CÁPSULA MOLECULAR DO TEMPO

ARMAZENANDO EM DNA DADOS A LONGO PRAZO

Colégio Militar de Manaus

Autores: OLIVEIRA, A. V. S.[1]; RAMOS, E. I. O.[1]; Orientadores: BARBOSA FILHO, R. A. A.[2]; GURGEL, P. M.[2]

[1]Alunos do Colégio Militar de Manaus

[2]Professores do Colégio Militar de Manaus

Colégio Militar de Manaus

criatividade e inovação FEBRACE 21ª feira brasileira de ciências e engenharia

INTRODUÇÃO

A humanidade sempre buscou formas de registrar sua existência e conhecimentos. Para isso, novas formas de armazenamento de dados continuaram surgindo até os dias atuais, onde a crescente demanda de informações digitais está superando a capacidade dos dispositivos, que além de possuírem uma vida útil curta, geram altos custos para serem mantidos[1]. Vários estudos científicos recentes vêm mostrando a utilidade da molécula de DNA como uma alternativa mais durável, compacta e eficiente para o armazenamento de dados[2], sendo capaz de se manter íntegra por milhares de anos em condições controladas.

A Cápsula Molecular do Tempo desenvolve um método de armazenar informações utilizando um sistema baseado em DNA, que une a linguagem de programação Python com o gene codificante da proteína spike (gene S) das variantes de preocupação (VOCs) do vírus causador da COVID-19, o SARS-CoV-2, assim criando um programa que converte dados em código alfanumérico e em sequências de DNA, sendo capaz de armazenar os formatos de arquivos txt, jpg, mp4, pdf e csv. Em laboratórios especializados, estas sequências podem ser sintetizadas em moléculas de DNA, que serão conservadas e guardarão a informação codificada por um longo período de tempo. Para demonstrar a funcionalidade do programa, foi gravado um vídeo registrando o nosso estilo de vida atual e com perguntas para as civilizações futuras.

DESENVOLVIMENTO

O desenvolvimento do projeto passou por duas áreas: a biológica e a de programação em Python; tendo as seguintes etapas:

1º: Alinhamento múltiplo das sequências genéticas do gene S das VOC's do SARS-CoV-2, com base em bancos de dados públicos;

2º: Análise de todas as sequências de DNA possíveis para cada variante de acordo com suas mutações características;

3º: Manipulação dos códons para criar 5 sequências hipotéticas com os mesmos aminoácidos correspondentes a cada variante, a partir de uma sequência de referência;

4º: Construção de um dicionário que será utilizado pelo programa contendo as sequências analisadas, para gravar as informações nas regiões de mutações silenciosas;

5º: Desenvolvimento do algoritmo com bibliotecas implementadas na linguagem de programação Python, para permitir a manipulação de arquivos;

6º: Criação de um dataframe interno que armazenará os dados dos arquivos codificados.

