Trabalho 1 de Algoritmos de Bioinformática
Definição de uma classe de manipulação de sequências
27 de Fevereiro de 2019

Entrega: 06/03/2019 até 24:00 (1 semana)

Este trabalho é para ser submetido via Moodle. Excepcionalmente para aqueles que não tenham acesso podem submeter via email para pgferreira@fc.up.pt com o seguinte text no assunto [AlgoritmosBioinformatica1819-Trabalho1] [Nome e número do Aluno].

O trabalho será desenvolvido como trabalho extra-classe. Os testes e exames terão perguntas relacionadas com este trabalho. Para saber o método e critério de avaliação, por favor consulte a ficha da unidade curricular na página do sigarra.

1 Descrição do Problema

O objectivo deste trabalho é o da implementação de uma classe com funcionalidades para representar e manipular sequências biológicas. A classe deverá chamar-se *BioSeq* e deverá incluir todas as funcionalidades apresentadas durante as aulas de *"Basic Processing of Biological Sequences".* A classe deverá lidar com três tipos de sequências: DNA, RNA e Protein e os métodos deverão ser adaptados a cada um destes tipos.

2 Implementação

A classe deverá conter pelo menos um atributo para representar a sequência, outro para o tipo de sequência e outros atributos que considere necessários.

Funcionalidades mínimas a implementar:

- Construtor de classe com biotipo
- Métodos de visualização da informação da classe
- Validação do tipo de sequências
- Frequência de símbolos
- GC content
- Reverse complement
- Transcription
- Translation
- Codon usage
- Find all ORFs (by minimum size)
- Pretty-printing da informação da classe
- Leitura a partir de um ficheiro Fasta (das sequências)
- Função Load e Save que guarda e lê em ficheiro o estado do objeto.

Notas:

1. Nem todas as funcionalidades fazem sentido aplicar nos três tipos pelo que terá que ter o cuidado de gerir a chamada de métodos que não façam sentido para um dado tipo de sequência.
2. Encontre exemplos de sequências para testar a sua classe (ver nota final).
3. Serão valorizados aplicações originais da classe.

3 Relatório para entrega

O trabalho deverá ser acompanhado de um pequeno relatório (em Português ou Inglês) com o máximo de duas páginas (tamanho de letra 11) e em formato pdf. Neste deve discutir os seguintes pontos:

- Introdução - Contextualizar e descrever brevemente o problema.
- Descrição e estratégias de implementação - Discutir abordagens relevantes ao problema.
- Resultados - Deve indicar que funcionalidades foram implementadas, se conseguiu implementar todas as funcionalidades pedidas e se implementou outras funcionalidades além das especificadas.
- Comentários e Conclusões.
- Referências Bibliográficas (precisam ser explicitamente citadas no texto para saberem de onde o texto foi retirado/adaptado! Copiar é crime e poderá transformar-se em processo disciplinar, portanto evitem copiar textos e códigos. Se utilizarem figuras retiradas da web ou de livros ou de artigos etc, é necessário colocar uma referência explícita e clara. Por favor tenham atenção aos erros ortográficos.

4 Entrega

Submeter através do Moodle um arquivo zip contendo todo o código fonte dos programas, e instruções de como executar('readme'). Todos os ficheiros devem ser colocados na mesma pasta incluindo os ficheiros com sequências de teste.

Importante: Deverão implementar um ficheiro run_me.py em que fazem a importação da classe implementada e através de várias exemplos demonstram a chamada dos vários métodos implementados. Para tal o programa deve imprimir mensagens a indicar a funcionalidade implementada. **O programa deve correr na linha de comando (python run_me.py)**.

O trabalho pode ser feito em grupo de no máximo duas pessoas. Trabalhos com cópia de código de outros grupos serão desclassificados!

Como obter sequências de genes

No site https://www.ncbi.nlm.nih.gov/genbank/ pode fazer a pesquisa por um gene de interesse. Por exemplo, CDH1, TP53, MDM2, KRAS,....
Pode por exemplo pesquisar por: CDH1 AND "Homo sapiens"
Na caixa de resultados que aparece no topo da página encontrará um botão de Download (Download FASTA sequences). Aqui encontrará sequências de DNA (RefSeqGene), de diferentes transcritos de mRNA (RefSeq Transcripts) ou de proteínas.