BIN 506: Protein & DNA Sequence Analysis

Assignment 02

By: Taha Ahmad

Student ID: 2546125

Instructor: Dr Yesim Aydin Son

1. What are the differences between local, global, and global with no end-gap penalty alignment methods?

Local alignment:

Local alignment is a dynamic programming algorithm that works by finding the best matching sequences within the larger input sequence ignoring regions of low similarity. Local alignment will penalise gaps anywhere in the alignment . local alignment is ideal for detecting conserved functional domains or motifs.

Global alignment:

Another dynamic programming algorithm which aligns sequences in an end-to-end manner, hence optimizing the entire length. Suitable for highly similar sequences of comparable lengths. Penalises gaps throughout, including termini. The algorithm is most useful when comparing closely related genes (homologous genes) or evolutionary studies which require full length alignment

Global with no end gap penalty:

Same working principle as Global Alignment however it does not penalise gaps at the start or end (terminal) hence making the algorithm useful for sequences with variable terminal ends but conserved internal regions. It is most useful when aligning sequences where ends are less significant or identifying short sequences integrated into a longer one e.g. primer binding in PCR.

2. Between local and global alignment methods, which one provides biologically more relevant information? Explain your reasoning.

Between local and global alignment tools local alignment methods generally provide more biologically relevant information in more practical situation. Especially when analysing sequences and identifying functional domains because of the following reasons.
Detecting conserved functional regions : local alignment tools like smith waterman succeed at identifying short regions of high similarity like functional domains amongst dissimilar sequences .These regions of similarity are critical for biological functions even if the overall region shares little similarity. These function domains may include but are not restricted to enzyme active sites and binding motifs.
Flexibility with structural variations: local alignment facilitates insertions, deletions and rearrangements which are common in biological sequences, thus making it suitable for comparing modular proteins or genes with shuffled exons..
Handling divergent sequences: this is especially practical in evolutionary studies where sequences from distantly related organisms often retain only key conserved regions.. Local alignment avoids strictly implementing an end to end match thus reducing noise from non homologous regions and focusing only on biologically meaningful similarities.
Thus local alignment is more broadly applicable for extracting biologically relevant insights, as it allows us to identify critical functional or evolutionary signatures without global similarity,

3. Use bl2seq to align NP_000509.1 to NP_000549.1

a) Which BLAST program did you choose? Why?
I chose BLASTP program because both sequences are proteins and BLASTP is specifically designed to compare amino acid sequences against protein databases

b) What is the sequence type in the final alignment?
The sequence type in the final alignment is amino acid sequences, however if the BLAST program was BLASTN then the sequences would be

DNA/RNA however BLASTP was chosen in this specific workflow as the input sequences are proteins.

c) What are the significance values and the total score of the alignment? Define them, and comment on the result you get.
Significance Value (E-value): 2e-38
Total Score: 114
Percentage Identity: 43.45%
Query Cover: 97%
The e-value of **2e-38** is very low which indicates a statistically significant alignment. It also suggests that the similarity is not due to random chance. The total score of **114** suggests moderate alignment strength. The **43.45%** identity indicates that almost half of the amino acids are identical in the alignment suggesting a degree of functional or evolutionary similarity. The **97%** query cover means that most of **NP-000509.1** aligns with **NP-000549.1**. This denotes significant overlap between the two protein sequences.

4. Use bl2seq to align NM_000558.3 to NP_000549.1

a) Which BLAST program did you choose? Why?
Since NM_000558.3 is a nucleotide sequence and NP_000549.1 is a protein sequence I chose the BLASTx program. BLASTx translates the nucleotide sequence in all six reading frames and compares it to protein sequences. This is very useful when working with DNA or RNA sequences and wanting to find potential protein products.

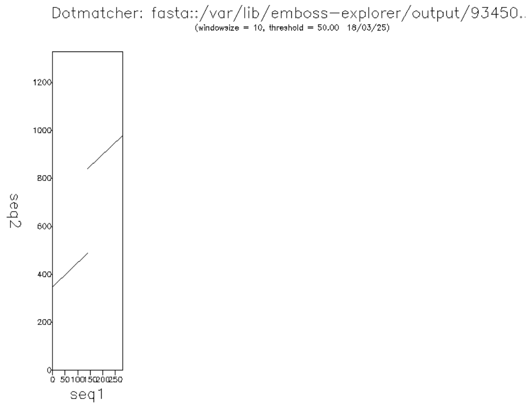
b) What is the sequence type in the final alignment?
The sequence type in the final alignment is protein. This is because BLASTx translates the nucleotide query sequence into amino acid sequences prior to the alignment of protein sequences from the database.

c) What are the significance value and the total score of the alignment? Comment on the result you get.

Significance Value (E-value): 2e-105
Total Score: 286
Percentage Identity: 100%
Query Cover: 74%
The e-value of **2e-105** is very low, this denotes that the alignment is highly significant and it is very unlikely that it has occurred by change. The total score of **286** further confirms strong alignment. The percent identity (Per. Ident) is **100%**, this means the sequences are identical over the aligned regions which essentially covers **74%** of the query sequence.
In short, the results suggest the nucleotide sequence **NPL000558.3** translates into a protein sequence which is similar to the protein sequence **NP_000549.1** over the aligned region. Therefore this indicates a strong match and a possibility of similar functionality.

5. Use seq1.fasta to seq2.fasta to answer the following questions.

a) Use dotmatcher to align the sequences with the given parameters and interpret the graph: Window size = 10, Threshold = 50.



The dotplot highlights a centrally conserved region between the two sequences. The threshold of 50 ensures high confidence matches. The gaps in the dotplot lines indicate regions where the sequences are diverging. These gaps are biologically meaningful as they indicate evolutionary changes or structural variations. The continuous diagonals highlight conserved regions which are critical for function.

b) <u>Use EMBOSS Needle and EMBOSS Water to align the sequences. Compare and elaborate on the results. Explain why the resulting alignment is shorter in length in EMBOSS Water.</u>

1. Alignment length:
   -EMBOSS Needle: 1330 residues
   -EMBOSS Water : 630 residues

   Explanation: the reason for the difference in alignment length of both tools is because of the fact that EMBOSS Water performs local alignment as it focuses only on highest similarity regions while ignoring the poorly conserved regions. EMBOSS Needle on the other hand performs a global alignment , thus stretching the alignment over the full length of both the sequences regardless of gaps.

2. Identity,Similarity,Gaps:
   -EMBOSS Needle: 21.1% identity (280/1330), 78.9% gaps.
   -EMBOSS Water: 44.4% identity (280/630), 55.6% gaps.

   Explanation: EMBOSS Water discards low-similarity regions which results in a shorter alignment but higher identity in the conserved regions. EMBOSS Needle on the other hand retains all the residues which also includes non matching regions.

3. Alignment Structure:
   - EMBOSS Needle: As shown in our output file Needle Alignment includes long stretched of gaps (- - - -) , forcing alignment from start to end
   - EMBOSS Water: Alings only a central conserved block and skipping divergent regions.

c) Look at the dotpot and the alignment results. Do the alignment results satisfy the information provided by the dotplot? Explain briefly.

Yes, the dot plot results are complementary to the alignment results:

- Local Alignment (EMBOSS Water) : The short , high identity block in water aligns with the central diagonal region in the dot plot. This region likely represents the conserved functional domain .
- Global Alignment (EMBOSS Needle) : The full length alignment (1330 residues) consists of divergent regions with low similarity (21.1% identity) which is shown in the dotplot by the scattered dots outside the central diagonal line in the dotplot