ORIENTAÇÕES PARA O PROJETO DE CURSO DE BIOINFORMÁTICA

CURSO DE SISTEMAS DE INFORMAÇÃO

MATÉRIA: BIOINFORMÁTICA

Professor: Diego Frias

Semestre: 2021.2

1. Introdução:

O projeto de curso tem como finalidades:

a) Consolidar os conhecimentos sobre o dogma central da Biologia Molecular, estrutura de DNA, genes e codons.
b) Treinar a mineração de sequências no Genbank
c) Treinar o processamento digital de sequências nucleotídicas em formato FASTA
d) Treinar a localização de motifs (sinais) em sequências genômicas
e) Desenvolver um anotador de genomas virais de fita simples aplicando os conceitos de Open Reading Frame (ORF) e Potencial Codificante
f) Opcional: Assimilar uso de bibliotecas para Bioinformática disponíveis nas liguagens modernas de scripting (Java, Perl, Python ou PHP).

2. Metodologia:

O projeto consta de 4 fases:

a) **Mineração de Dados:** Consiste em:
    1. Fazer download de 20 genomas completos de 2 espécies de vírus de fita simples (indicadas para cada equipe) em arquivo FASTA
    2. Fazer download da anotação de cada genoma e construir estrutura de dados (tabela) com: denominação, posição de início, posição terminal de cada ORF em cada genoma. Esta tabela será usada para estimar a "qualidade" da anotação realizada com o algortimo desenvolvido pela equipe.
b) **Leitura de arquivo FASTA e localização das ORFs putativas**: Após leitura do arquivo FASTA com os genomas de uma espécie viral, realizar a busca nos 3 frames dos inícios e fins das ORFs, localizando as posições dos Stop-Codons (TAA,TAG ou TGA) e do Start-Codon (ATG) em cada ORF.

    Levar em conta que:

1. Entre 2 Stops-Codons existe apenas 1 ORF, e
2. Que se não houver Start-Codon entre 2 Stop-codons, a ORF que finaliza no Stop-codon da direita deve ser desconsiderada.

c) **Identificação das ORFs mais prováveis segundo o potencial codificante**: Consiste em:
   1. Implementar algoritmo para o cálculo das "features" utilizadas no modelo de potencial em cada ORF
   2. Implementar o modelo de potencial (método que recebe as features e retorna Sim ou Não)
   3. Calcular o potencial de cada ORF candidata e selecionar aquelas com potencial codificante.
   4. Imprimir lista das ORFs identificadas (número consecutivo, incio, final) em cada genoma

d) Comparação de resultados com a anotação de referência:

Consiste em contar:

1. **TP - Verdadeiros Positivos:** O número de ORFs identificadas que possuem igual início e final que a ORFs de referência.
2. **pTP - Parcialmente Verdadeiros Positivos:** O número de ORFs identificadas que possuem apenas o início ou o final igual à ORF de referência.
3. **FP - Falsos Positivos:** Contar as ORFs identificadas que não se encontram na anotação de referência (nem início nem final iguais a nenhuma ORF de referência)
4. **FN - Falsos Negativos:** Contar as ORFs na anotação de referência que não foram identificadas pelo algoritmo.
5. **Cálculo da Acurácia, Precisão e Revocação** em cada espécie viral. Discussão dos resultados.

3. Resultados Esperados:

a) 2 arquivos FASTA com 20 genomas completos das 2 espécies virais usadas para teste

b) Algoritmo implementado pela equipe que processa os arquivos e lista as ORFs (inicio-fim) de cada genoma

c) Documento descrevendo a solução e os resultados. Os resultados devem ser tabelados para cada espécie viral e cada genoma, da forma:

Espécie A:

Genoma 1:

	Anotação		Algoritmo	
	Início	Fim	Início	Fim
ORF1	-	-	-	-
ORF2	-	-	-	-

…

ORFn ← n = n[Umero de ORFs anotadas na espécie A

… incluir linhas adicionais para ORFs identificadas mas não anotadas

---

Genoma 2:

…

Genoma 20:

…

Espécie B:

Genoma 1:

….

Os resultados devem conter também as métricas de desempenho do algoritmo tabeladas na forma:

Espécie	Acurácia	Precisão	Revocação
A	-	-	-
B	-	-	-

e finalizar com a discussão dos resultados respondendo, entre outras, às

questões:

1. Quão “eficiente' é o método utilizado (a partir das métricas médias das 2 espécies)?
2. Qual das 3 métricas é a maior? O que isso implica desde o ponto de vista prático para o usuário que usar esse método?
3. Em qual das espécies o método funcionou melhor?

4. Avaliação:

Cada fase do projeto descrita na seção de Metodologia terá o seguinte peso na nota:

a) Mineração de Dados: 0.15
b) Leitura de arquivo FASTA e localização das ORFs putativas: 0.35
c) Identificação das ORFs mais prováveis segundo o potencial codificante: 0.3
d) Comparação de resultados com a anotação de referência e discussão: 0.2

BOM TRABALHO!