Bioinformática

Exercícios TP3 : Comparação e alinhamento de sequências

1) O programa de comparação de sequências Dotlet (https://dotlet.vital-it.ch) permite identificar padrões de semelhança entre sequências de DNA ou proteína:

a) Obtenha, a partir de base de sequências Uniprot (www.uniprot.org), a sequência da tripsina 1 ("trypsin-1") humana e a da ratazana ("rat"). Faça a comparação das duas sequências no Dotlet usando os seguintes valores para os parâmetros: "Window size" igual a 15 e "Scoring matrix" igual a "BLOSUM 62".

b) Consegue distinguir um padrão claro no diagrama do dot plot? Varie o tamanho da janela ("Window size") e veja como este parâmetro afecta a visualização do dotplot. Utlize os sliders à direita do dotplot para ajustar a forma como os valores dos scores de comparação são mapeados em tons de cinzento no dotplot.

NOTA: o histograma a azul representa a distribuição de scores nas células do dotplot (a altura de cada barra do histograma denota o número de células com o score correspondente). O histograma deverá ter pelo menos um pico, o qual corresponde ao valor de maior ocorrência observado no diagrama (região de scores baixos). Um histograma com um 1 pico apenas indica que a distribuição de scores observada é a esperada num alinhamento aleatório, sugerindo a ausência de similaridades estatisticamente significantes entre as sequências. Um histograma com um segundo pico mais pequeno do lado direito indica scores potencialmente significantes (scores altos). A linha a vermelha representa o **logaritmo** dos valores do histograma, para facilitar a visualização do segundo pico, o qual pode ser muito pequeno quando comparado com o pico maior.

c) Repita o procedimento anterior, mas fazendo desta vez comparação da sequência da tripsina-1 humana com as sequências das tripsinas de salmão e da bactéria *Streptomyces griseus*. Compare os resultados da análise dos pares humana-salmão e humana-bactéria com o par humana-ratazana calculado anteriormente.

d) Para melhor entender o resultado da alínea anterior, calcule as percentagens de identidade de sequência para os 3 pares homem-ratazana, homem-salmão e homem-bactéria, utilizando o website do Uniprot (deverá selecionar as sequências pretendidas, adicioná-las ao "basket" (cesto de sequências) , selecionar um par de cada vez e usar o comando "Align".

e) A protrombina é uma proteína mais longa que a tripsina, mas que contém uma região bastante similar a esta última. Obtenha o precursor da protrombina a partir do Uniprot (código Uniprot: P00735) e faça a sua comparação com a tripsina humana no software Dotlet. Consegue identificar a região da sequência da protrombina que apresenta similaridade com a tripsina humana? Quais os números de sequência da início e fim desta região? (valores aproximados)

f) Os dotplots são particularmente bons para a deteção de repetições internas numa sequência (regiões de auto-similaridade), através da sua auto comparação (comparação da sequência com ela própria). Para sabermos identificar o tipo de padrões produzidos no Dotlet pela auto comparação de uma sequência com repetições, vamos começar com uma situação idealizada: sequência "artificial" constituída por várias repetições perfeitas de um determinado segmento. Geraremos esta sequência artificial a partir de repetições de um segmento de sequência aleatória de 200 aminoácidos. Para produzir este segmento, utilizamos a ferramenta "Randseq" na página http://expasy.org/tools/randseq.html, selecionando um comprimento de 200 aminoácidos e a opção de produzir o output em formato FASTA. Copie o resultado obtido para um ficheiro de texto utilizando o Bloco Notas do Windows. Neste ficheiro, produza três novas sequências, contendo respectivamente duas, três e quatro repetições da sequência aleatória original (utilize copy & paste). Faça a *auto-comparação* de cada uma dessas sequências com Dotlet e analise a relação entre o número de repetições e os padrões observados. (Nota: seleccione a matriz "Identity" da lista de matrizes disponíveis para o parâmetro "Scoring matrix" do Dotlet).

g) Como exemplo de uma situação real, faremos agora a auto-comparação do precursor da protrombina (código Uniprot: P00735), e também a auto-comparação do plasminogénio (P06868). Quantos regiões repetidas apresenta cada uma das sequências? (Esta pequenas regiões de sequência repetidas na família dos factores de coagulação são conhecidas como "domínios Kringle" (Kringle domains) e exemplificam a estrutura "modular" de muitas famílias de proteínas biológicas).

h) Faça a auto-comparação da Tripsina I humana com o Dotlet. Utilizando diferentes tamanos de janela ("Window size") e diferentes matrizes ("Scoring matrix") tenta identificar a possível existência de repetições internas na sequência da tripsina humana).

