Software de Comparação do Sequenciamento Genético

Luara Teodoro Lima, Paloma Ketnyn Machado, Uriel Gonçalves Paiva

Resumo: *Este relatório tem por finalidade explicitar a forma em que o projeto foi formulado, o funcionamento e exibir didaticamente explicando a forma como ele funciona. O mesmo se trata de um software que foi produzido em linguagem python 3, para mostrar a diferenciação de dna entre dois indivíduos, o programa é lido de códon em códon e mostra onde ocorre a diferenciação por bases nitrogenadas, foi-se obtido uma boa resposta mostrando um funcionamento sem erros. Neste relatório também é exibido um software similar (BLAST), onde foi informado como ele funciona e a diferenciação com o projeto desenvolvido pelos autores.*

Palavras chaves: *dna, bases nitrogenadas, software, python*

1. Introdução

O Ácido Desoxirribonucleico (DNA) é uma molécula de extrema importância considerado como portador da mensagem genética pois é nele que estão armazenadas todas as características de um ser vivo, que são únicas em cada indivíduo [1]. Desta forma, é através do código genético que se torna possível observar as particularidades físicas e fisiológicas de cada indivíduo, já que estas características, ou fenótipos, estão diretamente relacionadas à disposição e organização dos nucleotídeos através das fitas ou cadeias polinucleotídicas. Individualmente, cada nucleotídeos é formado por uma pentose (monossacarídeos formados por cinco átomos de carbono) chamada de desoxirribose, um grupo fosfato (HPO4) e bases nitrogenadas, que podem ser adenina (A), citosina (C), guanina (G) e timina (T). De modo que o DNA é constituído por duas fitas em espiral que se ligam através de pontes de hidrogênio entre as bases nitrogenadas, desenvolvendo o formato de dupla hélice.

Ou seja, código genético nada mais é que sequências de bases nitrogenadas que posteriormente serão traduzidos em aminoácidos. A combinação destas bases de três em três chama-se códon e faz com que seja determinado o aminoácido necessário para formação de uma proteína, e, portanto, diferentes combinações de aminoácidos formarão diferentes proteínas.

Desta maneira, estudar a composição do código genético a nível molecular, significa estudar a sequenciação dos códons, pois é a ordem deles que caracteriza as especificidades dos seres, portanto, ao analisá-la é possível apontar as semelhanças e distinções entre diferentes indivíduos ou até mesmo seres de diferentes categorias taxonômicas (espécie, gênero, família, etc) para poder rastrear onde os códigos de cada um se distinguiram ao ponto de atribuir diferentes características significativas.

Tendo em vista a dificuldade de analisar as sequências devido à extensão das cadeias, é possível automatizar esse processo através da criação de um programa que faça a leitura e comparação dos códigos genéticos e retorne quais sequências significativas de códons foram exclusivas à apenas um código, destacando então a parte que o torna diferente do outro.

Por exemplo, ao observar duas bactérias da mesma classe, mas de espécies diferentes, é possível perceber grande semelhança no código genético, porém, após

uma análise sistemática dos nucleotídeos, é possível determinar onde que elas se diferenciam molecularmente ao ponto de causar uma variedade na espécie. No caso, serão analisadas duas bactérias da classe Enterococcus, sendo uma delas conhecida como Enterococcus faecium e a outra Enterococcus faecalis, ambas são bactérias gram-positivas presentes no sistema digestivo e fazem parte da flora intestinal podendo adquirir características patogênicas, causando doenças como infecção do trato urinário, endocardite, infecção intra-abdominal, meningite e entre outras. As sequências dos códigos genéticos serão acessadas a partir do banco de dados NCBI - Centro Nacional de Informações Biotecnológicas, um GenBank público que armazena sequenciações de genomas e disponibiliza gratuitamente [1].

2. Objetivo

2.1. Objetivo geral

- Desenvolver um software que forneça a diferença no sequenciamento genético entre dois microorganismos de mesma classe.

