Universidade Federal de Viçosa – Campus UFV-Florestal
Ciência da Computação – Projeto e Análise de Algoritmos
Professor: Daniel Mendes Barbosa

Trabalho Prático 3

Este trabalho é **obrigatoriamente em grupo**. Os grupos já foram definidos nesta planilha e este trabalho deverá ser entregue no PVANet Moodle de acordo com as instruções presentes no final da especificação.

Atualmente a computação é ubíqua, ou seja, é onipresente e pode ser utilizada em diversos contextos para resolver problemas de forma simples e robusta. Durante a pandemia da COVID-19, a bioinformática foi de grande importância para reconhecer genomas diferentes e mapear as novas variantes. Um laboratório localizado em Wuhan tem pesquisado sobre os antepassados do ser humano. Os pesquisadores coletaram DNAs de seres humanos, cachorros e de diferentes espécies de chimpanzés. No entanto, o laboratório deseja saber qual é o DNA mais similar ao do ser humano, o do chimpanzé ou o do cachorro e para isso precisa do poder da computação para resolver esse problema. Os pesquisadores de Wuhan tem uma parceria com a UFV-CAF e encomendou esse estudo de similaridade de DNA. Os alunos de PAA foram escolhidos para terminar essa pesquisa e ao final enviarão os relatórios com os resultados obtidos.

Portanto, vocês deverão a partir dos DNAs dos animais **determinar a similaridade entre os pares**, ou seja, *similaridade(humano, chimpanzé)*, *similaridade(humano, cachorros)*, *similaridade(chimpanzé, cachorros)*.

Contextualização

O DNA é uma molécula presente no núcleo das células que contém as informações genéticas de um organismo. Ela é composta por uma fita dupla com nucleotídeos que possuem 4 bases nitrogenadas: A-Adenina, C-Citosina, T-Timina, G-Guanina.

Vocês irão receber 3 arquivos nomeados com o sequenciamento dos 3 animais (humano.txt, chimp.txt, cachorro.txt). Cada arquivo possui 4320, 1682 e 820 registros, respectivamente. Os DNAs também possuem tamanho dinâmico (**não é necessário utilizar alocação dinâmica**). Portanto, a entrada é composta de 3 arquivos e a saída é composta por 3 valores de similaridade que será apresentado a seguir.

Como calcular a similaridade

A similaridade deve ser calculada a partir da fórmula da similaridade por cossenos conforme abaixo:

$$\text{cosine similarity} = S_C(A,B) := \cos(\theta) = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$

Lembrando que **A** e **B** são vetores. Para definir esses vetores, é necessário combinar sequências das bases nitrogenadas e contar a frequência em cada sequência. Segue um exemplo abaixo.

Considere que exista uma sequência de DNA do humano *AACCCTG* e uma sequência do chimpanzé *AACCTG* e temos os produtos cartesianos das bases {*AA, CC, CT, TG}*. Logo, teremos dois vetores **A** = {1, 2, 1, 1} (1 *AA*, 2 *CC*, 1 *CT* e 1 *TG)* e **B** = {1, 1, 1, 1}. **Se existisse mais de uma sequência de DNA humano, o vetor A teria seus valores acrescidos.**

A partir disso, é possível calcular a similaridade que será igual a *0.944911182523068,* quanto mais próximo de 1, maior a similaridade. Note que para realizar a contagem, é necessário fazer a busca por padrões no texto, portanto, **vocês devem utilizar algum dos três algoritmos**: Boyer-Moore, Shift-And e Knuth-Morris-Pratt. Não faz sentido utilizar o shift-and aproximado.

Como obter os produtos cartesianos das bases

Cada produto cartesiano pode ser de tamanho variável, ou seja, produto com tamanho 1: {A, C, T, G}; com tamanho 2: {AA, AT, AC, AG, TA, TT, TC, TG, CA, CT, CC, CG, GA, GT, GC, GG}. Portanto, é necessário calcular o produto cartesiano das bases para depois, para cada elemento, buscar os padrões em cada sequência.

