Part a. Pairwise sequence alignments

Use the EMBOSS tools for sequence analysis at Wageningen University Laboratory of

Bioinformatics: http://emboss.bioinformatics.nl or at http://www.ebi.ac.uk/Tools/emboss/

Sequences can be compared by either global or local alignment methods. Global alignment

forces complete alignment of input sequences, whereas local alignment aligns only their

most similar sequences. In this exercise you will use two Dynamic Programming (DP)

algorithms for alignment and compare the result. There are two main variants of the DP

algorithm:

• Global alignment (Needleman-Wunsch algorithm) — the sequences are forced to be

aligned across their entire length. EMBOSS program Needle.

• Local alignment (Smith-Waterman algorithm) — only the best matching sub-part of

the sequences are aligned). EMBOSS program Water.

Align the sequences O32218 and Q9EYL5 with NeedleP and WaterP. Answer the following

questions.

a) What is the main difference between the results from using the two programs (NeedleP

and WaterP)? (look at the different scores (Identity, Similarity and Gaps)

As the table shown below the main difference is the value of gaps which is higher from the program Needle.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Identity | Similarity | Gaps |
| Water | 72/223 (32.3%) | 109/223 (48.9%) | 44/223 (19.7%) |
| Needle | 75/238 (31.5%) | 112/238 (47.1%) | 55/238 (23.1%) |

b) What is the difference between Identity and Similarity?

The identity and similarity from water are both less than Needle.

c) Modify the gap penalties. Explain how these parameters influence the score, occurrence

and length of gaps.

As the gap penalties increase, the score could be lower and occurrence becomes lower as well, but the length of gaps become shorter which are observed from the alignment results.

d) Explain why one cannot let gaps/insertions occur without penalty.

e) The gap penalty is composed of two components. Which?

opening(gap existence), gap extension

f) What is the difference between affine gap penalties and linear gap penalties?

Affine gap penalty is length dependent, and use a gap opening and a gap extension with different values, so it can be shown that in this penalty a few large gaps and many small gaps. Unlike the liner penalty.

g) How would you set the parameters to avoid long gaps in your alignment?

Setting a smaller extension value.

h) Do a local alignment and modify the scoring matrices used. Try in the optional fields

(Matrix file) to use BLOSUM35 and BLOSUM80 (In EMBOSS these are called EBLOSUM35 and

EBLSOMUM80, respectively. For comparison this is a link to BLOSUM62). How and why does

changing the scoring matrix influence the alignment scores? Which is the default matrix?

i) Align the sequences P54937 and Q01786 with NeedleP and WaterP. Which of the two

alignments is likely to be the most biologically relevant? Motivate your answer.

The results from Water is more likely to be biologically relevant since the gaps are much less. Alignment could be meaningless if there are many gaps.

j) Pairwise sequence alignment can be made by so called word methods which is a heuristic

algorithm. This method is used in the data base searching program BLAST. Here you will

compare the alignment of P54937 and Q01786 as well as that of NM\_000518 and M21825.

BLAST2seq (select Align two or more sequences). What differences from the previous

alignments did you find? (Use BLASTN for DNA and BLASTP for protein).

What is the principle difference between the dynamic programming based algorithms and the word based BLAST algorithm? (We discuss this in more detail in Sequence analysis part II).

Dot plot sequence comparison

1. Comparing a protein sequence with itself.

Get the sequence with the following accession number from the UniProtKB/Swiss-Prot

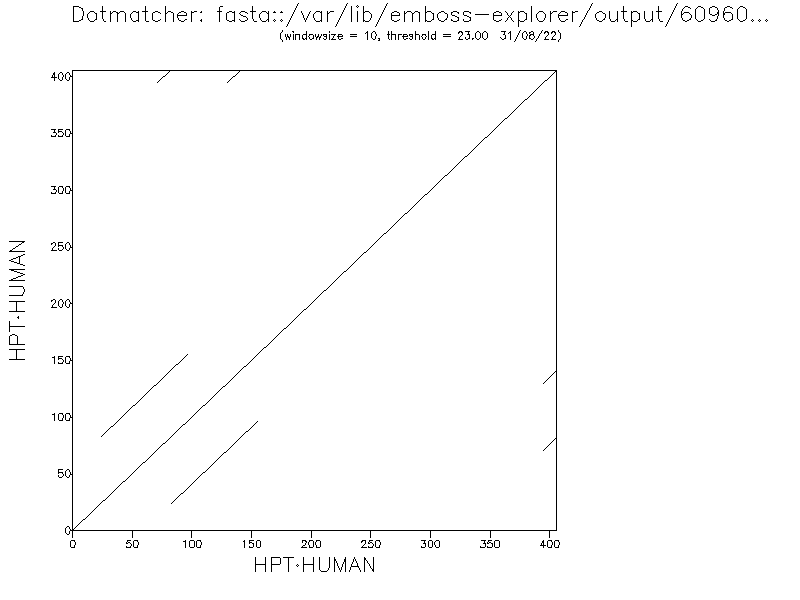
database: P00738. Use the program Dotlet Js or Dotmatcher to make a dotplot analysis of