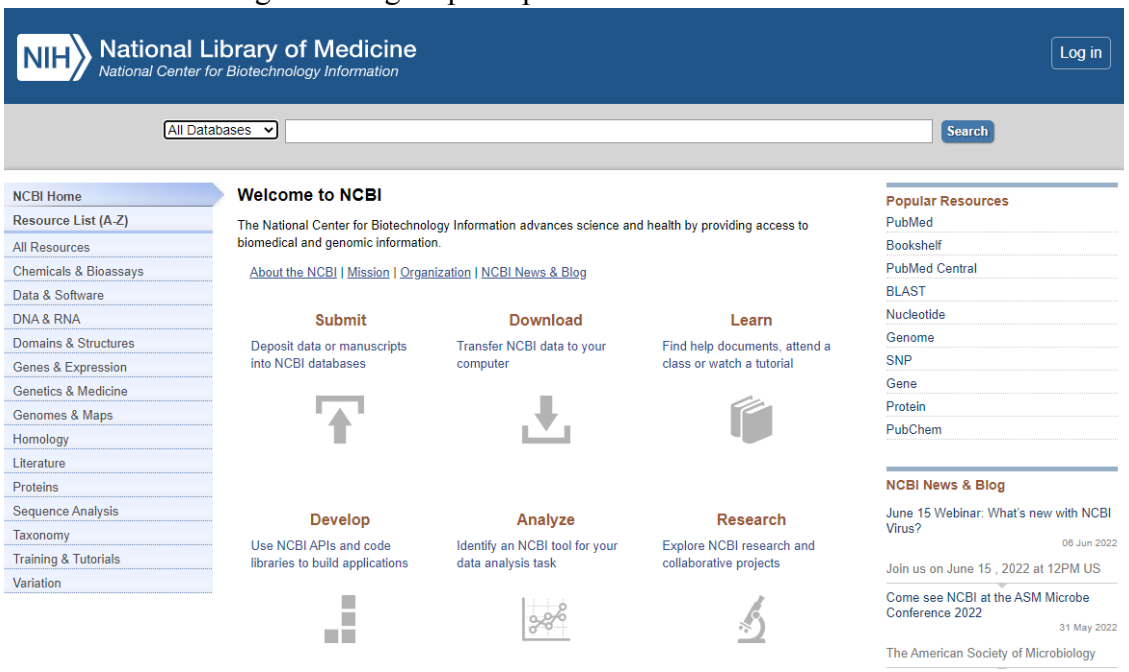
2.2. Objetivo específico

- Leitura das informações do banco de dados internacional **NCBI .**
- Análise e comparação da sequência de bases do DNA de dois códigos genéticos diferentes.
- Filtrar e expor as diferenças encontradas.

3. Metodologia

A partir dos banco de dados disponibilizados no GenBank NCBI (Figura 1), é possível encontrar o código genético das bactérias a serem estudadas**:** E. faecium e E. faecalis.

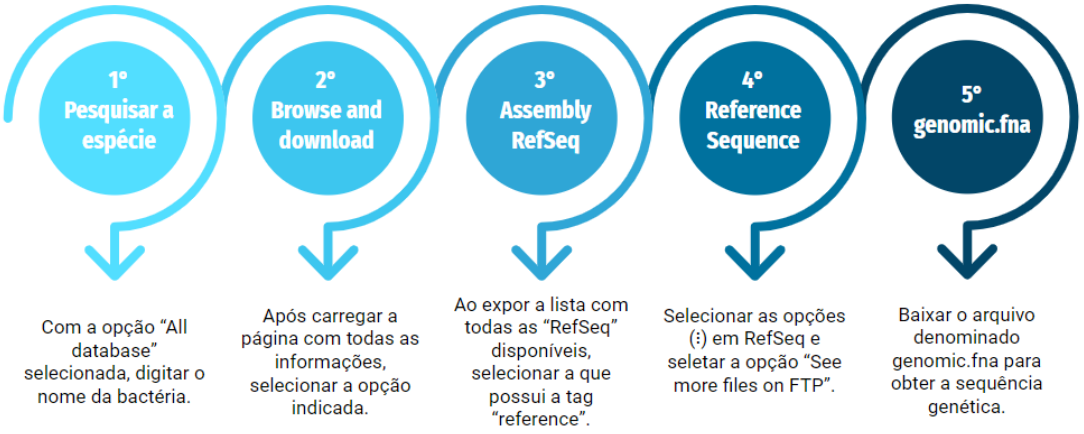
Figura 1: Página principal do banco de dados NCBI.