Não está definido nesse trabalho qual é o tamanho que deve ser utilizado para cada elemento do conjunto, logo vocês podem experimentar qual é o melhor tamanho de acordo com a similaridade obtida ao final.

Veja que como a cardinalidade do conjunto resultante da operação de produto cartesiano é muito grande, vocês devem escolher randomicamente uma quantidade de elementos que serão utilizados para a busca de padrões. No exemplo acima, foi calculado o produto com o tamanho dos elementos igual a 2 e escolhido randomicamente 4 elementos do conjunto.

Simulações

Como estamos lidando com um problema de combinação e aleatoriedade para determinar a similaridade entre sequências de DNA, utilizaremos uma técnica estatística para aproximar o valor o mais próximo possível da realidade.

Uma simulação é usada para calcular, por exemplo, a probabilidade de jogar 2 dados comuns. Existem 36 possibilidades, mas sabemos que é improvável que saia 2 dados com o número 6 na maioria das vezes. Podemos comprovar isso simulando essa jogada, com 100 tentativas. Podemos tentar ser mais fiel e fazer uma simulação com 1000 tentativas. Quanto maior o número de tentativas, mais fiel seremos ao ***resultado esperado***, isso é chamado de *simulações de Monte Carlo*. Uma técnica que nos permite simular experimentos aleatórios que são impossíveis de determinar o seu resultado analiticamente.

Então, vocês deverão utilizar dessa ideia para simular o comportamento da escolha das bases nitrogenadas aleatórias. Por exemplo, suponha que fizeram o produto cartesiano com elementos de tamanho 2, será gerado 16 elementos no conjunto. Para realizar o cálculo da similaridade dos vetores, escolheram 4 elementos do conjunto aleatoriamente. Para a simulação, então, essa escolha randômica e o cálculo da similaridade deve ser feito **X** vezes. Após a simulação feita, terá um vetor de similaridade entre **A** e **B** com **X** valores, e o resultado da similaridade será a média desses **X** valores. **A quantidade de vezes que o processo será repetido, será definido por cada grupo.**

Lembrem-se também que podem simular quantos elementos serão escolhidos aleatoriamente. Essa simulação é livre e será considerada no valor final do trabalho prático. Cada processo feito deve ser descrito no relatório final, vocês estão sendo avaliados pelo laboratório de Wuhan, um renomado centro de biologia molecular.

Resumo

- Ler os três arquivos de texto
- Calcular o produto cartesiano dos elementos
- Escolher aleatoriamente o número de elementos do conjunto
- Contar os padrões nas sequências genéticas utilizando um dos algoritmos de casamento de padrões (BM, Shift-And e KMP)
- Calcular a similaridade entre pares
- Simulação do processo
- Apresentar os resultados no relatório final

Tarefas Extras

1. Apresentar os resultados de forma gráfica com simulações variando a quantidade de vezes que foi repetido o processo. Houve variação do valor da similaridade?
2. O Laboratório de Wuhan pode ter mais pesquisas nesse sentido e precisa saber qual é o algoritmo mais eficiente para esse intuito. Faça um comparativo sobre a

eficiência desses algoritmos. Varie a quantidade de elementos do conjunto e escolha mais elementos para fazer o casamento de padrões. Qual algoritmo se saiu melhor?

Formato e data de entrega

Os arquivos com o código-fonte (projeto inteiro do Codeblocks ou arquivos .c, .h e makefile), juntamente com um arquivo PDF (**testado, para ver se não está corrompido**) contendo a **documentação**. A documentação deverá conter:

- explicação dos algoritmos projetados;
- implementação do algoritmo projetado (estruturas de dados criadas, etc);
- explicação de como compilar o programa.

Mais direcionamentos sobre o formato da documentação podem ser vistos no documento "Diretrizes para relatórios de documentação".

Importante: Entregar no formato **ZIP**. As datas de entrega estarão configuradas no PVANet Moodle. É necessário que apenas um aluno do grupo faça a entrega, mas o PDF da **documentação deve conter os nomes e números de matrícula de todos os alunos do grupo que efetivamente colaboraram com o trabalho em sua capa ou cabeçalho**.

Bom trabalho!