Home assignment - Master of Computer Science

Bioinformatics - 2002WETBIN

1 Local and global alignment with n sequences

Write a small program or script (preferably in Python, include clear install and/or compile instructions if you wish to use another language). The script should determine the optimal Smith-Waterman local alignment, the optimal Needleman-Wunsch global alignment, and their alignment scores for an arbitrary number of DNA/protein sequences.

The dynamic programming algorithm is completely analogous to the pairwise case, only now the scores for each position are equal to the sum of the individual pairwise comparisons (i.e. a position that is identical for three sequences has a score of $5_{s1,s2} + 5_{s1,s3} + 5_{s2,s3} = 15$).

- Calculate the score with the following scoring weights, and make sure you can easily change them (can be in the code, command line arguments are not required for this):
  - match: 5
  - mismatch: -2
  - indel (linear gap penalty): -4
  - two gaps (i.e. position 7 in s1 and s2 in the example below): 0
- Add an example input and output file using three sequences to illustrate your program. But make sure it also works for any number of (short) sequences!
- Your program should accept FASTA files as input.
- The output alignment should be in the following format. Each line starts with the **sequence ID** from the FASTA file (in this example s1, s2, and s3).

```
s1: ACTG.GT.CA
s2: .CAGGGT.CA
s3: CCAGGGACCA
```

- Make sure to also output the final alignment score

2 Application on biological problem

T cells are part of the adaptive immune system and are responsible for recognizing pathogens. 2 protein chains of the T cell receptor: the T cell receptor alpha (TCA) chain and T cell receptor beta (TCB) chain, are involved in the recognition, they both have a J region. Both sequences have multiple versions with small variations. The file *cs_assignment.fasta* contains 2 TCB J protein sequences and 1 TCA J protein sequence.

Perform these two tasks:

- Using global MSA, find which are the 2 TCB sequences and which is the TCA sequence.
- Using local MSA, find a conserved region in all three sequences.

Extra instructions:

- Use your own MSA from part 1 if you got it to work. Use the default parameters for the global alignment and change the mismatch score to -4 for the local alignment.
- If your own implementation does not work, use an online MSA tool (as seen in the practicals). Pick the one that you think will work best. Ready-to-use local MSA tools might not be available online, use a global MSA for both questions and explain how you found the conserved region.

3 Assignment

We expect you to hand in 1 .zip file with the following items:

- The source code, with clear comments.
- The output of your script for the given input (part 1).
- A report (2-3 pages) with:
  - A brief description of the idea of your algorithm and the time and memory complexity of your code (part 1).
  - The output of the alignments from part 2 (in the report or as a separate file), the answers to the questions, and a brief explanation of how you got them.

4 Submission

Assignments must be submitted to vincent.vandeuren@uantwerpen.be. Format your mail title as bioinformatics 2024 cs assignment name surname.

The deadline for handing in your assignment will be communicated during the class.

5 FAQ

- *Can I use an existing library?* Standard libraries are allowed for things like data structures, but you have to implement the actual alignment algorithm yourself. When in doubt, ask us for confirmation.