Fonte**:** Os autores

Para obter o código genético do microrganismo,segue ao esquema ilustrando:

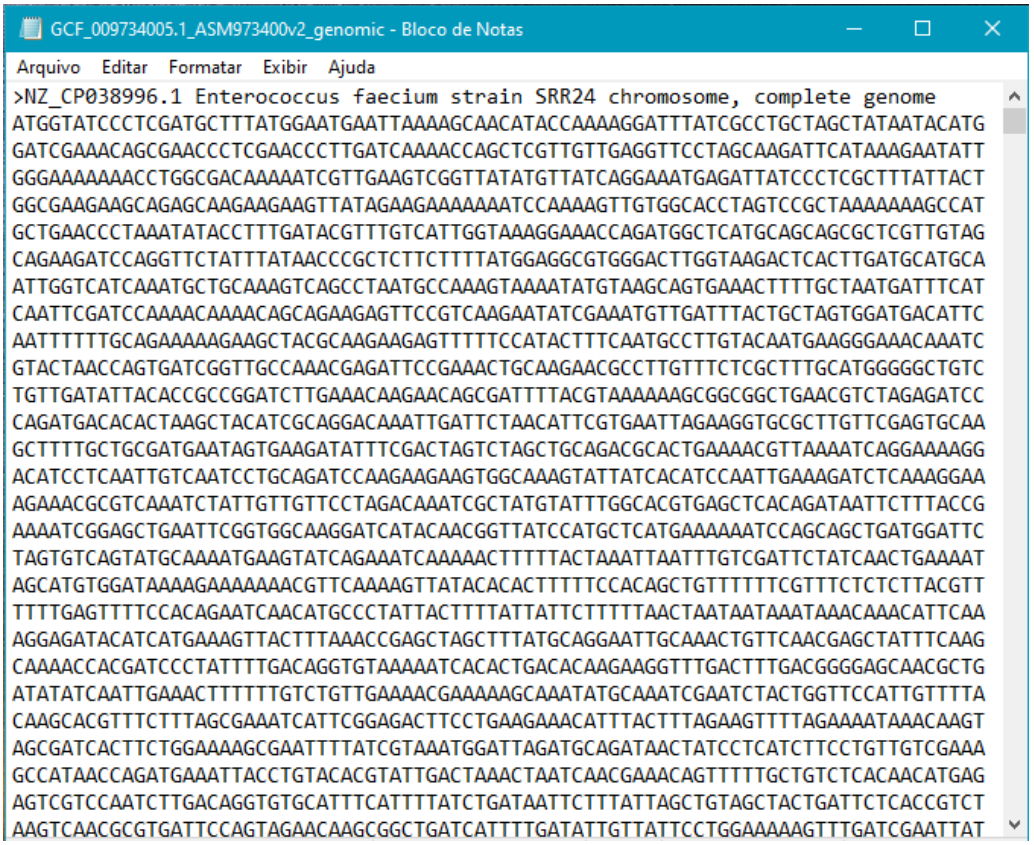
Figura 2: Esquema para baixar a sequência genética



Fonte: Os autores

A partir do passo a passo demonstrado, o resultado é um arquivo de texto de extensão .fna ou .txt com toda a sequenciação genética do microrganismo, no caso, os dois arquivos devem ser salvos em um diretório de escolha do usuário para que, posteriormente, sejam analisados pelo software desenvolvido de análise e comparação que realiza a leitura dos arquivos baixados do GenBank e os correlaciona.

