Genômica Computacional

Lista de Exercício 1

1. **Substrings repetidas**. Alinhamento local entre duas strings diferentes encontra pares de substrings das duas strings que possuem alta similaridade. É também importante encontrar substrings de uma determinada string com alta similaridade. Estas substrings representam substrings repetidas inexatas. Isto sugere que encontrar repetidas inexatas você deveria alinhar localmente uma string contra ela própria. Contudo, existe um problema com esta abordagem. Se nós realizamos um alinhamento local de uma string contra ela própria, a melhor substring será a string inteira. Mesmo utilizando outros valores da matriz, o melhor caminho pode ser fortemente influenciado pela diagonal principal. Existe uma forma simples que resolver este problema.

   (a) Encontre esta forma simples.

   (b) Ulizando a função de penalização de gap linear com $g = -2$, e a seguinte função de escore:

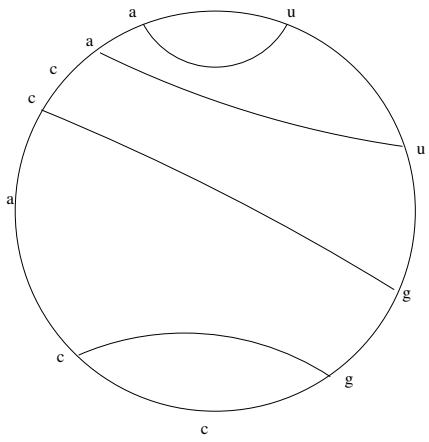
   $$s(x,y) = \left\{ \begin{array}{ll} 1 & se\ x = y \\ -1 & se\ x \neq y \end{array} \right)$$

   Utilize o seu algoritmo para encontrar ums substring inexata ótima em $S = gcgttccatagg$. Mostre a tabela de programação dinâmica resultante, incluindo as setas de "traceback".

   (c) Implemente a sua solução em Matlab (Você pode modificar a implementação existente do algoritmo de Waterman-Smith).

2. **Problema da Dobradura do tRNA**. A seguir nós apresentamos uma versão extremamente crua de um problema que surge na predição da estrutura secundária (planar) de móleculas de RNA transferidor. S é uma string de n carácteres no alfabeto RNA a, c, u, g . Nós definimos um pareamento como um conjunto de pares disjuntos de caracteres em S. Um pareamento é chamdo de apropriado se este contem somente pares da forma (a,u) ou (c,g) . Esta restrição surge porque em RNA a e u são nucleotídeos complementares, assim como c e g. Se nós desenhamos S, como uma string circular , nós pódemos definir um pareamento alinhado com um sendo um pareamento apropriado onde cada par no pareamento está conectado por uma linha dentro do círculo , e as linhas não se cruzam (veja a figura). O problema é encontrar um pareamento alinhado de maior cardinalidade. Frequentemente nós temos a restrição adicional que um carácter não pode formar um par com qualquer um dos seus dois vizinhos imediatos. Mostre como resolver este versão do problema de folding tRNA em tempo $O(n^3)$ utilizando programação dinâmica.
   Agora modifique o problema adicionando pesos a função objetivo de modo que o peso de um par $a - u$ é diferente de um par $c - g$. A meta agora

Figure 1: Pareamento apropriado



é encontrar um pareamento alinhado de peso total máximo. De um algoritmo eficiente para este problema ponderado.
Codifique o seu algoritmo em Matlab e aplique este aos dados do RNA transferidor da leucina no ser humano. Os dados podem ser obtidos da seguinte forma: vá para o NCBI Entrez, e encontre o registro com número de acesso X04700. Salve a sequência de 86 bases no formato FASTA. Mude todo T para U de modo a obter a sequência correta do RNA.
Encontre um pareamento ótimo se os pesos são uniformes. Encontre outro pareamento ótimo se os pesos são 3 para (a,u) e 1 para (c,g) .