Bioinformática 2021/2022

Exercise – Machine Learning

Discuta os seguintes cenários com os elementos do seu grupo e proponha soluções para a sua análise.

Cenário 1

De forma a poderem ser usadas em analise com métodos de Deep Learning as sequências biológicas precisam de ser convertidas num formato adequado. A codificação **One-Hot-Encoding (OHE)** permite que uma sequência de valores categóricos (e.g nucleotídeos) sejam convertidos num formato binário/numérico. Escreva uma função que dada uma sequência de DNA faça a conversão OHE da mesma, *dna_ohe(seq).*
A —> [1, 0, 0, 0]
C *—> [0, 0, 1, 0]*
G —> [0, 0, 1, 0]
T —> [0, 0, 0, 1]

Cenário 2
Considere que têm um conjunto de sequências de proteínas (identificadas pelo seu nome) e pretende saber como essas se distribuem e que tipo de relações e estrutura apresentam entre elas. Admita que além da sequência **não** possui mais informação sobre natureza das sequências (e.g. família de proteínas, presença de motifs funcionais, espécie, etc.).

- Escreva uma função que dada uma sequência retorna uma estrutura de dados que contém a frequência de cada k-mer na sequencia. Function: *word_to_kmer(word, k).*
- Escreva uma função que dado um ficheiro de sequencias em formato fasta, retorna uma tabela com as frequências de todos os k-mers em cada uma das sequências: *file_to_kmer_table(file_name).* Poderá retornar uma tabela com um dos objectos apresentados nas aulas ou numa tabela do tipo *pandas.*

Cenário 3
Considere um segundo cenário em que possui à partida um conjunto de proteínas previamente classificadas em 10 tipos de famílias. A ideia seria conseguir atribuir a novas sequências que vão sendo determinadas a um dos 10 tipos de famílias já conhecidas. Admita também que neste caso, não temos as sequências mas antes um conjunto de indicadores sobre a composição das sequências, como por exemplo: tamanho, GC%, presença/

Frequência de mais de 100 tipos de motifs da base de dados PFAM, número de hélices alpha, folhas beta, etc...

Descreva que tipo de abordagem baseada em Machine Learning usaria para cada análise em cima. Considere os seguintes pontos:
- natureza do problema de aprendizagem/análise;
- objectivo;
- dados de entrada e dados de saída?
- metodologia para definir a relação entre sequências (medidas, distâncias, score functions, etc....);
- considere como a sua proposta iria escalar para n =100, 1000, 10000, ...
- outros pontos que considere relevante.