Figura 3: Arquivo obtido do sequenciamento da bactéria E. faecium


GCF_009734005.1_ASM973400v2_genomic - Bloco de Notas

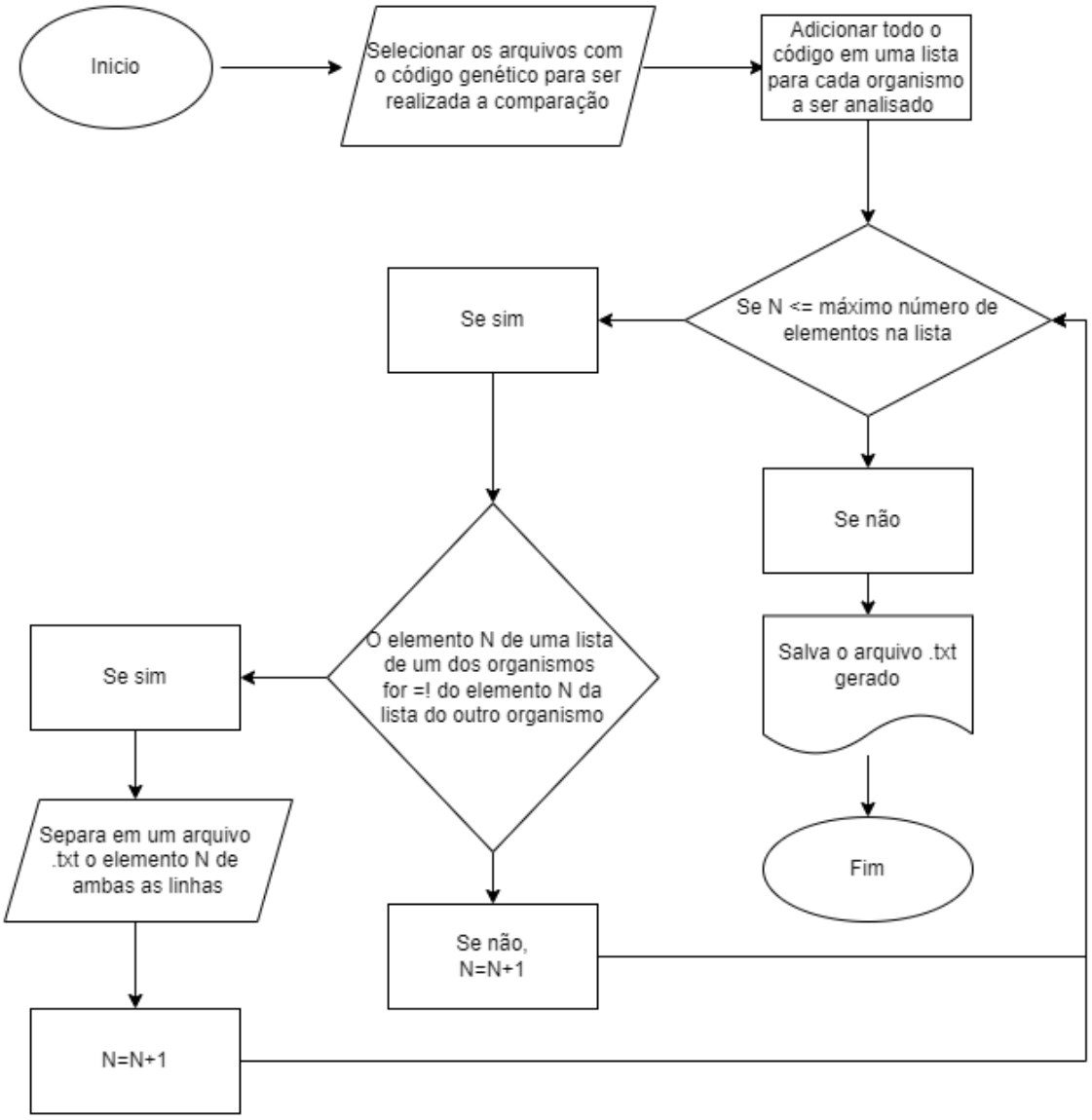
Arquivo Editar Formatar Exibir Ajuda

```
>NZ_CP038996.1 Enterococcus faecium strain SRR24 chromosome, complete genome
ATGGTATCCCTCGATGCTTTATGGAATGAATTAAAAGCAACATACCAAAAGGATTTATCGCCTGCTAGCTATAATACATG
GATCGAAACAGCGAACCCTCGAACCCTTGATCAAAACCAGCTCGTTGTTGAGGTTCCTAGCAAGATTCATAAAGAATATT
GGGAAAAAAAACCTGGCGACAAAAATCGTTGAAGTCGGTTATATGTTATCAGGAAATGAGATTATCCCTCGCTTTATTACT
GGCGAAGAAGCAGAGCAAGAAGAAGTTATAGAAGAAAAAAATCCAAAAGTTGTGGCACCTAGTCCGCTAAAAAAAGCCAT
GCTGAACCCTAAATATACCTTTGATACGTTTGTCATTGGTAAAGGAAACCAGATGGCTCATGCAGCAGCGCTCGTTGTAG
CAGAAGATCCAGGTTCTATTTATAACCCGCTCTTCTTTTATGGAGGCGTGGGACTTGGTAAGACTCACTTGATGCATGCA
ATTGGTCATCAAATGCTGCAAAGTCAGCCTAATGCCAAAGTAAAATATGTAAGCAGTGAAACTTTTGCTAATGATTTCAT
CAATTCGATCCAAAACAAAACAGCAGAAGAGTTCCGTCAAGAATATCGAAATGTTGATTTACTGCTAGTGGATGACATTC
AATTTTTTGCAGAAAAAGAAGCTACGCAAGAAGAGTTTTTCCATACTTTCAATGCCTTGTACAATGAAGGGAAACAAATC
GTACTAACCAGTGATCGGTTGCCAAACGAGATTCCGAAACTGCAAGAACGCCTTGTTTCTCGCTTTGCATGGGGGCTGTC
TGTTGATATTACACCGCCGGATCTTGAAACAAGAACAGCGATTTTACGTAAAAAAGCGGCGGCTGAACGTCTAGAGATCC
CAGATGACACACTAAGCTACATCGCAGGACAAATTGATTCTAACATTCGTGAATTAGAAGGTGCGCTTGTTCGAGTGCAA
GCTTTTGCTGCGATGAATAGTGAAGATATTTCGACTAGTCTAGCTGCAGACGCACTGAAAACGTTAAAATCAGGAAAAGG
ACATCCTCAATTGTCAATCCTGCAGATCCAAGAAGAAGTGGCAAAGTATTATCACATCCAATTGAAAGATCTCAAAGGAA
AGAAACGCGTCAAATCTATTGTTGTTCCTAGACAAATCGCTATGTATTTGGCACGTGAGCTCACAGATAATTCTTTACCG
AAAATCGGAGCTGAATTCGGTGGCAAGGATCATACAACGGTTATCCATGCTCATGAAAAAATCCAGCAGCTGATGGATTC
TAGTGTCAGTATGCAAAATGAAGTATCAGAAATCAAAAACTTTTTACTAAATTAATTTGTCGATTCTATCAACTGAAAAT
AGCATGTGGATAAAAGAAAAAAACGTTCAAAAGTTATACACACTTTTTCCACAGCTGTTTTTTCGTTTCTCTCTTACGTT
TTTTGAGTTTTCCACAGAATCAACATGCCCTATTACTTTTATTATTCTTTTTAACTAATAATAAATAAACAAACATTCAA
AGGAGATACATCATGAAAGTTACTTTAAACCGAGCTAGCTTTATGCAGGAATTGCAAACTGTTCAACGAGCTATTTCAAG
CAAAACCACGATCCCTATTTTGACAGGTGTAAAAATCACACTGACACAAGAAGGTTTGACTTTGACGGGGAGCAACGCTG
ATATATCAATTGAAACTTTTTTGTCTGTTGAAAACGAAAAAGCAAATATGCAAATCGAATCTACTGGTTCCATTGTTTTA
CAAGCACGTTTCTTTAGCGAAATCATTCGGAGACTTCCTGAAGAAACATTTACTTTAGAAGTTTTAGAAAATAAACAAGT
AGCGATCACTTCTGGAAAAGCGAATTTTATCGTAAATGGATTAGATGCAGATAACTATCCTCATCTTCCTGTTGTCGAAA
GCCATAACCAGATGAAATTACCTGTACACGTATTGACTAAACTAATCAACGAAACAGTTTTTGCTGTCTCACAACATGAG
AGTCGTCCAATCTTGACAGGTGTGCATTTCATTTTATCTGATAATTCTTTATTAGCTGTAGCTACTGATTCTCACCGTCT
AAGTCAACGCGTGATTCCAGTAGAACAAGCGGCTGATCATTTTGATATTGTTATTCCTGGAAAAAGTTTGATCGAATTAT
```

