

Computer Architecture

HW2

2019 FALL

PROF. Yi-Chang Lu

Smith-Waterman Algorithm

- A dynamic programming algorithm for sequence alignment
- Widely used in bioinformatics (生物資訊學)

$$H(i, j) = \max \left\{ \begin{array}{ll} 0 & \\ H(i-1, j-1) + s(a_i, b_j) & \text{Match/Mismatch} \\ H(i-1, j) + W & \text{Deletion} \\ H(i, j-1) + W & \text{Insertion} \end{array} \right\}, 1 \leq i \leq m, 1 \leq j \leq n$$

An Example (1/2)

- Sequence 1 = ACACACTA
- Sequence 2 = AGCACACA
- match: +2, mismatch/gap: -1

$$H = \begin{pmatrix} - & - & A & C & A & C & A & C & T & A \\ - & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ A & 0 & 2 & 1 & 2 & 1 & 2 & 1 & 0 & 2 \\ G & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 \\ C & 0 & 0 & 3 & 2 & 3 & 2 & 3 & 2 & 1 \\ A & 0 & 2 & 2 & 5 & 4 & 5 & 4 & 3 & 4 \\ C & 0 & 1 & 4 & 4 & 7 & 6 & 7 & 6 & 5 \\ A & 0 & 2 & 3 & 6 & 6 & 9 & 8 & 7 & 8 \\ C & 0 & 1 & 4 & 5 & 8 & 8 & 11 & 10 & 9 \\ A & 0 & 2 & 3 & 6 & 7 & 10 & 10 & 10 & 12 \end{pmatrix}$$

$$T = \begin{pmatrix} - & - & A & C & A & C & A & C & T & A \\ - & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ A & 0 & \nearrow & \leftarrow & \nearrow & \leftarrow & \nearrow & \leftarrow & \leftarrow & \nearrow \\ G & 0 & \uparrow & \nearrow & \uparrow & \nearrow & \uparrow & \nearrow & \nearrow & \uparrow \\ C & 0 & \uparrow & \nearrow & \leftarrow & \nearrow & \leftarrow & \nearrow & \leftarrow & \leftarrow \\ A & 0 & \nearrow & \uparrow & \nearrow & \leftarrow & \nearrow & \leftarrow & \leftarrow & \nearrow \\ C & 0 & \uparrow & \nearrow & \uparrow & \nearrow & \leftarrow & \nearrow & \leftarrow & \leftarrow \\ A & 0 & \nearrow & \uparrow & \nearrow & \uparrow & \nearrow & \leftarrow & \leftarrow & \nearrow \\ C & 0 & \uparrow & \nearrow & \uparrow & \nearrow & \uparrow & \nearrow & \leftarrow & \leftarrow \\ A & 0 & \nearrow & \uparrow & \nearrow & \uparrow & \nearrow & \uparrow & \nearrow & \nearrow \end{pmatrix}$$

An Example (2/2)

- Alignment result:
 - Score = 12
 - Sequence 1 => A-CACACTA
 - Sequence 2 => AGCACAC-A

$$H = \begin{pmatrix} - & A & C & A & C & A & C & T & A \\ - & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ A & 0 & 2 & 1 & 2 & 1 & 2 & 1 & 0 & 2 \\ G & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 \\ C & 0 & 0 & 3 & 2 & 3 & 2 & 3 & 2 & 1 \\ A & 0 & 2 & 2 & 5 & 4 & 5 & 4 & 3 & 4 \\ C & 0 & 1 & 4 & 4 & 7 & 6 & 7 & 6 & 5 \\ A & 0 & 2 & 3 & 6 & 6 & 9 & 8 & 7 & 8 \\ C & 0 & 1 & 4 & 5 & 8 & 8 & 11 & 10 & 9 \\ A & 0 & 2 & 3 & 6 & 7 & 10 & 10 & 10 & 12 \end{pmatrix}$$

