Instituto Federal Catarinense - Campus Videira
Aluno: Dyeizon Procopiuk da Silva
Turma: Ciência da Computação 2022/1
Disciplina: Algoritmos
Data: 02/07/2022

Metodologia utilizada para a solução do problema

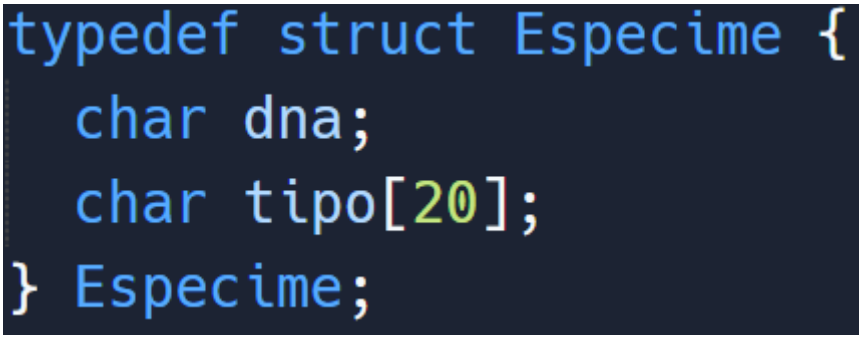
Para a entrada da sequência de DNA, foi criado um arquivo *.txt* que deve conter na primeira linha a dimensão da matriz quadrada, e nas linhas seguintes, a matriz quadrada a ser lida. Cada elemento da matriz deve estar sendo separado por um caractere de espaço. Então, a função **getDNAFromFile** interpreta o arquivo passado como parâmetro, e aplica as alterações necessárias para que a matriz seja utilizada pelo algoritmo, retornando um vetor com os elementos concatenados, resultando em uma sequência de DNA crua.

Figura 1 - Modelo de entrada

O próximo passo é a validação da sequência de DNA. A função **checkDNA** verifica se há alguma base nitrogenada inválida dentro da sequência de DNA, considerando que uma base nitrogenada inválida é qualquer caractere que não seja “A”, “G”, “C” ou “T”.

Após a validação da sequência de DNA, é criada uma instância do registro **Especime** (Figura 2), utilizando alocação dinâmica de memória. Este registro armazena a sequência de DNA, e a espécie que esta sequência corresponde.

Figura 2 - Registro Especime



A função **isSimian** submete a sequência de DNA a uma série de testes nas linhas, colunas, diagonais principais e diagonais secundárias em busca de repetições de 4 bases nitrogenadas coincidentes. Por motivos de performance, é verificado primeiramente as linhas e colunas, tendo a matriz inteira como referência, e caso seja encontrado repetições, a função responsável por realizar os testes das diagonais não é chamada.

Caso não sejam encontradas repetições nas linhas e colunas, a função recursiva **subMatrix** é chamada. Esta função é responsável por procurar repetições na diagonal principal e secundária de uma submatriz 4x4 que percorrerá toda a matriz (ambas as verificações são feitas ao mesmo tempo na submatriz), e chamando a si própria com valores diferentes até encontrar repetições, ou até chegar ao final da matriz.

O retorno da função **isSimian** determinará o tipo da espécie analisada pela sequência de DNA, sendo “humano” caso retorne 0, e “símio” caso retorne 1. Caso seja um símio, o algoritmo informa ao usuário onde foi encontrada repetição e qual caractere foi encontrado repetidamente.

O algoritmo pode ser facilmente modificado para considerar repetições mais extensas de bases nitrogenadas, assim como maiores submatrizes para as verificações das diagonais, caso necessário.

Resultados obtidos

Realizei uma sequência de testes com uma matriz quadrada de dimensões 20x20, em que inseri repetições de bases nitrogenadas em posições aleatórias da matriz, para verificar a precisão do algoritmo.

A primeira verificação foi nas linhas (Figura 3), o algoritmo encontrou uma repetição de 4 bases nitrogenadas “A” na linha 17 em 0.000238 segundos. O algoritmo finalizou a execução no momento em que a repetição foi encontrada, e **não** verificou as colunas e nem as diagonais, pois o fato de haver ou não repetições de bases nitrogenadas em outros sentidos não alteraria o resultado.

A próxima verificação é a das colunas (Figura 4), o algoritmo encontrou a repetição da base nitrogenada “G” em 0.000267 segundos.

Ao não encontrar repetições de bases nitrogenadas nas linhas e colunas, a função **subMatrix** é chamada para tentar localizar repetições nas diagonais. No próximo exemplo, foi encontrada uma repetição da base nitrogenada “C” na diagonal principal em 0.000330 segundos (Figura 5).

A última verificação da **subMatrix** é na diagonal secundária, para este teste em questão, coloquei a repetição na última posição possível da submatriz. A repetição da base nitrogenada “T” foi encontrada na diagonal secundária 0.000356 segundos (Figura 6).