Fonte: Os autores

O código será desenvolvido em linguagem Python 3, a qual é comum no quesito acadêmico podendo facilmente ser reproduzidas por diversos programas e IDLEs. No caso, será utilizado o programa Visual Studio para montar o seguinte algoritmo demonstrado abaixo. No caso, o inicializador do programa com o código deve ser salvo na mesma pasta que os arquivos que serão analisados.

Figura 4: Fluxograma

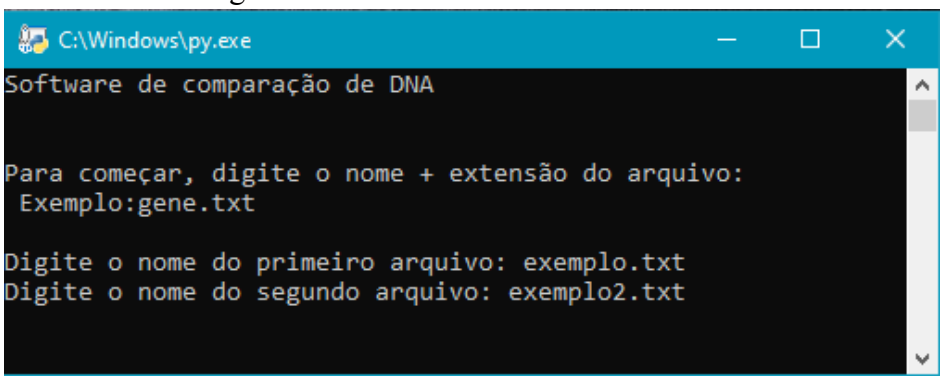


Fonte: Os autores

4. Resultados

Ao inicializar o código, aparece a janela inicial (Figura 5) pedindo o nome e a extensão do primeiro e do segundo arquivo, por isso devem estar no mesmo diretório que o programa, para que sejam facilmente localizados e consecutivamente lidos pelo programa para que possa ser realizada a comparação entre os dois. No caso, o primeiro arquivo será o código da bactéria E. faecium, e o segundo a bactéria E. faecalis.

