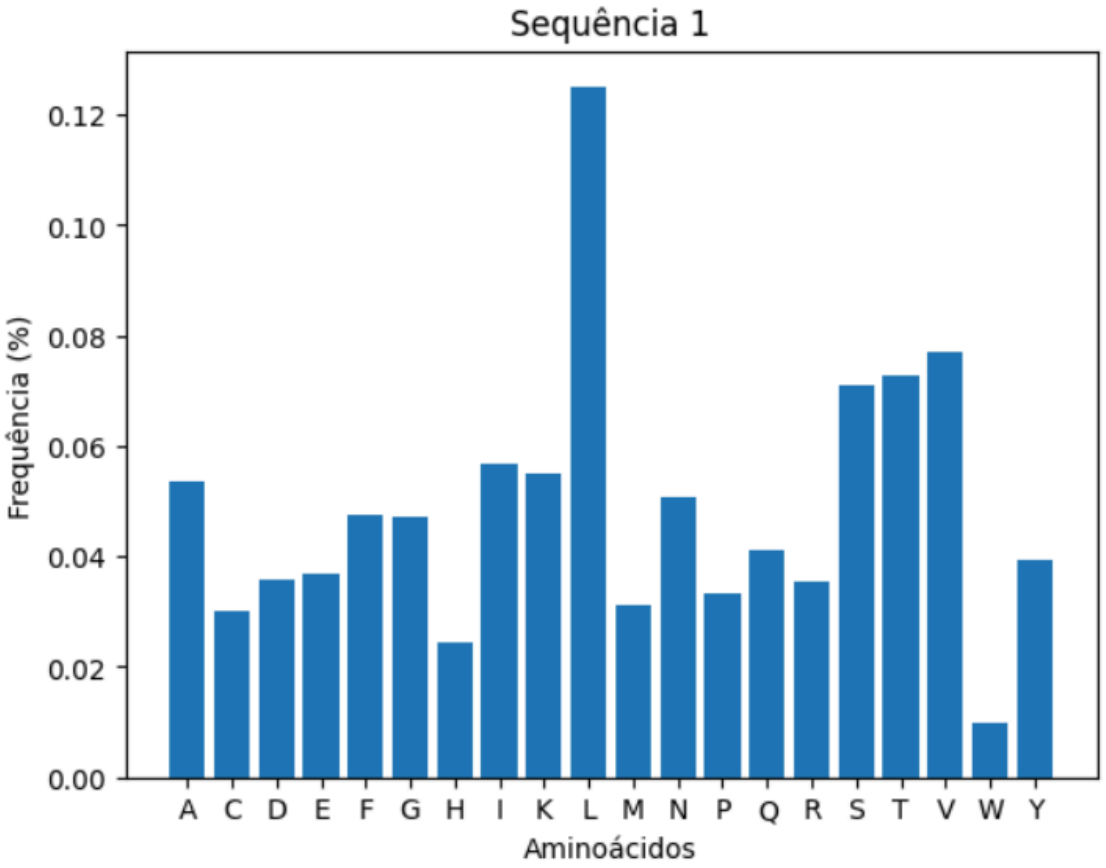

Imagem 6: mostra os aminoácidos produzidos pela sequência MT012098.1.