2) A detecção de similaridade entre sequências biológicas é um problema difícil e a inspecção visual das sequências e alinhamentos pode induzir em erro relativamente à sua real proximidade. Vamos exemplificar esta situação com as seguintes quatro sequências (que deverá obter a partir do site do Uniprot, www.uniprot.org): cadeias α e cadeia β da hemoglobina humana, leghemoglobina 1 do tremoço ("yellow lupin") e glutationo-S-transferase 7 (GST-7) do nemátodo *Caenorhabditis Elegans*.

a) Usando o alinhamento local (opção "water") em http://www.ebi.ac.uk/emboss/align, produza os 3 alinhamentos: cadeia α com a cadeia β, cadeia α com a leghemoglobina, e a cadeia α com GST-7. Tome nota das percentagens de identidade, semelhança, gaps e score em cada caso. Qual das sequências, leghemoglobina ou GST-7, parece ser mais aparentada com a cadeia α da hemoglobina?

b) Faça uma pesquisa BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins) usando como "query" a sequência da leghemoglobina 2. Consegue encontrar as cadeias α e β da hemoglobina humana?

c) Produza uma sequência aleatória de 100 aminoácidos de comprimento com Randseq e use-a como query em BLAST. Compare com o das pesquisas das alíneas anteriores.

d) Compare o alinhamento local e global da cadeia α com a leghemoglobina.

e) Quando as duas sequências apresentam uma similaridade fraca, pode ser difícil distinguir o alinhamento obtido de um outro produzido por duas sequências totalmente não aparentadas (alinhamento "aleatório"). Para simular o aspecto de um alinhamento verdadeiramente aleatório (sequências não-relacionadas, ou similaridade não detectável), vamos gerar duas sequências aleatórias de 200 aminoácidos com a ferramenta "Random Protein Sequence" do site SMS (https://www.bioinformatics.org/sms2/random_protein.html) Alinhe estas sequências local e globalmente, comparando os resultados com os obtidos em **a)**. Repita os alinhamentos algumas vezes, para ter uma ideia da variação possível dos parâmetros deste tipo de alinhamento (o alinhamento de sequências aleatórias é um bom modelo para o alinhamento de sequências biológicas não aparentadas).

3) Os algoritmos de alinnhmento óptimo de sequências de **Needleman-Wunsch** ou **Smith-Waterman** apresentam uma limitação importante: dado que o alinhamento progride numa direcção fixa, existem situações em que não é possível detectar todas as regiõe similares entre duas sequências. Isto pode acontecer porque as regiões similares ocorrem numa ordem diferente nas duas sequências, ou porque existem múltiplas cópias da região similar em uma ou nas duas sequências.

a) Vamos novamente recorrer au uma situação idealizada para ilustrar este problema. Comece por gerar duas sequências aleatórias, às quais chamaremos "A" e "B", com 50 aminoácidos de comprimento. Para tal, utilize a ferramenta "Random Protein Sequence" do problema anterior (Nota: não usar Randseq para este problema). Fazendo copy & paste num ficheiro de texto, gere duas sequências AB e BA, contendo os segmentos A e B nas duas ordens possíveis. Faça um alinhamento local e global das sequências AB e BA com as opções "needle" e "water" do webserver Emboss. Analize os resultados obtidos. Conseguem os dois algoritmos identificar todas as regiões similares entre as duas sequências?

b) Use o program Lalign (https://fastademo.bioch.virginia.edu/fasta_www2/fasta_www.cgi?rm=lalign) para listar alinhamentos sub-óptimos das sequências AB e BA. Compare os alinhamentos obtidos com Lalign com os alinhamentos anteriores. Terá este exemplo relevância biológica ?

c) Usando processo de a), construa uma sequência AAA e outra AB. Quantas regiões similares consegue detectar cada um dos algoritmos?

d) A detecção deste tipo de similaridades pode ser facilmente feita com um dotplot. Usando o **Dotlet**, faça a comparação dos pares de sequência usados em **a)** e **c)**.

4) Alinhe as sequências dos factores de coagulação IX e XII humanos local e globalmente. Use o programa Lalign (http://www.ch.embnet.org/software/LALIGN_form.html) para listar alinhamentos suboptimais. Compare os alinhamentos encontrados com as definições de domínio listadas nas entradas SWISSPROT dos factores IX e XII.