# Longest Common Subsequence (LCS)

TANJINA HELALY

# Subsequence

- A **subsequence** is a sequence that can be derived from another sequence by deleting some or no elements without changing the order of the remaining elements.
- For example, the sequence {A, B, D} is a subsequence of {A, B, C, D, E, F} obtained after removal of elements C, E, and F.

# Subsequences vs. Substring

- Subsequences can contain consecutive elements which were not consecutive in the original sequence.
- Substring contains consecutive elements which were also consecutive in the original sequence.
- Example:
  - “gramm” is both subsequence and substring of “programming”
  - “gammg” is a subsequence of “programming” but not substring.
  - All substrings are subsequences but all subsequences are not substrings.

# Common subsequence

- Given two sequences *X* and *Y*, a sequence *Z* is said to be a *common subsequence* of *X* and *Y*, if *Z* is a subsequence of both *X* and *Y*.
- For example, if
  - X = a c b d e g c e d b g
  - Y = c b e g j c f e k b
  - Z = b e b
  - then Z is the common subsequence of X and Y.
- Longest Common Subsequence will be **c b e g c e b**
  - It measured how Similar 2 strings are.

# Application

- To compare the DNA of two (or more) different organisms.
  - One reason to compare two strands of DNA is to determine how “similar” the two strands are, as some measure of how closely related the two organisms are.

# A recursive solution

- Assume 2 sequences $X=\{x_1, x_2,\ldots x_m\}$ and $Y = =\{y_1, y_2,\ldots y_n\}$
- If $x_m == y_n$
  - then find an LCS of $X_{m-1}$ and $Y_{n-1}$.
  - Appending $x_m$ ( or $y_n$) to this LCS yields an LCS of X and Y
- If $x_m != y_n$
  - then find the LCS(m-1, n) of $X_{m-1}$ & $Y_n$ and LCS(m, n-1) of $X_m$ & $Y_n$
  - LCS of X and Y will be the max of LCS(m-1,n) and LCS (m, n-1)

# Memoized Version

- To find the LCM of 2 sequences $X=\{x_1, x_2,\ldots x_m\}$ and Y = $=\{y_1, y_2,\ldots y_n\}$ follow the steps below
  - Create a Matrix C of m+1 by n+1 size.
  - C[i,j] represent the length of LCS of $X_i$ and $Y_j$
  - If either i = 0 or j = 0, one of the sequences has length 0
  - Hence, set C[i,0] and C[0,j] to 0.
  - Now traverse each cell row-wise or column-wise
    - If $x_i = y_j$ , set C[i, j] = C[i,j]+1 i.e. set the value of that cell 1 more than the value of the upper left diagonal cell.
    - If $x_i$ != $y_j$ , the C[i, j] will be either the value of the cell above it or left of it whichever is larger.

$$c[i,j] = \begin{cases} 0 & \text{if } i = 0 \text{ or } j = 0 , \\ c[i-1,j-1]+1 & \text{if } i,j > 0 \text{ and } x_i = y_j , \\ \max(c[i,j-1], c[i-1,j]) & \text{if } i,j > 0 \text{ and } x_i \neq y_j . \end{cases}$$

# Illustration

- set C[i,0] and C[0,j] to 0.

| | y[j] | b | d | c | a | b | a |
|---|---|---|---|---|---|---|---|
| x[i] | | | | | | | |
| a | | | | | | | |
| b | | | | | | | |
| c | | | | | | | |
| b | | | | | | | |
| d | | | | | | | |
| a | | | | | | | |
| b | | | | | | | |

# Illustration

- set C[i,0] and C[0,j] to 0.

| | y[j] | b | d | c | a | b | a |
|---|---|---|---|---|---|---|---|
| x[i] | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| a | 0 | | | | | | |
| b | 0 | | | | | | |
| c | 0 | | | | | | |
| b | 0 | | | | | | |
| d | 0 | | | | | | |
| a | 0 | | | | | | |
| b | 0 | | | | | | |

# Illustration

- $x_1$ (a) != $y_1$ (b). So, take the max of the cell above it and left of it. As both of those are 0, C[1,1] will be 0.

| | y[j] | b | d | c | a | b | a |
|---|---|---|---|---|---|---|---|
| x[i] | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| a | 0 | 0 | | | | | |
| b | 0 | | | | | | |
| c | 0 | | | | | | |
| b | 0 | | | | | | |
| d | 0 | | | | | | |
| a | 0 | | | | | | |
| b | 0 | | | | | | |

# Illustration

- Similarly C[1,2] and C[1,3] will be 0.

| | y[j] | b | d | c | a | b | a |
|---|---|---|---|---|---|---|---|
| x[i] | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| a | 0 | 0 | 0 | 0 | | | |
| b | 0 | | | | | | |
| c | 0 | | | | | | |
| b | 0 | | | | | | |
| d | 0 | | | | | | |
| a | 0 | | | | | | |
| b | 0 | | | | | | |

# Illustration

- As $x_1 = y_4 = a$, C[1,4] will be 1 larger than the diagonal cell (Yellow highligted one.

| | y[j] | b | d | c | a | b | a |
|---|---|---|---|---|---|---|---|
| x[i] | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| a | 0 | 0 | 0 | 0 | 1 | | |
| b | 0 | | | | | | |
| c | 0 | | | | | | |
| b | 0 | | | | | | |
| d | 0 | | | | | | |
| a | 0 | | | | | | |
| b | 0 | | | | | | |

