Algoritmos para Análise de Sequências Biológicas

Ficha 6

Objetivo

- Implementar uma versão simplificada do Blast

Versão simplificada

- Considera apenas matches perfeitos entre a query e as sequências da BD
- Critérios simples usados para a extensão dos hits
- Score será a contagem do nº de matches

Exercício

1. Crie uma função chamada `query_map` que recebe a sequência e o `w` e que devolve um dicionário em que as chaves são as sequências e os valores são uma lista dos índices
2. Crie uma função chamada `hits` que recebe o dicionário da função anterior e uma sequência da BD e devolve uma lista de *hits* em que cada elemento é um tuplo com os índices
3. Crie uma função chamada `extend_hit` que recebe a query, a sequência da BD, o hit e o valor de w e o estende um hit em cada direção se o nº de matches for de pelo menos metade do tamanho da extensão; a função devolve um tuplo com o índice do início do hit estendido na query, na sequência, o tamanho e o nº de matches
4. Crie uma função chamada `best_hit` que recebe uma query, uma sequência da BD e o `w` e que devolve a extensão de maior score (no caso de empate, deverá devolver a de menor tamanho que aparece primeiro)

Exemplo

```
>>> query = "AATATAT"
>>> seq = "AATATGTTATATAATAATATTT"
>>> w = 3
>>> qm = query_map(query, w)
>>> qm
{'AAT': [0], 'ATA': [1, 3], 'TAT': [2, 4]}
>>> hits(qm, seq)
[(0, 0), (0, 12), (0, 15), (1, 1), (1, 8), (1, 10), (1, 13), (1, 16),
 (3, 1), (3, 8), (3, 10), (3, 13), (3, 16), (2, 2), (2, 7), (2, 9),
 (2, 17), (4, 2), (4, 7), (4, 9), (4, 17)]
>>> extend_hit(query, seq, (1, 16), 3)
(0, 15, 7, 6)
>>> best_hit(query, seq, 3)
(0, 0, 7, 6)
```

Exercício

5. Crie uma classe chamada `SimpleBlast`
6. Construtor recebe uma lista de sequências ou um ficheiro com sequências e o `w`
7. Crie uma classe chamada "SimpleBlastHit"' para encapsular os hits do Blast
8. Crie uma classe chamada "SimpleBlastMatch"' para encapsular os matches do Blast
9. Crie um método chamado `best_alignment` que recebe a sequência de query e devolve a sequência da BD que corresponde ao maior match

Sugestões de melhoria

- Implemente uma versão que considere uma matriz de substituição. Nesse caso, deverá definir um parâmetro T e considerar os hits de tamanho W e score $>=$ T (poderá alterar o mapeamento da query para testar todas as hipóteses ou apenas a função hits)
- Na extensão dos hits, poderá considerar como critério estender enquanto a contribuição for positiva ou nula
- Poderá ainda implementar formas de normalizar o score considerando o tamanho do alinhamento, bem como uma função para poder ligar dois hits que tenham distâncias entre si pequenas (abaixo de um valor definido como parâmetro)