FACULTY OF SCIENCE,
TECHNOLOGY
AND MEDICINE

Longest Common Subsequences

Seminar 2

Joris LIMONIER

Supervised by George KERCHEV

May 31, 2021

Table of Contents

1. Introduction

1. Introduction

1.1 What are LCS ?

What are LCS ?

Notation

"LCS" = Longest Common Subsequence(s)

What are LCS ?

Notation

"LCS" = Longest Common Subsequence(s)

Example 1

S_1 : A B A B B

What are LCS ?

Notation

"LCS" = Longest Common Subsequence(s)

Example 1

$$S_1 : \quad A \quad B \quad A \quad B \quad B$$
$$S_2 : \quad A \quad A \quad B \quad A \quad B$$

What are LCS ?

Notation

"LCS" = Longest Common Subsequence(s)

Example 1

S_1 : **A** B A B B

S_2 : **A** A B A B

What are LCS ?

Notation

"LCS" = Longest Common Subsequence(s)

Example 1

S_1 : **A** **B** A B B

S_2 : **A** A **B** A B

What are LCS ?

Notation

"LCS" = Longest Common Subsequence(s)

Example 1

S_1 : **A** **B** **A** B B

S_2 : **A** A **B** **A** B

What are LCS ?

Notation

"LCS" = Longest Common Subsequence(s)

Example 1

S_1 : **A** **B** **A** **B** B

S_2 : **A** A **B** **A** **B**

What are LCS ?

Notation

"LCS" = Longest Common Subsequence(s)

Example 1

S_1 : **A** **B** **A** **B** B

S_2 : **A** A **B** **A** **B**

$\Longrightarrow$ The LCS between S_1 and S_2 is **A B A B**

What are LCS ?

Notation

"LCS" = Longest Common Subsequence(s)

Example 1

$$S_1 : \quad \mathbf{A}\ \mathbf{B}\ \mathbf{A}\ \mathbf{B}\ \mathrm{B}$$
$$S_2 : \quad \mathbf{A}\ \mathrm{A}\ \mathbf{B}\ \mathbf{A}\ \mathbf{B}$$

$\Longrightarrow$ The LCS between S_1 and S_2 is **A B A B**

NB: LCS may not be unique, A A B B also works.

Example 2

What is the LCS of the following sequences ?

