

Bioinformática

Exercícios TP5 : Pesquisa de sequências

1. Numa aula TP anterior, verificou-se que a leghemoglobina 1 do tremçoço (*Lupinus luteus*) apresenta uma similaridade de sequência muito fraca com a hemoglobina humana, produzindo alinhamentos de baixa significância. Todavia, a comparação das estruturas das duas proteínas revelou uma clara similaridade estrutural, evidenciando uma origem comum. Vamos tentar evidenciar essa relação numa pesquisa de sequências, usando o software de pesquisa [BLAST](#);
 - a) Faça uma pesquisa BLAST com os parâmetros por defeito, usando como “query” a leghemoglobina 1 de *Lupinus luteus*. Consegue encontrar as cadeias alfa e beta da hemoglobina na lista de resultados?
 - b) Na pesquisa realizada em a) deve ter obtido 100 resultados. Observe o primeiro resultado, considerando os parâmetros “query coverage” e “% identity”. Em face dos valores observados, que sequência deverá ser esta? Observe também o E-value deste primeiro hit.
 - c) Para verificar se se encontram entre os resultados as proteínas humanas (e particularmente as cadeias alfa e beta da hemoglobina), vá para a tab “Taxonomy” na lista de resultados e observe atentamente os grupos taxonómicos lá listados. Há alguma possibilidade de as cadeias da hemoglobina humana se encontrarem entre as 100 sequências listadas? Porquê?
 - d) Usando o botão “Edit Search” para voltar à página de submissão da sequência. Click no botão “Algorithm parameters” e observe os parâmetros “Max target sequences” e “Expect threshold” (limiar de E-value). Para obter um maior número de resultados, aumente o valor de “Max target sequences” para 1000. Repita a pesquisa. Quantas sequências são listadas? Existirá alguma hemoglobina humana entre as sequências listadas? (Use a tab “Taxonomy” novamente para verificar os grupos taxonómicos aos quais pertencem os organismos listados).
 - e) Existe claramente um número de sequências não humanas muito vasto, as quais aparecem como resultados nesta pesquisa. Isto acontece porque, por defeito, a pesquisa está a ser feita na base de dados **nr** (non-redundant protein sequences), a qual contém mais de 600 milhões de sequências. Para além do enorme número de resultados obtidos, os quais dificultam a identificação dos potenciais “hits” humanos, as pesquisas são muito mais demoradas por serem feitas num espaço de sequências tão vasto. Antes de fazermos qualquer outra pesquisa, vamos restringir as entradas a pesquisar na base de dados ao conjunto das sequências humanas. Para tal, digite “Homo Sapiens” na caixa “Organism”, a qual se encontra imediatamente abaixo da caixa “Database”.
 - f) Mantendo inalterados os anteriores parâmetros, realize nova pesquisa. O que observa? Porquê?
 - g) Use o botão “Edit Search” para voltar à página de entrada. Nesta última pesquisa, a ausência de resultados poderá indicar um “Expect Threshold” demasiado baixo, limitando a exibição de sequências com E-values mais altos, de entre as quais poderão estar as sequências dejetadas na nossa pesquisa. O E-value por defeito é 0.05, um valor que garante uma taxa de falsos positivos relativamente baixa, mas a troca de uma sensibilidade também baixa. Vamos alterar este valor para “1000” e repetir a pesquisa. O que observa neste caso? Os valores de E-value são consistentes com a ausência de resultados na pesquisa anterior?
 - h) A lista de hits apresenta apenas 185 resultados, e não 1000. Porquê?

