

Estrutura de Dados II

Trabalho Prático

Data de entrega: 03/12/2023

1 Descrição do Problema

Durante sua excursão pela Floresta Amazônica, Dr. Manhattan percebeu que um grupo de mamíferos de diferentes espécies estavam doentes. Ele resolveu recolher alguns exemplares para estudo e montou um laboratório improvisado no meio da floresta para tal objetivo. Ele identificou que o causador da doença é um vírus que se transmite por fluidos do corpo de seu hospedeiro, se adapta facilmente a diferentes espécies de mamíferos e se propaga com mais rapidez em espécies que vivem em bandos. Ao analisar o comportamento do vírus nos corpos dos animais infectados, o pesquisador descobriu que ele se utiliza do ciclo lisogênico para sua reprodução, ou seja, ele agrega seu material genético ao genoma da célula hospedeira.

Após algumas semanas, o pesquisador observou que o vírus sofreu várias mutações e que algumas delas prevaleciam sobre outras. No estudo genético sobre as variações do vírus, ele observou que há mais de uma ocorrência do padrão do vírus em uma sequência de DNA de uma célula hospedeira. Além disso, foi detectado em um mesmo animal mais de uma variante do vírus.

Para identificar as variantes do vírus nos animais e estudar o comportamento do vírus nos genes dos hospedeiros, ele solicitou a você que implementasse um programa que recebe dois arquivos: (1) um arquivo com várias sequências de DNA (obtidos dos animais analisados); e (2) um arquivo com vários padrões de DNA obtidos das variações do vírus. A saída do programa deve ser a listagem de todas as ocorrências de cada padrão de variações do vírus.

2 Objetivo do trabalho

Escreva um programa que recebe dois arquivos:

- (1) BaseDadosDNA.txt: contém as sequências de DNA coletadas dos animais,
- (2) PadroesVirus.txt: contém as sequências de DNA dos variantes do vírus.

Para cada padrão de DNA do vírus, o programa deve imprimir o número de ocorrências do mesmo na sequência do DNA do animal seguido de uma lista com as posições onde o primeiro caractere do padrão (do vírus) ocorre na sequência de DNA do animal.

2.1 Padrão do arquivo BaseDadosDNA.txt

Cada sequência de DNA será precedido por uma linha começando com o caractere “>” (maior que) seguida por um código e por uma breve descrição do animal do qual o DNA foi obtido.

A sequência de DNA começará então na próxima linha por um certo número de linhas. Cada linha conterá exatamente 70 caracteres do conjunto A, C, G e T, exceto a última linha, que poderá conter menos de 70 caracteres. Após esta última linha estará a linha descritora da próxima sequência de DNA. O final do arquivo será indicado pelo descritor (“>EOF”).

Exemplo:

```
>0989487.1 onca pintada
GGTACCTCCTACGGGAGGCAGCAGTGAGGAATTTTCCGCAATGGGCGAAAGCCTGACGGAGCAATACCGC
GTGGAGGATGACGGCCTGTGGGTCTTAAACTCCTTTTCTTAGAAAAGAAATTTTGACGGTATCTAAGGAA
TAAGCATCGGCTAACTCCGTGCCAGCAGCCGCGTAATACGGGGGATGCAAGCGTTATCCGGAATGATTG
GGCGTAAAGCGTCTGTAGGTGGCTTTCCAAGT
>L034233.4 lobo guara
TGTACAATTGCTAGCGTACGGCCGTCGACGTCCGGATCCTCGAGAAAGGAGAAGAAGCTTTTCACTGGAGT
TGTCCCAATTCTTGTGAATTAGATGGTGATGTTAATGGGCACAAATTTTCTGTCAGTGGAGAGGGTGAA
GGTGATGCAACATACGGAAAACTTACCCTTAAATTTATTTGCACTACTGGAAAACTACCTGTTCCATGGC
CAACACTTGTCACTACTTTCTTATGGTGTTCAATGCTTTTCAAG
>EOF
```

2.2 Padrão do arquivo PadroesVirus.txt

Cada sequência de DNA será precedido por uma linha começando com o caractere “>” (maior que) seguida por uma breve descrição do padrão. A sequência de DNA começará então na próxima linha e conterá até 70 caracteres do conjunto A, C, G e T. Após esta última linha estará a linha descritora da próxima sequência de DNA. O final do arquivo será indicado pelo descritor (“>EOF”).

Exemplo:

```
>variante B12.1
GGAGA
>variante C4.3
AAAG
>EOF
```

2.3 Padrão de saída

Para cada padrão contido no arquivo PadroesVirus.txt, imprima (no terminal) uma linha com a descrição do padrão e, nas linhas seguintes, imprima a descrição da sequência de DNA (do arquivo BaseDadosDNA.txt) entre colchetes seguido do número de ocorrências e das posições onde o padrão ocorre na sequência de DNA (uma linha para cada sequência de DNA). Veja o exemplo a seguir.

```
variante B12.1
[L034233.4 lobo guara] no. de ocorrencias: 2 posicoes: 47 128

variante C4.3
[0989487.1 onca pintada] no. de ocorrencias: 3 posicoes: 47 113 215
[L034233.4 lobo guara] no. de ocorrencias: 1 posicoes: 44
```

3 Entrega do trabalho

A avaliação do trabalho será realizada no sistema Linux. Logo, caso o programa implementado não compile no Linux a nota será **zero** (compilação e execução em linha de comando utilizando o programa **make**).

Todos os arquivos do trabalho deverão ser compactados em um arquivo, cujo nome deve obedecer o formato:

ED2_Grupo<número do grupo>_[Aluno1][Aluno2][Aluno3][Aluno4].zip,

onde <número do grupo> deve ser substituído pelo número do grupo definido na lista de grupos. A primeira página do relatório, assim como **cada arquivo de código do programa**, devem conter o número do grupo e os nomes dos componentes do grupo.

Além dos arquivos de código fonte, os grupos deverão implementar o arquivo **makefile** para compilar o arquivo no linux e deverão escrever um **relatório** com a documentação do trabalho. O trabalho deverá ser desenvolvido em grupo de quatro alunos e deverá ser entregue até o dia **03/12/2023 às 23:55 hs no sistema ava.ufes.br**.

4 Apresentação:

As apresentações dos trabalhos ocorrerão nos dias 04/12/23 e 06/12/23 em sala de aula. Os grupos devem preparar *slides* e realizar a apresentação de forma mais didática possível. A qualidade da apresentação será considerada na avaliação do trabalho.