Example 2

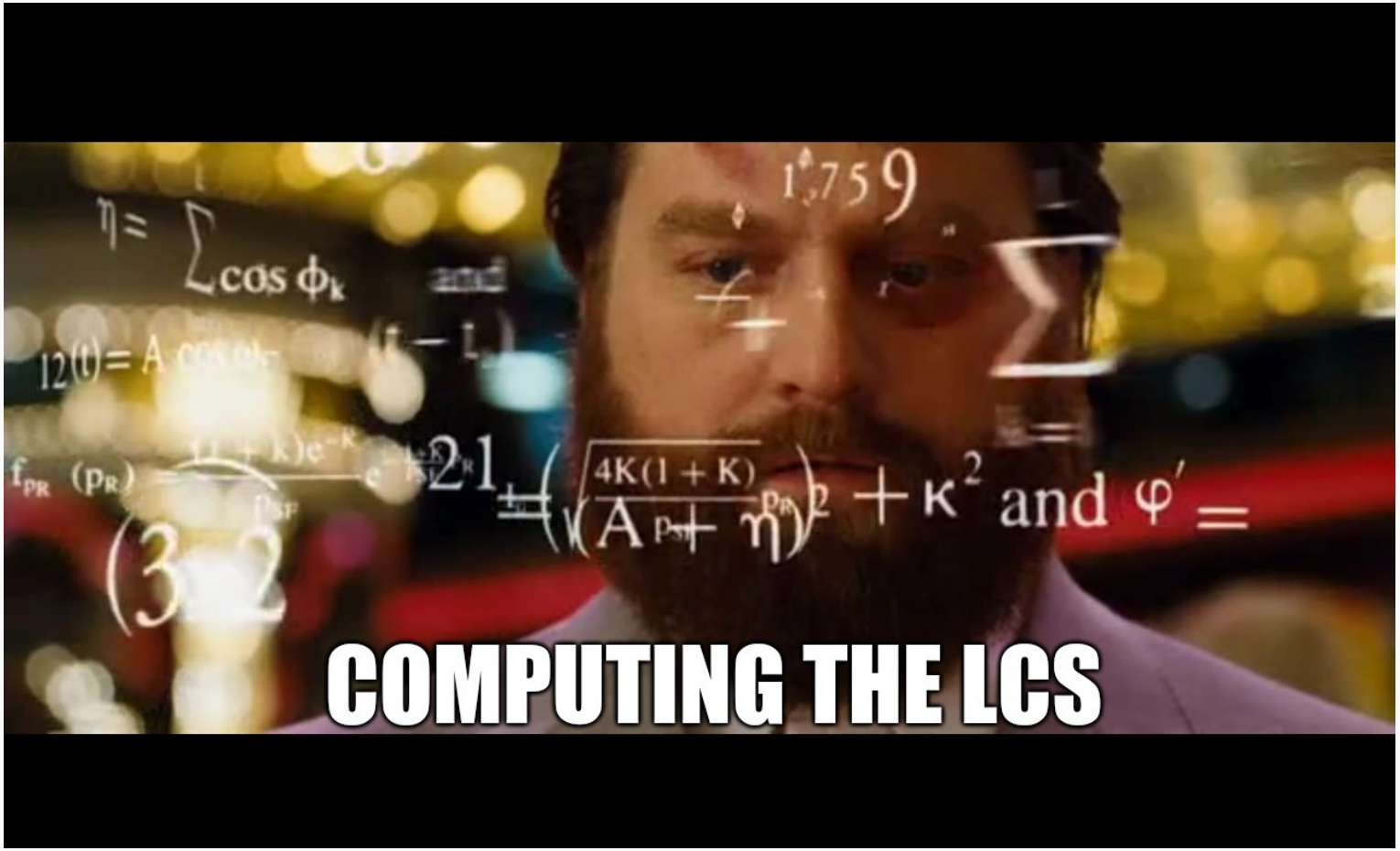
What is the LCS of the following sequences ?

S_3 : AABBAABAAABABAAABAAABBABAAABAAABAAA
AAAABBABBBBAAABABBAABBAABBBBBAAAABA
BBABAAAABABAABBBABABBBBBAAABBBAABBB
AABAABBABABAABABBBBBBAABBBBBBBAAAAAB
AABAAAAABAABAABAAABBABBBBABBAAAABBB

S_4 : BABBBABAABAABBBABBABBBBBBBABABAAABB
BBABBABBABBBABBBBABBABBABABABBAABABA
BAABABAAAABABBABABBAAAABABBAABABABB
BABBBBBBAABAAABBABBBBAAAAABBBBBAAAB
ABBAAAABBBABABAABBABBBAABABBBABAABA

Example 2

What is the LCS of the following sequences ?