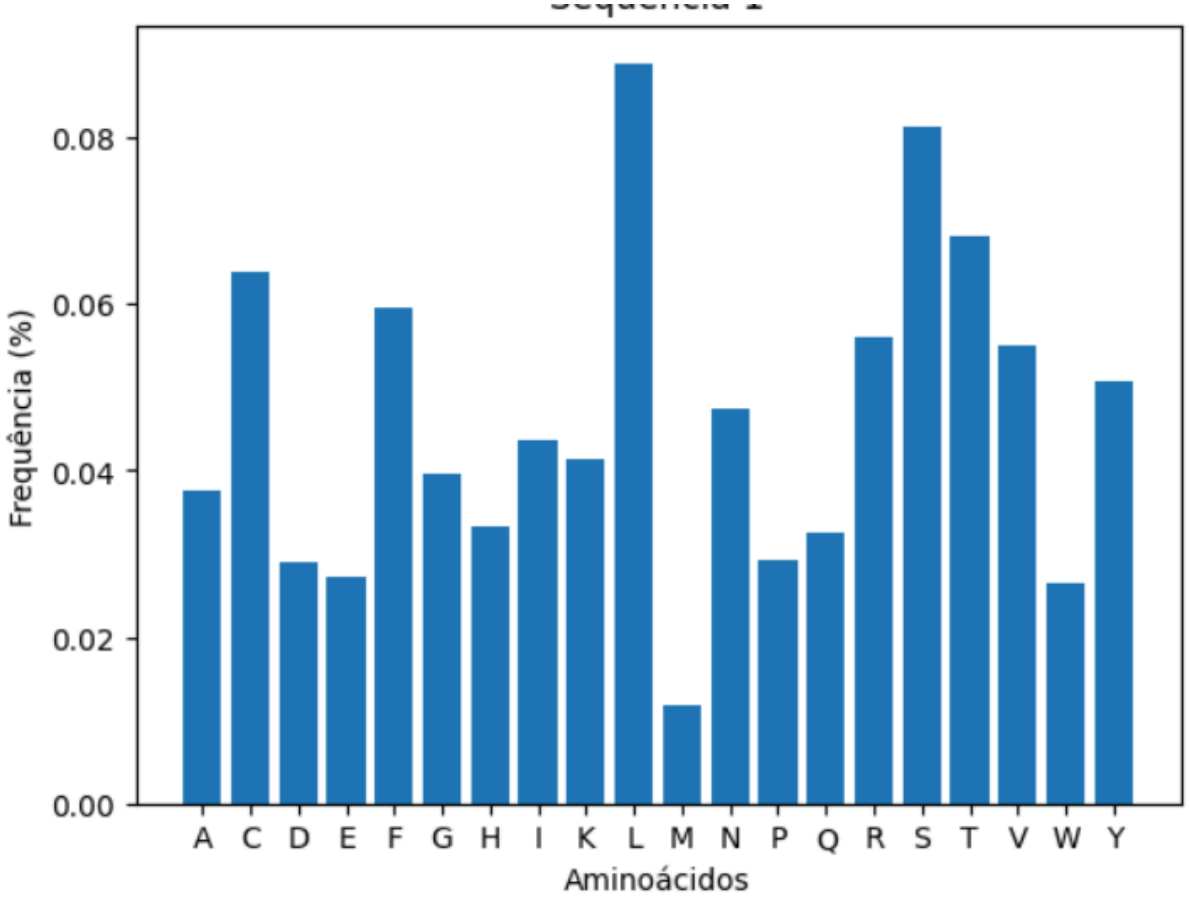

Imagem 7: mostra os aminoácidos produzidos pela sequência MN908947.3.

