Smith-Waterman Algorithm

COMP5211 Assignment Tutorial

Smith-Waterman Algorithm

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

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

Algorithm Description

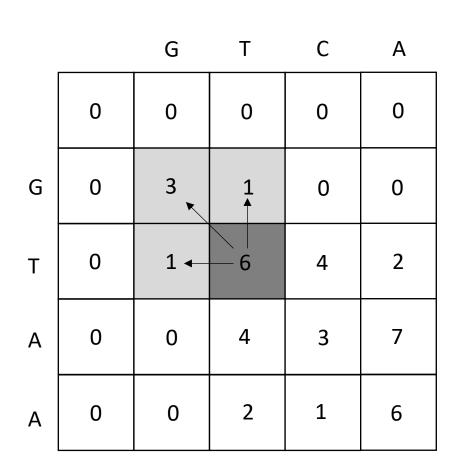
- Input: string $A=a_1,a_2,\ldots,a_n$, $B=b_1,b_2,\ldots,b_m$
- match score u, mismatch score v, gap penalty w
- Compute scoring matrix *H* by dynamic programming:

$$H_{ij} = \max \begin{cases} H_{i-1,j-1} + s(a_i, b_j) \\ H_{i-1,j} - w \\ H_{i,j-1} - w \\ 0 \end{cases}$$
 $(1 \le i \le n, 1 \le j \le m)$

$$s(a_i, b_j) = \begin{cases} u, & a_i = b_j \\ v, & a_i \ne b_j \end{cases}$$

Illustration

```
\begin{array}{c|c} \textbf{for } i \leftarrow 1 \textbf{ to } |A| \textbf{ do} \\ & \textbf{ for } j \leftarrow 1 \textbf{ to } |B| \textbf{ 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]); \\ & \textbf{ end} \\ & \textbf{end} \end{array}
```



Scoring matrix H

Parallelization

 Hint: cells on the same antidiagonal can be computed in parallel

