

Smith-Waterman Algorithm

COMP5211 Assignment Tutorial

Smith-Waterman Algorithm

- Measure the similarity between two strings (e.g., DNA sequence)
- Example: AAGTTAC and AACTTGAC

AAGTTAC
| | | | |
AACTTGAC

- Assume that Match: +3, Mismatch: -3, Gap: -2
- Similarity score: $6 \times 3 - 3 - 2 = 13$

Algorithm Description

- Input: string $A = a_1, a_2, \dots, a_n, B = b_1, b_2, \dots, b_m$
- match score u , mismatch score v , gap penalty w
- Compute scoring matrix H by dynamic programming:

$$H_{ij} = \max \begin{cases} H_{i-1,j-1} + s(a_i, b_j) \\ H_{i-1,j} - w \\ H_{i,j-1} - w \\ 0 \end{cases} \quad (1 \leq i \leq n, 1 \leq j \leq m)$$
$$s(a_i, b_j) = \begin{cases} u, & a_i = b_j \\ v, & a_i \neq b_j \end{cases}$$

Illustration

```
for  $i \leftarrow 1$  to  $|A|$  do
  for  $j \leftarrow 1$  to  $|B|$  do
     $score[i][j] \leftarrow \max(0,$   

                                 $score[i - 1][j] - w,$   

                                 $score[i][j - 1] - w,$   

                                 $score[i - 1][j - 1] + sub\_mat(a_i, b_j));$   

     $max\_score \leftarrow \max(max\_score, score[i][j]);$ 
  end
end
```

| | | G | T | C | A | |
|---|---|---|---|---|---|---|
| | | 0 | 0 | 0 | 0 | 0 |
| G | 0 | 3 | 1 | 0 | 0 | |
| T | 0 | 1 | 6 | 4 | 2 | |
| A | 0 | 0 | 4 | 3 | 7 | |
| A | 0 | 0 | 2 | 1 | 6 | |

Scoring matrix H

Parallelization

- Hint: cells on the same anti-diagonal can be computed in parallel