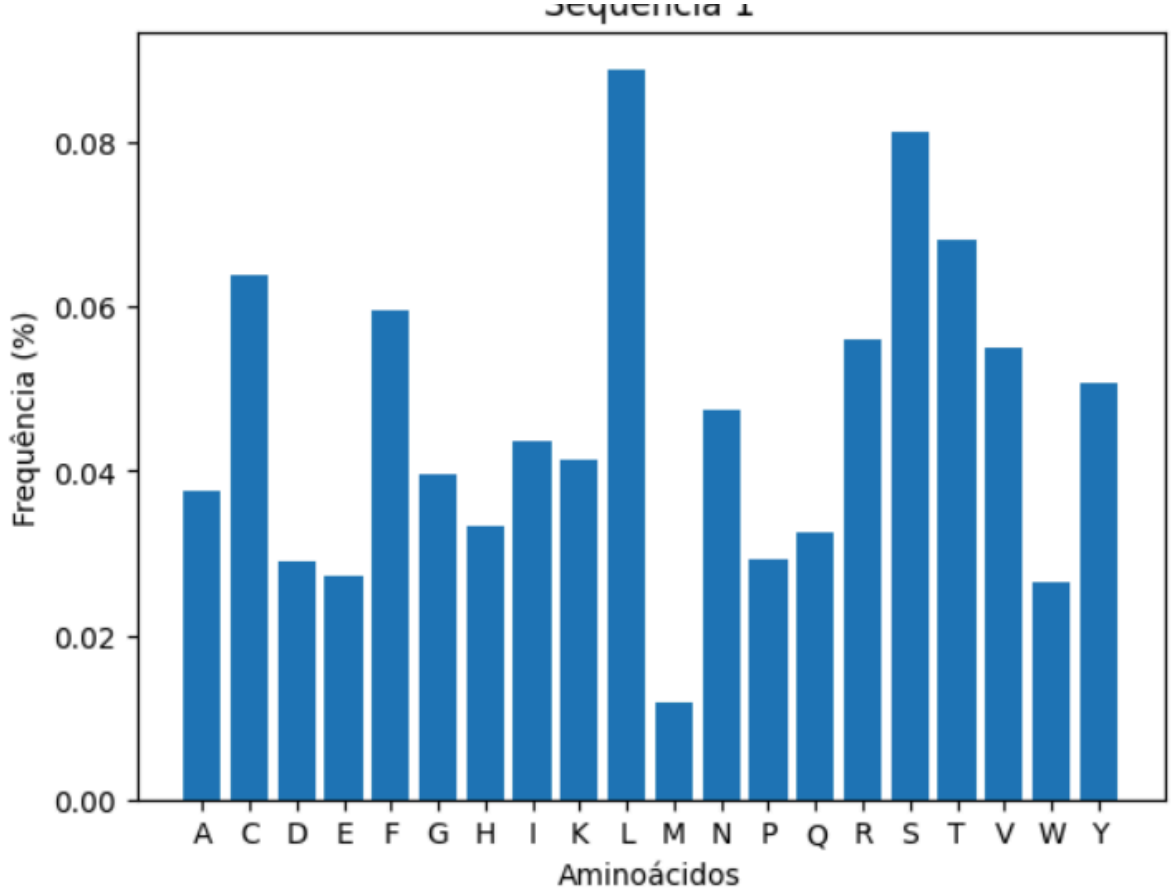

Imagem 8: mostra os aminoácidos produzidos pela sequência MT324062.1.