$$T = \begin{pmatrix} - & A & C & A & C & A & C & T & A \\ - & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ A & 0 & \swarrow & \leftarrow & \swarrow & \leftarrow & \swarrow & \leftarrow & \leftarrow & \swarrow \\ G & 0 & \uparrow & \swarrow & \uparrow & \swarrow & \uparrow & \swarrow & \swarrow & \uparrow \\ C & 0 & \uparrow & \swarrow & \leftarrow & \swarrow & \leftarrow & \swarrow & \leftarrow & \leftarrow \\ A & 0 & \swarrow & \uparrow & \swarrow & \leftarrow & \swarrow & \leftarrow & \leftarrow & \swarrow \\ C & 0 & \uparrow & \swarrow & \uparrow & \swarrow & \leftarrow & \swarrow & \leftarrow & \leftarrow \\ A & 0 & \swarrow & \uparrow & \swarrow & \uparrow & \swarrow & \leftarrow & \leftarrow & \swarrow \\ C & 0 & \uparrow & \swarrow & \uparrow & \swarrow & \uparrow & \swarrow & \leftarrow & \leftarrow \\ A & 0 & \swarrow & \uparrow & \swarrow & \uparrow & \swarrow & \uparrow & \swarrow & \swarrow \end{pmatrix}$$

Homework

- Implement Smith-Waterman algorithm on Qtspim with following requirements
 - Lengths of two input sequences can be changed or unequal
 - Scores: match: +3 , mismatch: -1 , gap: -2
 - Traceback: Starting at the highest score in the scoring matrix and ending at a matrix cell that has a score of 0

Homework

- Output should include:
 - The highest score in the scoring matrix
 - The traceback direction from the highest score to the 0 score
 - (direction: ↖ (3), ↑ (2), ← (1))
 - Recommended priority: 3 > 2 > 1

- Example:

The highest score is: 12

Traceback result:
313333323

seq 1

seq 2

$$T = \begin{pmatrix} - & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ A & 0 & \swarrow & \leftarrow & \swarrow & \leftarrow & \swarrow & \leftarrow & \leftarrow & \swarrow \\ G & 0 & \uparrow & \swarrow & \uparrow & \swarrow & \uparrow & \swarrow & \swarrow & \uparrow \\ C & 0 & \uparrow & \swarrow & \leftarrow & \swarrow & \leftarrow & \swarrow & \leftarrow & \leftarrow \\ A & 0 & \swarrow & \uparrow & \swarrow & \leftarrow & \swarrow & \leftarrow & \leftarrow & \swarrow \\ C & 0 & \uparrow & \swarrow & \uparrow & \swarrow & \leftarrow & \swarrow & \leftarrow & \leftarrow \\ A & 0 & \swarrow & \uparrow & \swarrow & \uparrow & \swarrow & \leftarrow & \leftarrow & \swarrow \\ C & 0 & \uparrow & \swarrow & \uparrow & \swarrow & \uparrow & \swarrow & \leftarrow & \leftarrow \\ A & 0 & \swarrow & \uparrow & \swarrow & \uparrow & \swarrow & \uparrow & \swarrow & \swarrow \end{pmatrix}$$

Homework

- 80% : correctness of the program
 - 60% scoring matrix (highest score), 20% traceback
 - 2 testing data provided with answer, you should screenshot your result of both 2 and paste them on your report
 - Other sequences will be tested, you should pass all testing data to get full credit
- 20%: report which includes
 - screenshots of your result
 - how you design this program
 - how to run your program

1. load byte
2. print function

Homework

- Due: 2019/10/29 13:00
- submitted in a compressed file **CA1081_HW2_yourID.zip** to CEIBA
- The file includes:
 - HW2_yourID/
 - HW2.s
 - Report.pdf
- If you have any question, please contact r08943010@ntu.edu.tw

References

- Smith, Temple F. & Waterman, Michael S. (1981). "Identification of Common Molecular Subsequences". Journal of Molecular Biology. 147: 195–197.
- https://cs.stanford.edu/people/eroberts/courses/soco/projects/computers-and-the-hgp/smith_waterman.html