# Illustration

- $x_1$ (a) != $y_5$ (b). So, take the max of the cell above it and left of it (both highlighted yellow). As Left cell has the bigger value (1), C[1,5] will be 1.

| | y[j] | b | d | c | a | b | a |
|---|---|---|---|---|---|---|---|
| x[i] | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| a | 0 | 0 | 0 | 0 | 1 | | |
| b | 0 | | | | | | |
| c | 0 | | | | | | |
| b | 0 | | | | | | |
| d | 0 | | | | | | |
| a | 0 | | | | | | |
| b | 0 | | | | | | |

# Illustration

- $x_1$ (a) != $y_5$ (b). So, take the max of the cell above it and left of it (both highlighted yellow). As Left cell has the bigger value (1), C[1,5] will be 1.

| | y[j] | b | d | c | a | b | a |
|---|---|---|---|---|---|---|---|
| x[i] | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| a | 0 | 0 | 0 | 0 | 1 | 1 | |
| b | 0 | | | | | | |
| c | 0 | | | | | | |
| b | 0 | | | | | | |
| d | 0 | | | | | | |
| a | 0 | | | | | | |
| b | 0 | | | | | | |

# Illustration

- Similarly fill up the rest of the table.

| | y[j] | b | d | c | a | b | a |
|---|---|---|---|---|---|---|---|
| x[i] | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| a | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| b | 0 | 1 | 1 | 1 | 1 | 2 | 2 |
| c | 0 | 1 | 1 | 2 | 2 | 2 | 2 |
| b | 0 | 1 | 1 | 2 | 2 | 3 | 3 |
| d | 0 | 1 | 2 | 2 | 2 | 3 | 3 |
| a | 0 | 1 | 2 | 2 | 3 | 3 | 4 |
| b | 0 | 1 | 2 | 2 | 3 | 4 | 4 |

# Illustration

| | y[j] | b | d | c | a | b | a |
|---|---|---|---|---|---|---|---|
| x[i] | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| a | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| b | 0 | 1 | 1 | 1 | 1 | 2 | 2 |
| c | 0 | 1 | 1 | 2 | 2 | 2 | 2 |
| b | 0 | 1 | 1 | 2 | 2 | 3 | 3 |
| d | 0 | 1 | 2 | 2 | 2 | 3 | 3 |
| a | 0 | 1 | 2 | 2 | 3 | 3 | 4 |
| b | 0 | 1 | 2 | 2 | 3 | 4 | 4 |

# Illustration

| | y[j] | b | d | c | a | b | a |
|---|---|---|---|---|---|---|---|
| x[i] | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| a | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| b | 0 | 1 | 1 | 1 | 1 | 2 | 2 |
| c | 0 | 1 | 1 | 2 | 2 | 2 | 2 |
| b | 0 | 1 | 1 | 2 | 2 | 3 | 3 |
| d | 0 | 1 | 2 | 2 | 2 | 3 | 3 |
| a | 0 | 1 | 2 | 2 | 3 | 3 | 4 |
| b | 0 | 1 | 2 | 2 | 3 | 4 | 4 |

LCS= bcba

# Illustration

| | y[j] | b | D | c | a | b | a |
|---|---|---|---|---|---|---|---|
| x[i] | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| a | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| b | 0 | 1 | 1 | 1 | 1 | 2 | 2 |
| c | 0 | 1 | 1 | 2 | 2 | 2 | 2 |
| b | 0 | 1 | 1 | 2 | 2 | 3 | 3 |
| d | 0 | 1 | 2 | 2 | 2 | 3 | 3 |
| a | 0 | 1 | 2 | 2 | 3 | 3 | 4 |
| b | 0 | 1 | 2 | 2 | 3 | 4 | 4 |

LCS= bcab

## ALGORITHM

```
LCS-LENGTH(X, Y)
m = X.length
n = Y.length
let b[1..m, 1..n] and c[0..m, 0..n] be new tables
for i = 1 to m
    c[i, 0] = 0
for j = 0 to n
    c[0, j] = 0
for i = 1 to m
    for j = 1 to n
        if x_i == y_j
            c[i, j] = c[i − 1, j − 1] + 1
            b[i, j] = "↖"
        elseif c[i − 1, j] ≥ c[i, j − 1]
            c[i, j] = c[i − 1, j]
            b[i, j] = "↑"
        else c[i, j] = c[i, j − 1]
            b[i, j] = "←"
return c and b
```

## SIMULATION

| | $j$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|---|
| $i$ | | $y_j$ | $B$ | $D$ | $C$ | $A$ | $B$ | $A$ |
| 0 | $x_i$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | $A$ | 0 | ↑0 | ↑0 | ↑0 | ↖1 | ←1 | ↖1 |
| 2 | $B$ | 0 | ↖1 | ←1 | ←1 | ↑1 | ↖2 | ←2 |
| 3 | $C$ | 0 | ↑1 | ↑1 | ↖2 | ←2 | ↑2 | ↑2 |
| 4 | $B$ | 0 | ↖1 | ↑1 | ↑2 | ↑2 | ↖3 | ←3 |
| 5 | $D$ | 0 | ↑1 | ↖2 | ↑2 | ↑2 | ↑3 | ↑3 |
| 6 | $A$ | 0 | ↑1 | ↑2 | ↑2 | ↖3 | ↑3 | ↖4 |
| 7 | $B$ | 0 | ↖1 | ↑2 | ↑2 | ↑3 | ↖4 | ↑4 |