the sequence. A self-comparison of sequences by dot a matrix comparison can be useful for

finding repeated patterns. Examine the effect of changing the window size and the threshold

(stringency). Do a comparison with the default values. Do you find the presence of any

repeated elements and where are they in the sequence? Yes, there is one repeated elementin the position of seqence 1:29 and sequence 2:88. when the window size change



Get UniProtKB/Swiss-Prot entry P69193. Do an analysis using Dotlet Js or Dotmatcher.

Explain the result.

2. Comparing DNA sequences.

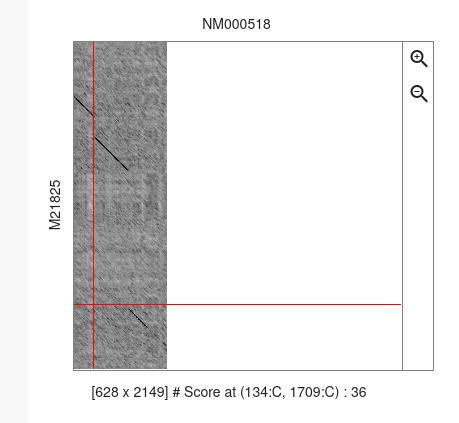
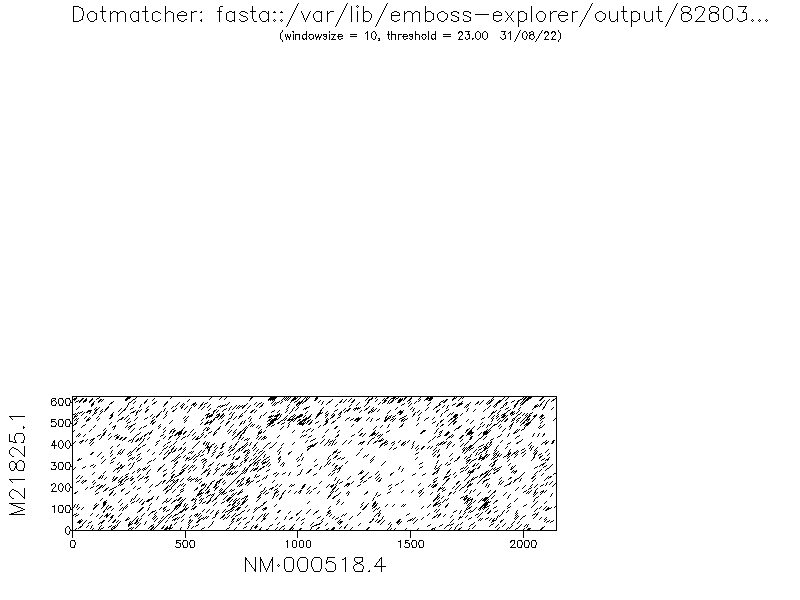
Retrieve the Genbank entries NM\_000518 and M21825. Do a dot plot analysis. Use different

window size and thresholds. Find the corresponding protein sequences and do a dot plot.

