Bioinformatics

Exercises – Sequence comparison and alignment

1. The sequence comparison program Dotlet (http://myhits.isb-sib.ch/cgi-bin/dotlet) allows for the identification of similarity patterns between two protein or DNA sequences:
   a) Using the Uniprot database (www.uniprot.org) retrieve the human and rat trypsin-1 sequences, and compare them using Dotlet (use a windows of size 5 and an identity matrix). What can you see ? Use the Uniprot site to compute the percent identity between the two sequences.
   b) Repeat the procedure in a), but comparing now the human trypsin with those from salmon and *Streptomyces griseus* (a bacterium). Compare with the previous results.
   c) Compare the prothrombin precursor (Uniprot code **P00735**) with the human trypsin sequence. Can you understand the result ?
   d) Generate a 200 amino acid random sequence using the Randseq tool (http://expasy.org/tools/randseq.html), and store it in a FASTA text file. Create three sequences with 400, 600 and 800 amino acids using respectively a two, three and four repeat of the original sequence - store these sequences in the text file. Do *self*-comparisons of these last three sequences with Dotlet. How does the number of repeats correlate with the observed patterns ?
   e) Run a self-comparison of human Trypsin-1 with Dotlet. Can you conclude something from the results ?
   f) Run a self comparison of the prothrombin precursors (Uniprot code P00735), and plasminogen (P06868). How many Kringle domains can you identify in each one ? (Kringle domain are short repeat sequences that can be found in some blood plasma proteins)
   g) Retrieve the sequence of the calmodulin gene from the fungus *Aspergillus nidulans* (GeneBank code J05545) and the corresponding protein sequence (P60204). Use Dotlet to find the position of the introns within the calmodulin gene (Suggestion: use a high selectivity matrix, Blosum100, and a small-sized window, 7).

2. Retrieve the following sequences: α and β hemoglobin chains, leghemoglobin 1 from yellow lupin and Glutathione S-transferase 2 (GST-2) from *Caenorhabditis Elegans*.

a) Using the local alignment tool (option "water") from http://www.ebi.ac.uk/emboss/align, create the following 3 alignments: α chain with β chain α chain with leghemoglobin, and α chain with GST-2. Write down the percent identities, similarities and gaps for each alignment. Which of the sequences, leghemoglobin or GST-2, seems more related with the hemoglobin α chain ?
b) Run a BLAST (http://www.ncbi.nlm.nih.gov/BLAST) search, using as query the leghemoglobin sequence. Can you find the hemoglobin α or β chains ?
c) Compare the local and global alignments of the hemoglobin α chain against leghemoglobin.
d) Using the Randseq tool, generate two random sequences of lenght 200 and align them locally and globally, comparing the results with those obtained in a).