

CSCI 291T Spring 2015 Programming Assignment 2 25 pts. Due: 2/10 (Tue)

Implement the Smith-Waterman local alignment algorithm with the linear gap scoring scheme. Since the dynamic programming needs a large size table, it is impractical to implement such a $O(nm)$ space algorithm. For this programming assignment, implement the space optimized version, i.e., using only two rows or two columns for the score table.

With this space optimized version, we cannot traceback the alignment, but can obtain the optimum score and the ending positions of the optimum local alignment.

In this practice, you should write a driver, which reads a database file (chr1.fa) and a query file, which includes multiple fasta format sequences, and calls Smith-Waterman function for each query sequence.

Linear gap scoring scheme is described in the textbook, i.e., score (-1) for each indel (insertion or deletion). As used in the textbook, use score (+2) for match and score (-1) for mismatch.

Program input: chr1.fa and a query file containing gene NM_032291's first 10 exon sequences; the exon sequences file is placed in the Blackboard by the instructor.

Program output: The optimum local alignment score for each exon sequence against chr1.fa, together with the corresponding ending positions of the query and database sequences.

In this practice, only nucleotide (DNA) data are used, and a sample output is shown below.

Submission:

1. Include good documentations in the source code and submit the hard copy of the source code. Documentation should include a global documentation (at least, program description – what it does, input/output description, methods/algorithms used, how to compile and run) and each function head documentation (what the function does, input/output, etc.).
2. Run your program with given input data and submit the hard copy of the run time output.

Sample output for nucleotide data

```
>chr1.66999824.67000051.NM_032291_exon_0_0_chr1_66999825_f.+      len=227
ref(1-10) = NNNNNNNNNN ; pattern(1-10) = TTTCTCTCAG
--- Optimum Smith-Waterman score = 454 (i=227, j=67000051)
>chr1.67091529.67091593.NM_032291_exon_1_0_chr1_67091530_f.+      len=64
ref(1-10) = NNNNNNNNNN ; pattern(1-10) = GATTGAAAAA
--- Optimum Smith-Waterman score = 128 (i=64, j=67091593)
>chr1.67098752.67098777.NM_032291_exon_2_0_chr1_67098753_f.+      len=25
ref(1-10) = NNNNNNNNNN ; pattern(1-10) = AGG TTCACCA
--- Optimum Smith-Waterman score = 50 (i=25, j=67098777)
>chr1.67101626.67101698.NM_032291_exon_3_0_chr1_67101627_f.+      len=72
ref(1-10) = NNNNNNNNNN ; pattern(1-10) = CAGCCCAGCC
--- Optimum Smith-Waterman score = 144 (i=72, j=67101698)
>chr1.67105459.67105516.NM_032291_exon_4_0_chr1_67105460_f.+      len=57
ref(1-10) = NNNNNNNNNN ; pattern(1-10) = AAGAAAAGCA
--- Optimum Smith-Waterman score = 114 (i=57, j=67105516)
```