- i) Na lista de resultados, tente encontrar sequências de globinas (sugestão: use CTRL-F para pesquisar na lista de *hits*, usando a palavra “globin”).
 - j) Repita esta pesquisa com valores mais elevados de “Expect Threshold” e “Max target sequences”. Consegue encontrar alguma globina?
 - k) O parâmetro “Word size” afecta a sensibilidade do algoritmo de BLAST. Baixando o valor por defeito, podemos obter uma maior sensibilidade a troco de uma menor velocidade no desempenho da pesquisa. Reduza o “Word size” para 2, ajustando o “Expect Threshold” e “Max target sequences” para os seus valores por defeito (respectivamente 0.05 e 100). Quantas hits obtém agora? Aumente o “Expect threshold” para 1000.
 - l) No limite máximo de sequências exibidas (5000) e mediante um “Expect threshold” suficientemente alto, deverá ser possível encontrar a cadeia beta da hemoglobina (entre várias outras formas da hemoglobina e também outras proteínas da família das globinas). Note a presença óbvia de inúmeros falsos positivos. Estes falsos positivos apresentam frequentemente alinhamentos com uma percentagem de identidade elevada, mas com um comprimento da região alinhada muito curto, conduzindo a scores muito baixos (e E-values extremamente altos).
2. No número anterior verificou-se a necessidade de aumentar a sensibilidade do software BLAST (mediante ajuste do “Word size”) de modo a permitir detetar a similaridade entre a leghemoglobina e as cadeias alfa e beta da hemoglobina humana. Na verdade, apenas a cadeia beta pode ser detetada dentro do número máximo de sequências exibidas (“Max target sequences”) permitido pela versão web do software BLAST. Para identificação de homólogas mais distantes (como a cadeia alfa da hemoglobina humana) torna-se necessário recorrer a métodos mais sensíveis. Um destes métodos encontra-se implementado na suite de BLAST com a designação PSI-BLAST.
- a) Faça uma pesquisa usando como sequência de busca a leghemoglobina 1 e os parâmetros por defeito de BLAST, mas usando a opção “PSI-BLAST”. Que resultados obtém?
 - b) Aumente a sensibilidade da busca usando um “Expect Threshold” de 100. Observe as sequências listadas pelo PSI-Blast. Selecione para a segunda interação as sequências de globulinas. Repita o processo várias iterações de PSI-BLAST e verifique se há convergência dos resultados.
 - c) Usa a opção de filtrar de PSI-BLAST para eliminar resultados de outras famílias de proteínas que não as globulinas (potenciais falsos positivos).
3. Obtenha uma lista de sequências da “Matrix Gla Protein” (MGP) a partir do portal Uniprot (www.uniprot.org), (**Search in:** UniprotKB ; **Query:** matrix gla protein)
- a) Examine os resultados. Parece-lhe que todas as sequências obtidas pertencem à família das MGP’s. Porquê?
 - b) Repita a pesquisa anterior, mas sem preencher a caixa **Query**, use a opção **Advanced Search** usando como **Field** “Protein Name” e preenchendo a caixa correspondente com “matrix gla protein”. Compare com os resultados obtidos anteriormente.
 - c) Seleccione a MGP humana, faça uma pesquisa BLAST com esta sequência contra toda a base de dados Uniprot usando o portal SRS (www.ebi.ac.uk/blast2) (no Step3 – Parameters, mude o valor “Scores” de 50 para 100. Uma das sequências obtidas na alínea a não deveria aparecer na lista agora obtida, a sequência *MGP_PRIGL* do “blue shark”. Qual será a razão porque esta sequência não é agora listada ?
 - d) Use a sequência *MGP_PRIGL* para pesquisar com blast2 a base de dados UniprotKB. Analise os resultados e copie uma das sequências obtidas, a *A8YQS4_PRIGL*
 - e) Use a sequência gravada na alínea anterior para uma nova pesquisa blast2 contra UniprotKB.

- f) De acordo com os resultados obtidos em c), não é claro se a sequência *D2BNG8_LACLK* pertence de facto à família das MGPs. Para responder a esta questão, faça uma nova pesquisa blast2 desta sequência contra a base de dados UniprotKB e analise o resultado obtido.
- g) Produza uma sequência aleatória de 100 aminoácidos de comprimento com Randseq e use-a como query em blast2. Compare os resultados com os obtidos nas alíneas anteriores.