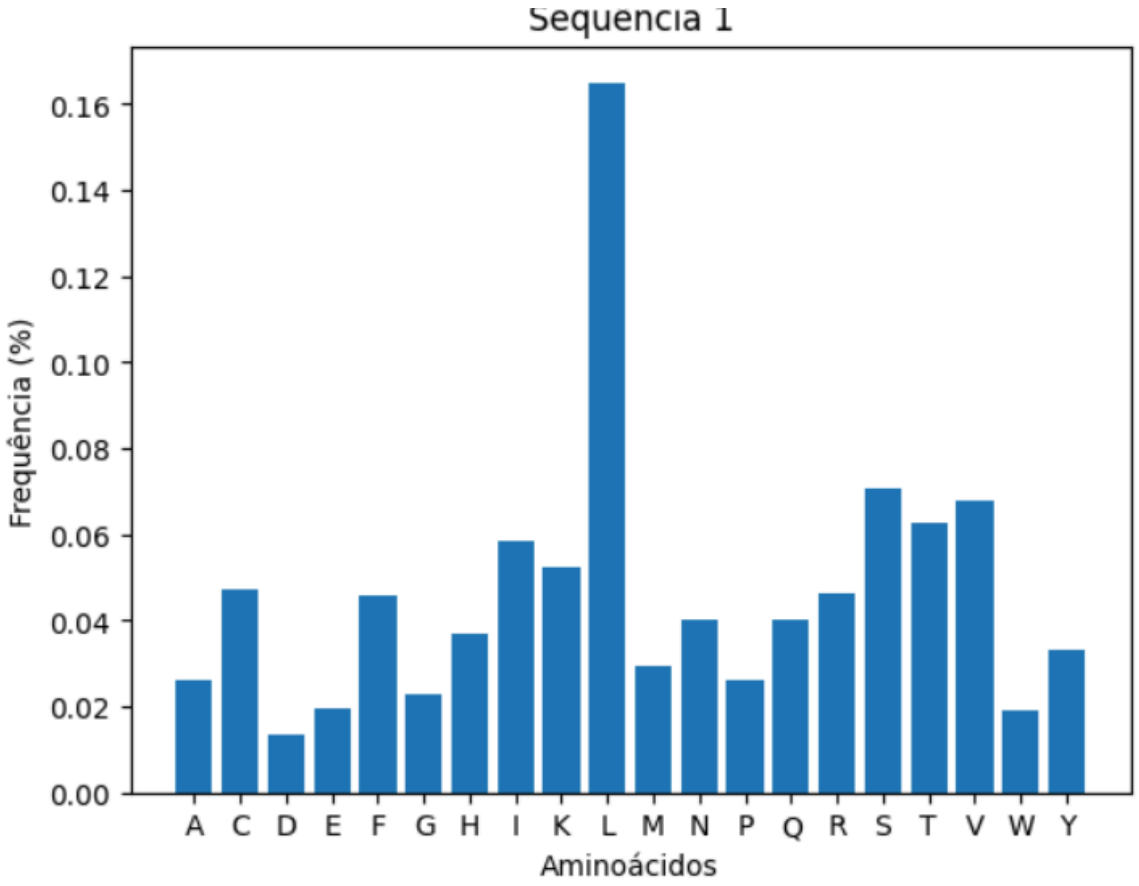

Imagem 9: mostra os aminoácidos produzidos pela sequência MZ264787.1

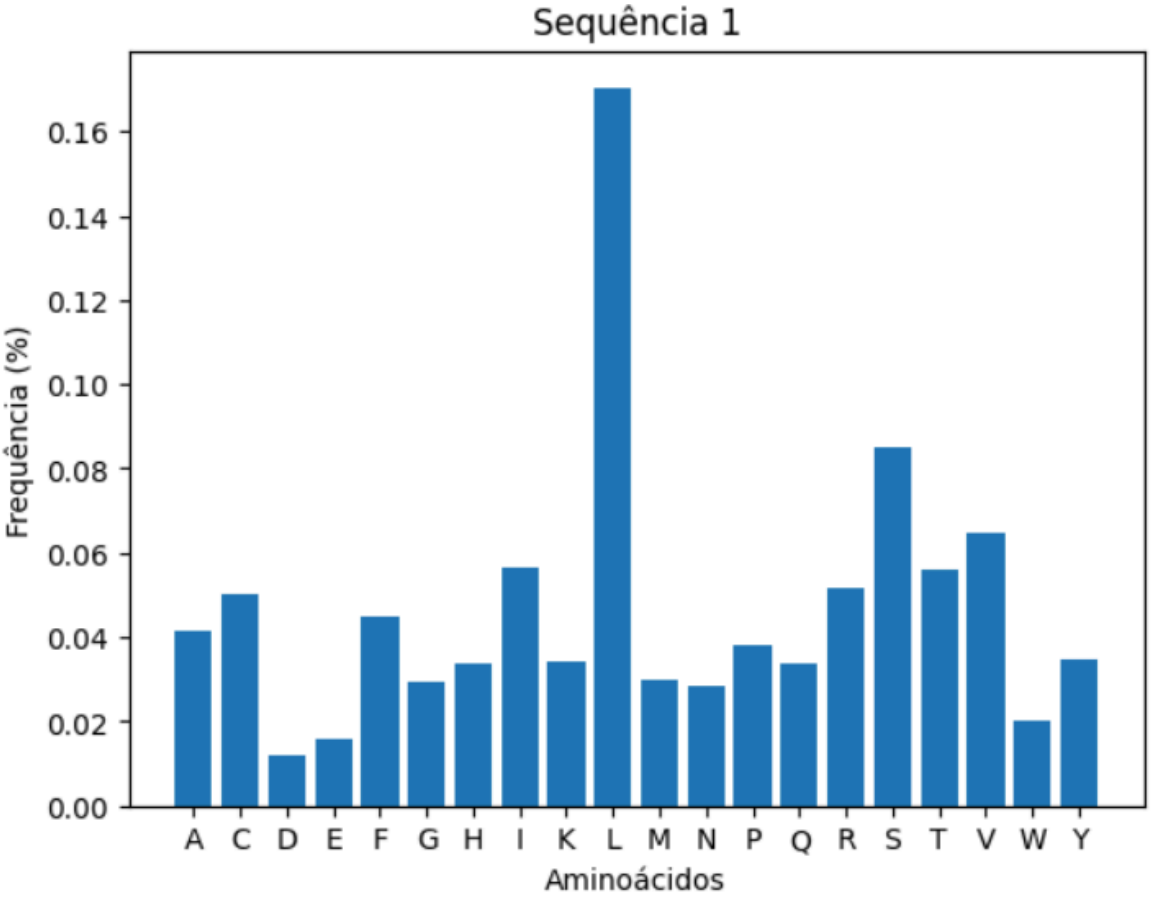

Figura 10: mostra os aminoácidos produzidos pela sequência NC_019843.3.

5. E em relação ao a estrutura secundária dos aminoácidos foi usado uma função da biblioteca do python que identifica a estrutura secundária, no qual será analisado a estrutura de hélice, a folha beta e a volta que é a última das formas secundárias.

```
Total de proteínas contidas na sequência 1: 9951
Estrutura secundária da proteína da sequência 1:
Helix: 0.36
Turn: 0.20
Sheet: 0.25
```

Figura 16: a figura mostra a porcentagem de cada uma das estruturas possíveis para a sequência MT012098.1.

```
Total de proteínas contidas na sequência 1: 9967
Estrutura secundária da proteína da sequência 1:
Helix: 0.32
Turn: 0.20
Sheet: 0.17
```

Figura 17: a figura mostra a porcentagem de cada uma das estruturas possíveis para a sequência MN908947.3.

```
Total de proteínas contidas na sequência 1: 9967
Estrutura secundária da proteína da sequência 1:
Helix: 0.32
Turn: 0.20
Sheet: 0.17
```

Figura 18: a figura mostra a porcentagem de cada uma das estruturas possíveis para a sequência MT324062.1.

```
