Trabalho 2 de Bioinformática

Alinhamento Pareado de Sequências

27 de Março de 2019

Entrega: 12/04/2019 até 24:00

Este trabalho é para ser submetido via Moodle. O trabalho será desenvolvido como trabalho extra-classe. Os testes e exames poderão conter 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 alinhamento pareado de sequências têm como objectivo organizar as sequências de forma a maximizar o nível de semelhança entre estas. Para tal podemos realizar operações de edição de alinhamento de símbolos, substituição, deleção e inserção. Este é um problema de otimização que têm como objetivo encontrar a melhor solução. No entanto, é frequente existir mais do que uma solução ótima.

No preenchimento da matriz de Score (S) sempre que na relação de recorrência do algoritmo de DP existirem duas possibilidades com o mesmo score máximo podemos vir a ter a possibilidade de alinhar as sequências de formas alternativas. Estes casos podem gerar múltiplas soluções ótimas.

No código desenvolvido nas aulas é contemplado apenas uma das soluções ótimas (pode dar-se o caso de facto só existir uma). O objetivo deste trabalho é estender esse código no sentido de contemplar múltiplas soluções ótimas.

2 Implementação

A) Alinhamento global com possibilidade de múltiplas soluções. Crie as funções **global_align_multiple_solutions** e **recover_global_align_multiple_solutions**. Na primeira função deverá permitir que a matriz trace-back (T) guarde múltiplas alternativas de movimentos (diagonal, vertical, horizontal). Na segunda função deverá retornar uma lista de alinhamentos ótimos e não apenas um único alinhamento ótimo.

Nota: A função max3t deverá ser alterada para em vez de retornar um único valor retornar agora uma lista de valores representando os movimentos com valores máximos. Para a função de recuperação de alinhamento é apresentado como sugestão o pseudo-código no final deste enunciado.

B) Alinhamento local com possibilidade de múltiplas soluções. Crie as funções **local_align_multiple_solutions** e **recover_local_align_multiple_solutions**. Na primeira função deverá permitir que a matriz trace-back (T) guarde múltiplas alternativas de movimentos (diagonal, vertical, horizontal). Na segunda função deverá retornar uma lista de alinhamentos ótimos e não apenas um único alinhamento ótimo. Note que a matriz T poderá ter agora quatro valores.

C) Escreva funções de teste para os dois casos anteriores demonstrando múltiplos alinhamentos ótimos.

C.1) Das sequências no ficheiro protein_sequence.fas indique um par sequências com vários alinhamentos ótimos. Use a matriz de substituição BLOSUM62 e um gap de -3.

C.2) Teste as sequências GATTACA e GCATGCT com match = 1, mismatch = -1 e gap = -1 e indique se contém mais do que um alinhamento ótimo.

D) Escreva as funções **compare_pairwise_global_align** e **compare_pairwise_local_align** que dada uma lista de sequências retorna uma matriz com os valores de score entre cada par de sequências. Faça o pretty print dessa matriz.

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 funções desenvolvidas 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

Para testar a funcionalidade C) e D) poderá usar como exemplo as sequências no ficheiro protein_sequences.fas.

```
 # PSEUDO-CODE
tmp_aligns = [["","",|seq1|, |seq2|]]   # || represents length
final_aligns = []
while tmp_aligns not_empty do
    align = tmp_aligns.pop()   # pop first element
    i = align[2]    # indices to the matrix T
    j = align[3]
    if i == 0 & j == 0 do # reached upper left cell
        final_aligns.push([align[0], align[1]])
    else
        for every move in T[i,j] do
            if t == 1 do
                new_tmp_align = [seq1[i-1] + align[0], seq2[j-1] + align[1], i-1, j-1]
            if t==3 do
                new_tmp_align = ["-" + align[0], seq2[j-1] + align[1], i, j-1]
            if t==2 do
                new_tmp_align = [seq1[i-1] + align[0], "-" + align[1], i-1, j]
        tmp_aligns.push(new_tmp_align) # add new incremented alignment

return final_aligns
```

Pseudo-código para alterar a função de recuperação dos vários alinhamentos possíveis. Note que terá que implementar segundo a sintaxe do Python e que algumas destas funções terão que ser adaptadas às mais adequadas em Python.