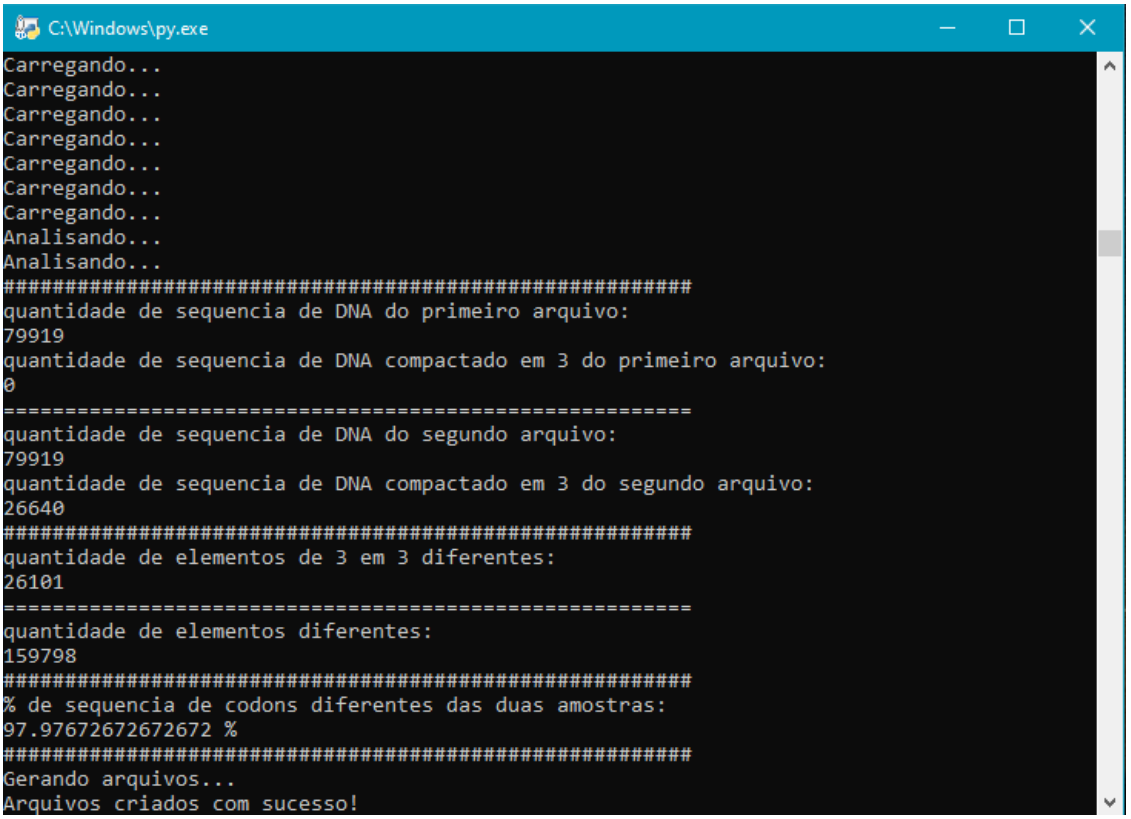
Figura 5: Janela inicial do software criado



Fonte: Os autores

Após selecionar separadamente os dois arquivos dos códigos genético com a sequência das bases nitrogenadas e executar o programa, o usuário recebe as seguintes informações:

Figura 6: Resposta obtida pelo software



Fonte: Os autores

Ou seja, durante a comparação, é informado todas a informações quantitativas ao longo do processo, como a quantidade total de bases nitrogenadas de cada arquivo, no caso ambos possuem 79.919 pois foi utilizado uma região encurtada do código para fins demonstrativos, a quantidade das trincas de DNA, compactando as bases de 3 em 3, a quantidade de elementos diferentes entre os dois arquivos e a porcentagem dessa diferença. Além de anunciar essas informações, o programa cria também 3 arquivos mostrando as comparações, sendo elas:

4.1. Análises de 3 em 3

Neste caso, é realizado a contabilização e comparação de códon a códon, por exemplo: o primeiro códon do primeiro arquivo com o primeiro códon do segundo arquivo; o segundo códon do primeiro arquivo com o segundo códon do segundo arquivo e assim sucessivamente. Dessa maneira, é possível realizar a comparação de áreas mais específicas.