Total de proteínas contidas na sequência 1: 9955
Estrutura secundária da proteína da sequência 1:
Helix: 0.39
Turn: 0.16
Sheet: 0.24
```

Figura 19: a figura mostra a porcentagem de cada uma das estruturas possíveis para a sequência MZ264787.1

```
Total de proteínas contidas na sequência 1: 10039
Estrutura secundária da proteína da sequência 1:
Helix: 0.39
Turn: 0.18
Sheet: 0.26
```

Figura 20: a figura mostra a porcentagem de cada uma das estruturas possíveis para a sequência NC_019843.3.

5.Conclusão

A partir da análise feita pelo algoritmo é possível concluir o nível de similaridade dessas sequência de nucleotídeos, pois o que se percebe é que elas têm propriedades físicas parecidas e uma estrutura secundária semelhante, o que faz total sentido uma vez que são amostras da Covid-19 coletados em lugares diferentes, mas como são decorrentes de um mesmo organismo sua mutação deve ser pouca, uma vez que não teve tanto tempo para sofrer o processo evolutivo. Isso é mostrado no alinhamento global resultado, além de perceber-se um alto score em todos os testes de alinhamento e perceba também que a temperatura de melting é próxima.

6.Referência

1. MOUNT, David W. **Bioinformatics: Sequence and Genome Analysis**. [s.l.]: CSHL Press, 2004.