Bioinformatics

Exercises – Sequence comparison and alignment

1) The sequence comparison software Dotlet (https://dotlet.vital-it.ch) can be used to identify similarity patterns between DNA or protein sequences:

a) From the Uniprot protein sequence database (https://www.uniprot.org) retrieve the human trypsin-1 sequence and the rat trypsin sequence (you can use these terms in the search box). Compare the two sequences using the Dotlet software with the following parameters: "Window Size" 15, and "Scoring Matrix" BLOSUM62

b) Are you able to discern a clear pattern in the dotplot diagram? Change the size of the comparison window ("Window Size" parameter) and observe how it affects the visual representation of the dotplot. Use the sliders to the right of the dotplot windows to change the mapping of comparison scores to shades of gray on the dotplot window.

NOTE: the blue histogram represents the distribution of scores over the dotplot cells (the heights of the bars on the histogram represent the number of cells with a given score). The histogram should have at least on peak, corresponding to the most frequently observed score on the dotplot (normally a low score). An histogram with just one peak is to be expected for the comparison of unrelated sequences, where there are no statistically significant similarities. If the histogram displays a second peak to the right of the first peak (higher scores), this is and indication of potentially significant similarities producing scores that are far higher than expected in a comparison of unrelated sequences. The red line is the logarithm of the histogram bar heights , a "compressed " version of the histogram so that low height peaks are easier to spot.

c) Repeat the previous procedure but this time comparing the human trypsin-1 sequence with those of a fish trypsin (Salmon) and a bacterial trypsin (*Streptomyces griseus).* Compare the diagrams for human-bacteria and human-fish with the previous human-rat diagram.

d) To have a better insight on the previous results, compute the percent identities for the three pairs of sequences: human-rat, human-salmon and human-*S.griseus,* using the alignment tool on the Uniprot website (you should select the required sequences, add them to the "sequence basket" and, selecting each pair in turn, use the command "Align").

e) Even though prothrombin is a much longer protein than trypsin, it contains a segment that is highly similar to the latter. Retrieve the sequence of the prothrombin precursor from Uniprot (Uniprot code: P00735) and compare it with the human trypsin sequence. Are you able to identify the prothrombin segment that is similar to trypsin. What are the approximate sequence numbers for the first and last aminoacids of this segment?

f) Dotplots are especially good at detecting internal repetitions within a sequence (self-similarity regions), by means of sequence self-comparison. In order to learn how to interpret self-similarity patterns, we shall start with an idealized situation: artificially crafted sequences containing a varying number of perfectly identical repeats. To avoid spurious similarities, the repeating segment will be a random sequence of 200 aminoacids produced with the Randseq tool (http://expasy.org/tools/randseq.html). After generating this sequence (make sure you selected the "FASTA" format in Randseq), produce sequences of 2, 3 and 4 identical repeats by cutting and pasting the random segment on a text file (e.g., on the Windows Notepad app). Perform the self-comparison of these three sequences using Dotlet. Compare the observed patterns and figure out how to calculate the number of repeats from the observed pattern. **NOTE:** for this particular case, select "Identify" from the choice of scoring matrices in Dotlet.

g) Now we move to a real-world example: the self-comparison of the prothrombin precursor (Unirpot code: P00735) and plasminogen (Uniprot code: P06868). How many self-identical regions did you find withing each of these protein sequences? Are they perfect matches? Can you sketch a diagram of the sequences indicating the approximate location and number of repeats?

h) Now we attempt a much harder example of self-similarity detection: human trypsin-1. Performs the self-comparison of human trypsin-1 in Dotlet. Try to use different windows sizes and scoring matrices and try also to play with the color mapping of scores by using the sliders to the right of the dotplot. Use the histogram to assess the significance of the observed results. Can you conclude anything?

2) The detection of similarities between biological sequences is not an easy problem and visual inspection of aligned sequences may lead to a wrong appraisal of their real similarity. We shall exemplify the problem with following 4 sequences: α and β chains of human hemoglobin, leghemoglobin from yellow lupin and the glutathione-S-transferase 2 (GST-2) from *Caenorhabditis Elegans*.

a) Using the local alignment option (option "water") in http://www.ebi.ac.uk/emboss/align, produce the following three alignments: hemoglobin α chain with β chain, hemoglobin α chain with leghemoglobin and hemoglobin α chain with GST-2. Write down the percent identities, percent similarity, score and percentage of gaps. Which of the sequences, leghemoglobin or GST-2, appear to be the most related to chain α of hemoglobin ?

b) Run a Blast search (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins) using the sequence of yellow lupin hemoglobin as query. Can you find the α and β chain among the hit results of Blast search ?

c) Unsing the Randseq tool, generate a random sequence of 100 amino acids and perform a Blast search with it. Compare with the results of the previous search.

d) Using the Emboss sit, compare the local (option "water") and global (option "needle") alignments of the α and β chains of hemoglobin.

e) When two sequences are only weakly similar, it may be difficult to distinguish their alignment from that of a pair of completely unrelated sequences (random alignment). In order to simulate the appearance of a random alignment, generate two sequences of length 200 using the "Random Protein Sequence" tool from the site SMS (https://www.bioinformatics.org/sms2/random_protein.html). Align the two random sequences locally and globally, and compare with the results obtained in **d)** . Repeat the procedure two or three times, to get a better impression of the extent of variation of the alignment parameters (the alignment of random sequences is a good model for the alignment of truly unrelated biological sequences).

3) The Needleman-Wunsch and Smith-Waterman optimal sequence alignment algorithms have one important shortcoming: since the alignment moves forward in one direction only, there are situation where it may be unable to reveal all the similar regions between two sequences. This may happen for instance if the similar regions occur in a different order in the two sequences (like in domain-swapped proteins) or if multiple copies of a similar region are present in one or both sequences.

a) We shall again make use of an idealized situation to demonstrate the problem. Generate two random sequences of length 50 that we will call "A" and "B". Copying and pasting on a text file, produce to sequences containing both segments it the two possible orders, AB and BA (NOTE: use the SMS random sequence generator for this exercise, not Randseq). Align the sequence AB and BA locally and globally in Emboss. Are the two alignment algorithms showing all possible matches between the two sequences AB and BA ?

b) Use the sub-optimal alignment finder program Lalign (http://www.ch.embnet.org/software/LALIGN_form.html) to list the suboptimal aligments of sequences AB and BA. Compare the results with those obtained in the previous exercise. Do you think this example may be biologically relevant?

c) Using the procedure described in a), generate two sequences of the type AAA and AB. How many similar regions can be detected between the two sequences?

d) Detection of this type of similarities can easily be achieved with dotplots. Use to Dotlet to compare the sequence pairs of **a)** and **c)**.

4) Align the sequences of the human coagulation factors IX and XII locally and globally. Use Lalign list the suboptimal alignments. Compare the resulting alignments with the domain definitions listed on the Uniprot entries of the two proteins.