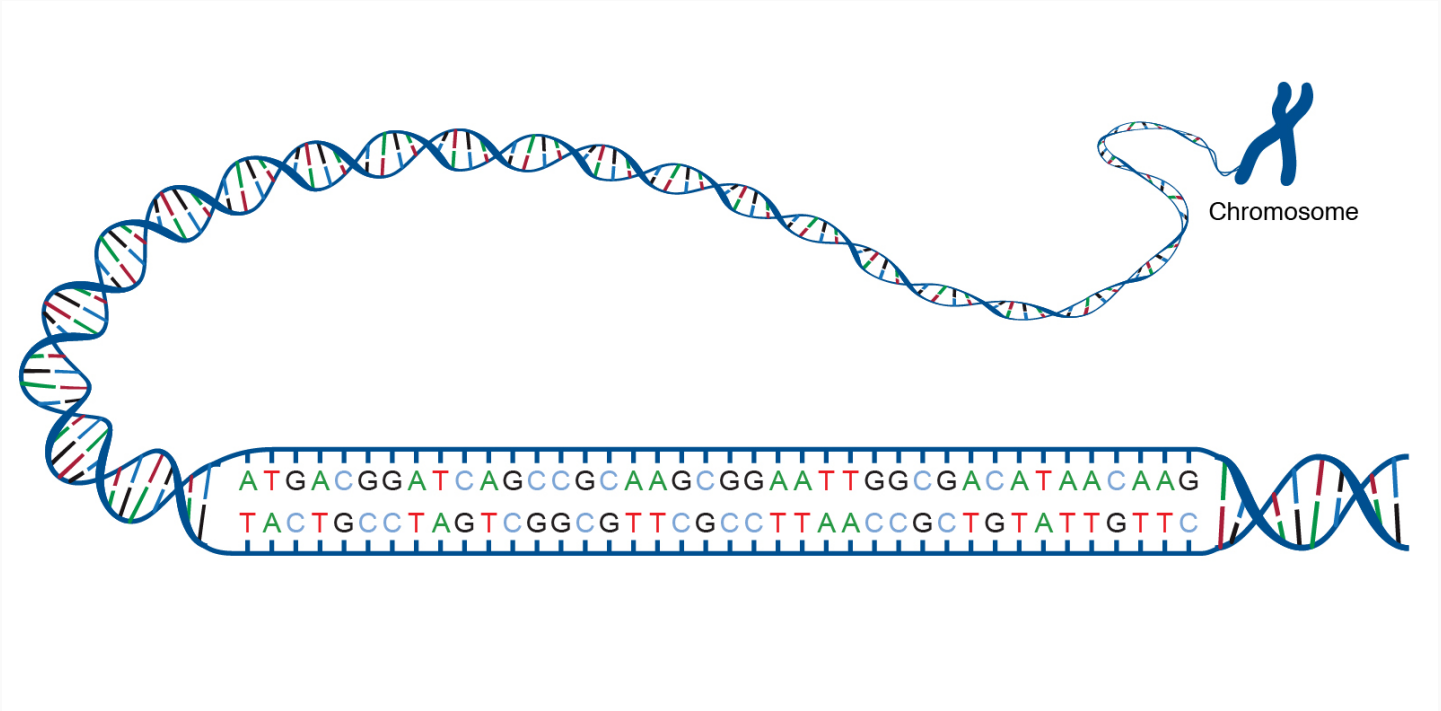
1. Introduction

1.2 Why are we interested in LCS ?

Applications

Applications

- Bioinformatics: Compare sequences of nucleotides (DNA)