Figura 6 e 7: Código para realizar a comparação de cada códon e arquivo criado a partir dele

```
    ff=len(LG110F)
    ff2=len(LG21F)
    if q<q2:
        tt=(len(LG210F)/q2)*100
    else:
       tt=(len(LG110F)/q)*100
    if len(LG110F) == 0:
        LG110F.append("N/D")    ##separação de 3 em 3
        LG210F.append("N/D")
        R.append("N/D")
#################################################################
    arq=open("analise_de_3_em_3_{}.txt".format(g),"w")
    arq.write(str("codon"+"\t"+"DNA1"+"\t"+"DNA2"+"\n"))
    while h<len(LG110F):
        arq.write(str(R[h]+"\t"+LG110F[h]+"\t"+LG210F[h]+"\n"))
        h=h+1
    arq.close()
```

codon	DNA1	DNA2
1	ATG	GTT
2	GTA	CAT
3	TCC	TGA
4	CTC	AAA
5	GAT	CTG
6	GCT	GAT
7	TTA	ATT
8	TGG	GAA
9	AAT	GTA
10	GAA	AAA
11	TTA	AGA
12	AAA	ATC

Fonte: Os autores

Observação: A criação das listas foi realizada anteriormente no programa.

4.2. Análise de 9 em 9

Essa comparação utiliza o mesmo princípio citado no item 4.1, porém em vez de comparar cada sequência de trinca, são comparados os trios de trinca, ou seja, um conjunto de 9 bases nitrogenadas

Figura 8 e 9: Parte do código para realizar a comparação a cada 3 códons e arquivo criado a partir dele, respectivamente.

```
#################################################################
arq=open("analise_de_9_em_9_{}.txt".format(g),"w")
arq.write(str("região"+"\t"+"DNA1"+"\t"+"DNA2"+"\n"))
while d<len(LG29):
    arq.write(str(D[d]+"\t"+LG19[d]+"\t"+LG29[d]+"\n"))
    d=d+1
arq.close()
g=g+1
print("Arquivos criados com sucesso!")
```

região	DNA1	DNA2
1	ATGGTATCC	GTTCATTGA
2	GTATCCCTC	CATTGAAAA
3	TCCCTCGAT	TGAAAACTG
4	CTCGATGCT	AAACTGGAT
5	GATGCTTTA	CTGGATATT
6	GCTTTATGG	GATATTGAA
7	TTATGGAAT	ATTGAAGTA
8	TGGAATGAA	GAAGTAAAA
9	AATGAATTA	GTAAAAAGA
10	GAATTAAAA	AAAAGAATC

Fonte: Os autores

4.3. Diferenças totais

Neste caso, foi realizada a comparação com junções de 7 trincas, o programa condensa as bases nitrogenadas em grupos de 21 e compara todos os conjuntos do primeiro arquivo com os conjuntos obtidos do segundo arquivo e adiciona em novo documento a sequência que existe apenas em um arquivo, ou seja, separa as sequências exclusivas de cada arquivo. Por exemplo, o primeiro conjunto do primeiro arquivo é comparado com todos os conjuntos do arquivo dois, se ele se repetir em qualquer região, o programa passa para o segundo conjunto de 21 bases e assim sucessivamente. O processo é repetido para o segundo arquivo 2 em relação ao primeiro.

Figura 10: Parte do código para realizar a comparação dos conjuntos.

```
151     ff=len(LG110F)
152     ff2=len(LG21F)
153     if q<q2:
154         tt=(len(LG210F)/q2)*100
155     if len(LG1TTUF) == 0:
156         LG1TTUF.append("N/D")      ##separação unitaria
157         LG2TTUF.append("N/D")
158         E.append("N/D")
159 ##############################################################################
160     arq=open("analise_de_diferença_{}.txt".format(g),"w")
161     arq.write(str("elemento"+"\t"+"DNA1"+"\t"+"DNA2"+"\n"))
162     while i<len(LG121F):
163         arq.write(str("\t"+LG121F[i]+"\t"+LG221F[i]+"\n"))
164         i=i+1
165     arq.close()
```

Fonte: Os autores

Figura 11: Arquivo criado com os conjuntos de códons exclusivos em cada arquivo.

```
elemento        DNA1    DNA2
        ATGGTATCCCTCGATGCTTTA   GTTCATTGAAAACTGGATATT
        TGGTATCCCTCGATGCTTTAT   TTCATTGAAAACTGGATATTG
        GGTATCCCTCGATGCTTTATG   TCATTGAAAACTGGATATTGA
        GTATCCCTCGATGCTTTATGG   CATTGAAAACTGGATATTGAA
        TATCCCTCGATGCTTTATGGA   ATTGAAAACTGGATATTGAAG
        ATCCCTCGATGCTTTATGGAA   TTGAAAACTGGATATTGAAGT
        TCCCTCGATGCTTTATGGAAT   TGAAAACTGGATATTGAAGTA
        CCCTCGATGCTTTATGGAATG   GAAAACTGGATATTGAAGTAA
        CCTCGATGCTTTATGGAATGA   AAAACTGGATATTGAAGTAAA
        CTCGATGCTTTATGGAATGAA   AAACTGGATATTGAAGTAAAA
        TCGATGCTTTATGGAATGAAT   AACTGGATATTGAAGTAAAAA
```

