

Exercises - Week 8

Genomics and bioinformatics

1 MSA with ClustalW

Here is a summary of the main steps followed by ClustalW to generate the MSA. ClustalW first computes the optimal global alignment for every pair of sequences and then the distance score is set to be $1 - y/x$, where x and y are the number of non-gap positions and the number of identical positions, respectively. The guide tree is then built according to the distance matrix (and not the raw alignment scores) using the neighbour joining algorithm presented in the lectures. The last step involves doing profile-profile alignments and may be computed using dynamic programming to with some PSP score.

```
>sequence1
PPGVKSDCAS
>sequence2
PADGVKDCAS
>sequence3
PPDGKSDS
>sequence4
GADGKDCCS
>sequence5
GADGKDCAS
```

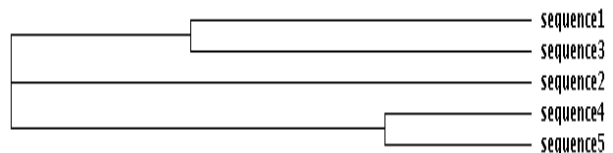


SeqA	Name	Length	SeqB	Name	Length	Score
1	sequence1	10	2	sequence2	10	50.0
1	sequence1	10	3	sequence3	8	62.5
1	sequence1	10	4	sequence4	9	33.33
1	sequence1	10	5	sequence5	9	44.44
2	sequence2	10	3	sequence3	8	50.0
2	sequence2	10	4	sequence4	9	44.44
2	sequence2	10	5	sequence5	9	55.56
3	sequence3	8	4	sequence4	9	37.5
3	sequence3	8	5	sequence5	9	37.5
4	sequence4	9	5	sequence5	9	88.89



CLUSTAL 2.1 multiple sequence alignment

```
sequence1      PPGVKSDCAS 10
sequence3      PPDGKSD--S 8
sequence2      PADGVKDCAS 10
sequence4      GADGK-DCCS 9
sequence5      GADGK-DCAS 9
               .. * *
```



Using the PAM250 scoring matrix with a gap penalty of -10, one can compute the SP score of the MSA by hands using the formula presented in the course. It is however more convenient to write a code (see `MSA_SP_score.py`) to do this task. We find a SP score of 101.