MB 1A Assessed Practical 3: Dynamic programming

Andrew Firth

Assignment Due: 17.00 on 15 May 2023

Tasks

In Practical 16, you developed a dynamic programming algorithm for performing a ***global*** alignment of two sequences. For this Assessed Practical you are asked to:

1) Develop a dynamic programming algorithm for performing a ***semi-global*** alignment of two sequences.

You should closely follow the program introduced in Practical 16 to create a program (R function) that reads in two nucleotide sequences (e.g. CGGATG and CAGTG) as well as "match", "mismatch" and "gap" scores. The program should then build a dynamic programming matrix, fill it out according to the sequences and scoring scheme, perform the trace back to find the optimal semi-global alignment, and write out this optimal alignment.

Write the program so that it does NOT show terminal-gap regions, e.g. if aligning ACGGTAACGAAGCCCG with TAAGAGGC your alignment might look like

```
TAACGAAGC
TAA-GAGGC
```

but not

```
ACGGTAACGAAGCCCG
----TAA-GAGGC---
```

Here are some example alignments for checking your program:

```
DP_nt(1,-1,-2,"CTTCACACTCAAAGGCGGTGCACCAACAAAGGTTACTTTTGGTGATGACACTGTGATAGA",
"CCTACTTTGTTCAGACTCAAAGGTGGTGCGCCTCCCAAAGGAGTTAAGTTTGGTGGCGAA")
[1] "CTTCACACTCAAAGGCGGTGCACC-AACAAA-G-GTTACTTTTGGTGATGAC"
[1] "GTTCAGACTCAAAGGTGGTGCGCCTCCCAAAGGAGTTAAGTTTGGTGGCGAA"

DP_nt(1,-1,-2,"CTTCACACTCAAAGGCGGTGCACCAACAAAGGTTACTTTTGGTGATGACACTGTGATAGA",
"AGGTGCCCCTACAAAGGGAGTCACATTTGGTGAAGACACAGTTGTGGAAATCCAGGGTTA")
[1] "CGGTGCACCAACAAA--G-GTTACTTTTGGTGATGACACTGTGATAGA"
[1] "AGGTGCCCCTACAAAGGGAGTCACATTTGGTGAAGACACAGTTGTGGA"

DP_nt(1,-1,-1,"CTTCACACTCAAAGGCGGTGCACCAACAAAGGTTACTTTTGGTGATGACACTGTGATAGA",
"AGGTGCCCCTACAAAGGGAGTCACATTTGGTGAAGACACAGTTGTGGAAATCCAGGGTTA")
[1] "CGGTGCACCAACAAA--G-GTTACTTTTGGTGATGACACTG-TGATAGA"
[1] "AGGTGCCCCTACAAAGGGAGTCACATTTGGTGAAGACACAGTTG-TGGA"

DP_nt(1,-2,-1,"CTTCACACTCAAAGGCGGTGCACCAACAAAGGTTACTTTTGGTGATGACACTGTGATAGA",
"AGGTGCCCCTACAAAGGGAGTCACATTTGGTGAAGACACAGTTGTGGAAATCCAGGGTTA")
[1] "-GGTGCACCAACAAA--G-GTTACTTTTGGTGATGACAC---TGT-GATAGA"
[1] "AGGTGCCCCTACAAAGGGAGTCACATTTGGTGAAGACACAGTTGTGGA-A-A"
```

Note that, depending on exactly how you wrote your code, if there are multiple equally optimal solutions you might potentially get slightly different answers from the above.

2) Use your program to perform ***semi-global*** alignments of the following sequence pairs:

   i) AGTGTT and CAATG (scoring scheme: match +1, mismatch -1, gap -2) [this is the example in the lectures]

   ii) GGTAATG and CTAGTGTT (scoring scheme: match +1, mismatch -1, gap -2)

   iii) GGGGGGCTCCAAGCCCAGAACACCAAGGGGCCCAAAAA and CTCCGACCCAGCACCACGTGGC (scoring scheme: match +1, mismatch -1, gap -2)

   iv) ATGGAGAGCCTTGTCCCTGGTTTCAACGAGAAAACACACGTCCTC and GTCCCCGGGTTTAATGAGAGGACTCATGTCCTCCTCAGTTTGCCTGTT (scoring scheme: match +1, mismatch -1, gap -2)

   v) ATGGAGAGCCTTGTCCCTGGTTTCAACGAGAAAACACACGTCCTC and GTCCCCGGGTTTAATGAGAGGACTCATGTCCTCCTCAGTTTGCCTGTT (scoring scheme: match +1, mismatch -1, gap -1)

Assessment

Your report should be submitted electronically via Moodle, in two different formats - both as an R markdown document (in .Rmd format) and formatted as a .html or .pdf document - **before 17.00 on 15 May 2023**. If you are unable to produce the .html or .pdf version, it is OK to submit only the .Rmd version.

Please make sure to include your name and CRsid in the body of your report; you should also include your CRSID in the name of the files, which should be titled **'MathBiolAssessedPractical3_XXXX' (replacing the XXXX with your CRSID)**."

Your assignment will be marked by (a) loading the R markdown into a clean environment and checking it produces the expected output, and (b) your answers to question 2 (i.e. the example sequence alignments). You might therefore wish to check that your code works in an entirely self-contained manner and does not rely on details of the environment that were altered during development (e.g. by shutting R, starting R, loading your .Rmd file, and knitting it).

Notes

- You will remember from Practical 16 that, for simplicity, we decided to just output one alignment if there were multiple possibilities, and we arbitrarily gave traceback_d priority over traceback_h, and traceback_h priority over traceback_v. You are advised to follow the same scheme for the Assessed Practical.
- If there are multiple equally optimal solutions for a given pair of input sequences (e.g. multiple equally high scores in the last column and last row), your program only needs to find one of the solutions.
- This assignment does not require any external data files, so there is no need to specify a working directory (in other words, don't use the setwd function).
- The code you have to write is very similar to that from Practical 16 (which is available on Moodle); feel free to reuse the code as needed. Keep in mind that there may be different ways to write code to get correct results.