Figura 3 - Linhas

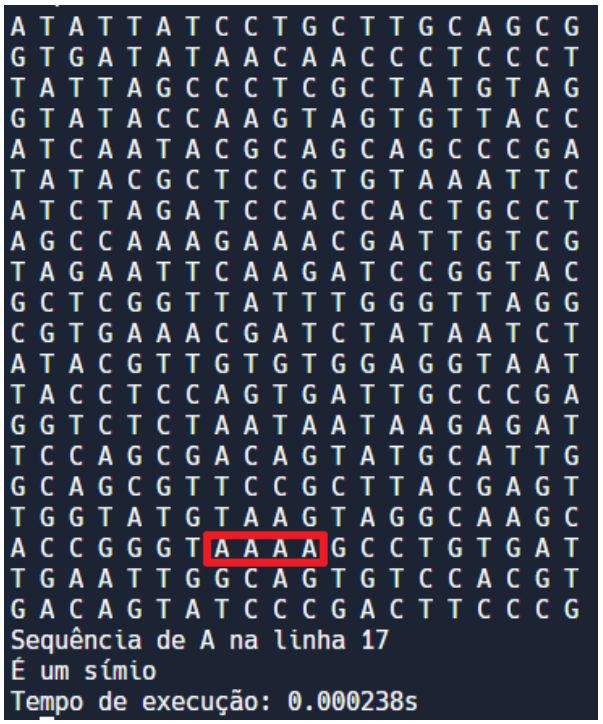


Figura 4 - Colunas

Figura 5 - Diagonais principais

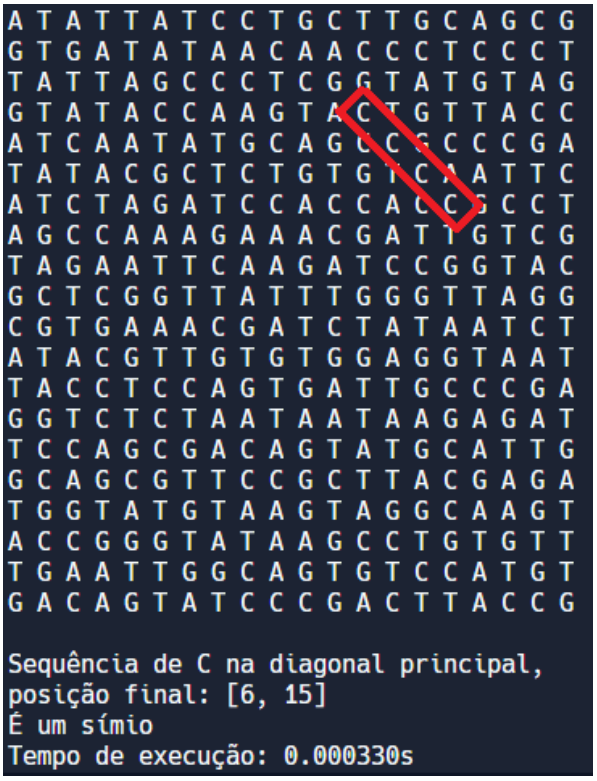
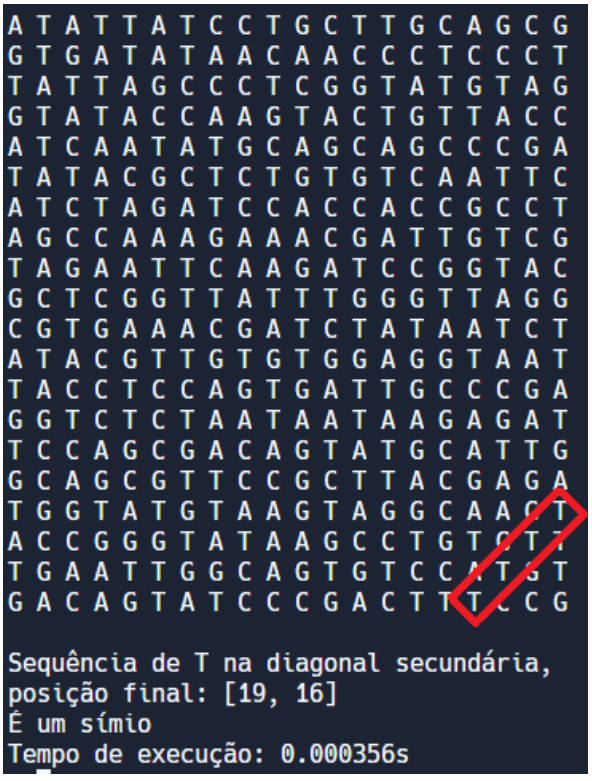


Figura 6 - Diagonais secundárias



Caso sejam feitas todas as verificações e não tenham sido encontradas repetições, o espécime é considerado “humano”, como verificado na Figura 7, em que a verificação levou 0.000390 segundos para passar por todas as etapas de verificação, e chegar à conclusão de que a sequência de DNA corresponde a de um ser humano.

Figura 7 - Sequência de DNA humano