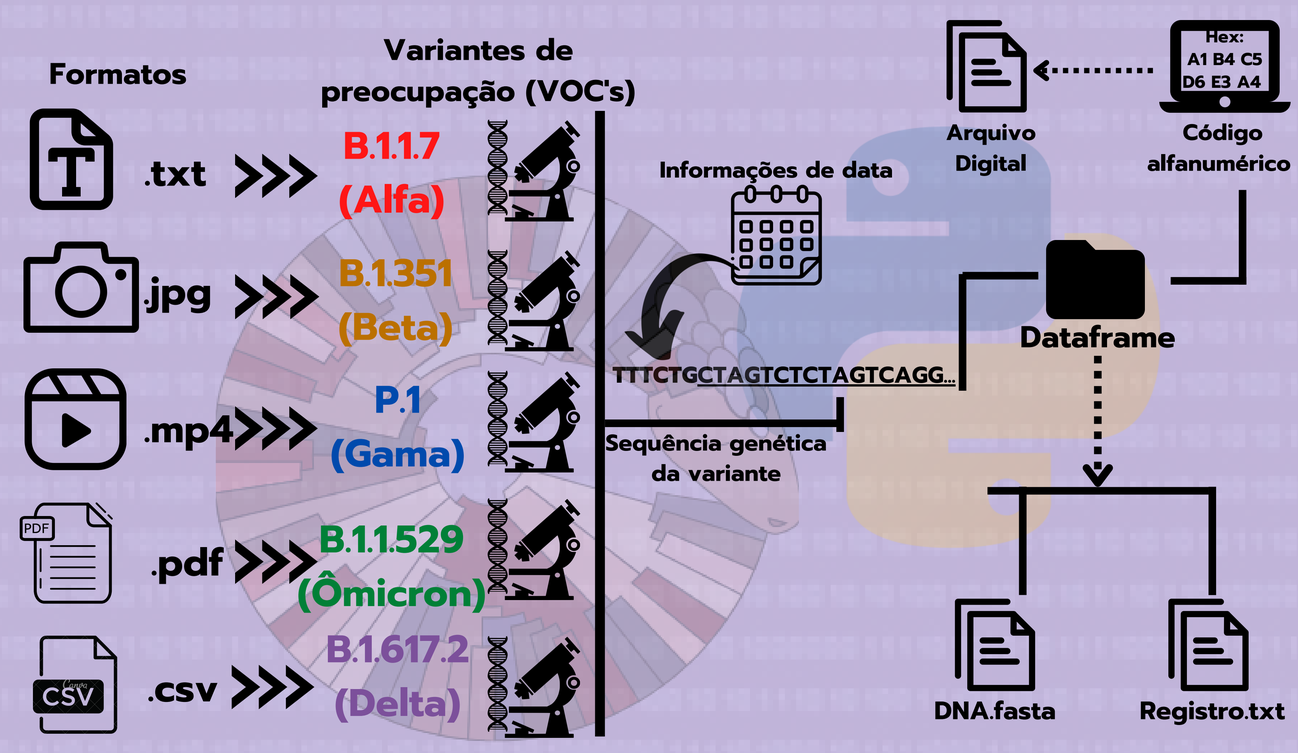

Imagem 1 - Fluxograma de funcionamento do programa

RESULTADOS

- O programa funciona em sua integridade, gerando um arquivo de registro, que exibe o dataframe, e um arquivo no formato fasta, que contém a sequência de DNA.
- O vídeo foi gravado e codificado usando o programa, para sua sequência ser futuramente sintetizada em uma molécula de DNA.
- Os testes estatísticos confirmaram que o tempo de conversão e recuperação de um arquivo é diretamente proporcional ao seu tamanho.
- Os documentos fas gerados sempre possuem 4kb, independente do tamanho do arquivo original, já que o gene S é sempre utilizado como modelo.

Arquivo	Formato	Tamanho	Tempo para conversão (s)	Tempo de recuperação (s)
1	txt	1 kb	0,02	0,003
2	txt	23 mb	24,9	9,39
3	jpg	52 kb	0,08	0,01
4	jpg	143 kb	3,13	0,06
5	mp4	7,5 mb	12,1	3,7
6	mp4	23.2 mb	39,3	25
7	pdf	7 mb	9,9	4,0
8	pdf	91 kb	6,57	0.07

Imagem 2 - Relação do tamanho dos arquivos com o tempo de conversão e recuperação.

CONCLUSÕES E PERSPECTIVAS

- A implementação de bibliotecas que possibilitem o registro de coordenadas geográficas e sua posterior inserção na sequência de DNA permitirá identificar o local do armazenamento, melhorando a precisão.
- A implementação de uma biblioteca de aprendizado de máquina do Python, a scikit-learn, poderá otimizar a geração de sequências de DNA a partir dos dados.
- Mesmo utilizando o gene S do SARS-CoV-2 como modelo, o algoritmo tem potencial de utilizar qualquer outro gene, possibilitando a síntese de uma molécula de DNA.
- O programa é uma alternativa tanto para a segurança de informações, utilizando a sequência de DNA como criptografia, quanto para o armazenamento de dados.
- O armazenamento em DNA já é uma realidade[4], e um potencial meio mais duradouro e capaz de guardar dados.
- O aperfeiçoamento desta técnica é uma possível revolução tecnológica e social.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] DONG, Yiming et al. DNA storage: research landscape and future prospects. National Science Review, v. 7, n. 6, p. 1092-1107, 2020

[2] CEZE, Luis; NIVALA, Jeff; STRAUSS, Karin. Molecular data storage using DNA. Nature Reviews Genetics, v. 20, n. 8, p. 456-466, 2019

[3] BRASIL. Ministério da Saúde - Secretaria de Vigilância em Saúde. Boletim Epidemiológico Especial – COVID-19 Nº 131. Brasil, 23 de setembro de 2022. 158 p.

[4] Extance, A. How DNA could store all the world´s data. Nature 537, 22-24 (2016). https://doi.org/10.1038/537022a Rincon, Paul. BBC News, Como cientistas querem usar DNA para armazenar dados. https://www.bbc.com>internacional-59572594