Interpret the results and explain the main differences between using the nucleotide

sequences and the protein sequences? The main difference is that

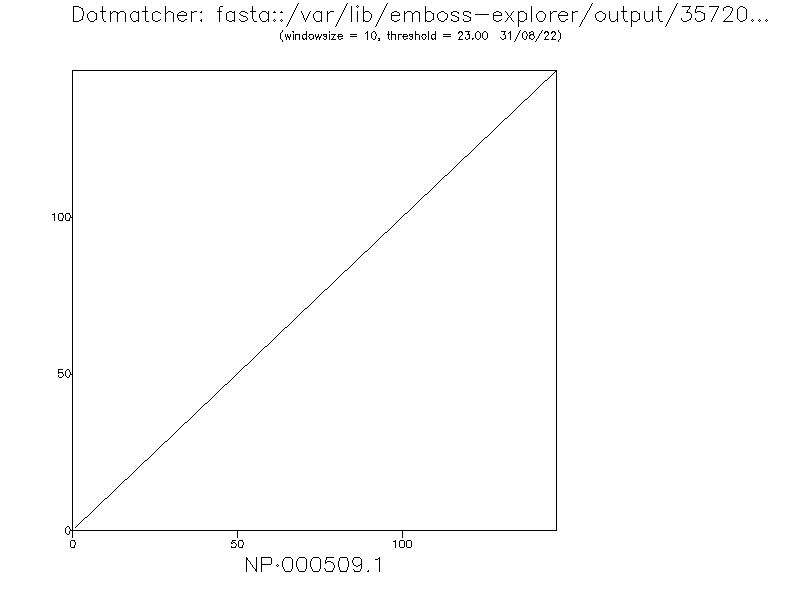
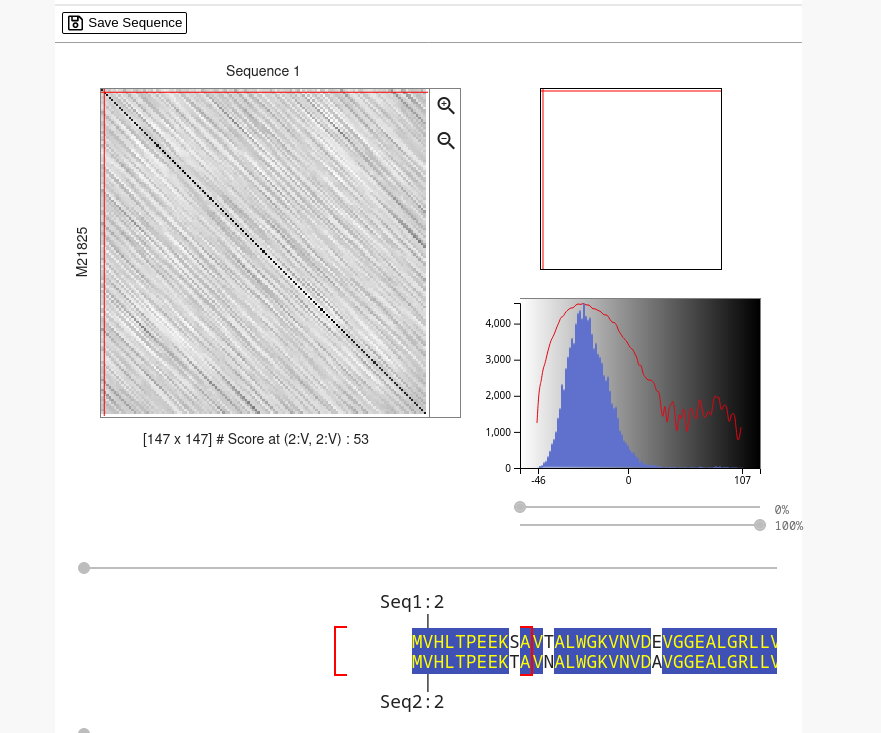
DNA sequences



NM000518 DNA Protein

M21825 DNA Protein

protein sequences



Manual sequence alignment using dynamic programming.   
1. Make an optimal alignment of ACTCG with ACAGTAG.   
The gap penalty is -1, the match score is +1 and the mismatch score is 0. Use the   
Needleman-Wunsch algorithm. Show the complete scores table!

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 0 | A | C | T | C | G |
| A |  |  |  |  |  |
| C |  |  |  |  |  |
| A |  |  |  |  |  |
| G |  |  |  |  |  |
| T |  |  |  |  |  |
| A |  |  |  |  |  |
| G |  |  |  |  |  |

2. What was the gap penalty, the match score and the mismatch score that was used to   
produce the scores table shown below?

3. Which are the two alternative optimal alignments?   
4. What is the difference between the Needleman-Wunsch algorithm and the Smith-  
Waterman algorithm?   
5. Make an optimal alignment of AACCTATAGCT with GCGATATA. The gap penalty is -1,   
the match score is +1 and the mismatch score is -1. Use the Smith-Waterman   
algorithm. Show the scores table