Fonte: Os autores

Ou seja, no caso da bactéria E. faecium, o primeiro conjunto de códons exclusivo encontrado foi: **ATG GTA TCC CTC GAT GCT TTA**.

No caso da bactéria E. faecalis, o primeiro conjunto exclusivo encontrado foi: **GTT CAT TGA AAA CTG GAT ATT** .

5. Discussão

Um software que contém similaridade com o que foi feito neste projeto é o BLAST. O BLAST, ferramenta básica de alinhamento local [2], é um programa em que foi elaborado para realizar buscas de comparação de sequência biológica primária e apresentar através de um banco de dados o DNA / RNA que contém maior equivalência e maior relevância em relação a sequência genética submetida.

No BLAST é possível realizar a pesquisa das seguintes formas:

- BLASTn: comparação através de nucleotídeos em um banco de dados de nucleotídeos, geralmente é utilizado para identificar sequências mais distantes e identificar sequências desconhecidas geradas de sequenciamento e PCR;
- BLASTp: consulta por meio de proteína - proteína e o resultado é dado por outras pesquisas BLAST, como BLASTx e tBLASTn;
- BLASTx: a pesquisa é realizada através de um nucleotídeo e a resposta é dada a partir de banco de dados da proteínas, esta é mais sensível que o BLASTn por conter comparação a nível proteico;
- tBLASTn: busca proteínas em um banco de dados de nucleotídeos para verificar quais genes estão relacionados a uma determinada proteína;
- tBLASTx: a sequência de nucleotídeos são convertidas em 6 sequência de aminoácidos que serão contrastados as 6 possíveis fases de leitura para identificar o nucleotídeo dentro de um banco de dados.

Este sistema operacional é encontrado dentro do NCBI e coletam informações do banco de dados da plataforma. O funcionamento dele ocorre por meio de seus algoritmos que realizam uma pesquisa fundamentada em alinhamentos locais, sendo eles confiáveis e rápidos. A rapidez dentro deste sistema é acarretada pela varredura por similaridade dentro do algoritmo do BLAST.

Comparando as duas ferramentas, o BLAST e a criada neste projeto, observa-se que o programa realizado pelos autores faz a separação de diferença por linha a lista a cada códon, cada 3 códons e também compara as diferença em geral analisando se existe ou não tal sequência na lista do outro arquivo.

Pensando em como realizar as melhorias neste projeto, pode-se aprimorar na forma de seleção de arquivos a serem utilizados e fazer uma interface mais gráfica, para facilitar e ser mais amigável para a utilização do usuário, conforme utilizando e observando alguns erros serão aperfeiçoados de acordo com a busca de um programas sem erros durante a utilização.

6. Conclusão

Por fim, este relatório tem o objetivo primeiramente explicar o que é o material genético e a parte biológica e informar o processo de criação deste software que tem por objetivo facilitar na varredura para comparação e mostrar as diferenças e semelhanças entre diferentes indivíduos ou seres que possuem somente diferentes categorias taxonômicas. No relatório, foi mostrado passo a passo de como utilizar o programa, após um árduo trabalho, tendo em vista que ele foi gerado do zero, obteve-se um software sem erros e com um bom funcionamento, gerando ao fim dois novos arquivos mostrando a resposta que se pede (a diferenciação entre os dois indivíduos informados ao programa).

7. Referências

[1] ***Frantz, Nilo. DNA: saiba suas principais características e funções.*** *Nilo Frantz 07/2020. Disponível em: https://www.nilofrantz.com.br/dna-caracteristicas-e-funcoes/ || Acessado em: 02/05/2022*

[2] ***NCBI, National Center for Biotechnology of Information.*** *Disponível em: https://www.ncbi.nlm.nih.gov/ || Acessado em: 02/05/2022*

[3] ***BLAST, Basic Local Alignment Search Tool.*** *Disponível em: https://medium.com/omixdata/entendendo-blast-parte-i-conceitos-principais-4711e34cc2b6 || Acessado em: 05/06/2022*