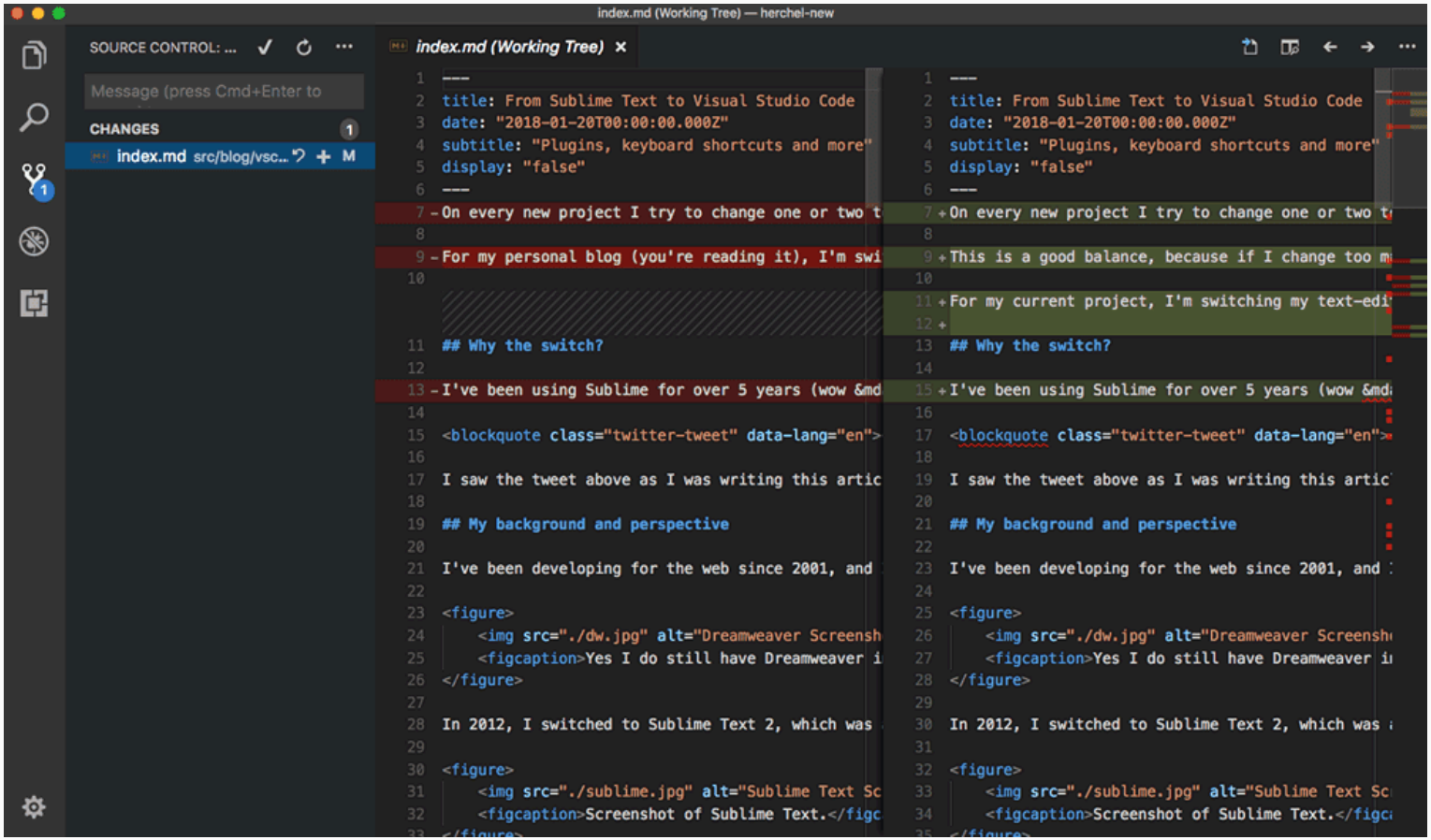


Applications

- Bioinformatics: Compare sequences of nucleotides (DNA)
- Natural Language Processing: Compare texts

Applications

- Bioinformatics: Compare sequences of nucleotides (DNA)
- Natural Language Processing: Compare texts
- Computer Science: Detect differences in texts



2. How to find LCS ?

2. How to find LCS ?

2.1 Step A: Building the table

Set-up

Let $S_1 = ABABB$ and $S_2 = AABAB$.

Set-up

Let $S_1 = ABABB$ and $S_2 = AABAB$.

- Make a table where S_1 and S_2 are the column and row names respectively.

	A	B	A	B	B
A					
A					
B					
A					
B					

Set-up

Let $S_1 = ABABB$ and $S_2 = AABAB$.

- Make a table where S_1 and S_2 are the column and row names respectively.
- Add a row (resp. column) at the top (resp. left) of the table. Fill them with 0's.

	$\varnothing$	A	B	A	B	B
$\varnothing$	0	0	0	0	0	0
A	0					
A	0					
B	0					
A	0					
B	0					

Procedure

Start from top-left corner. Move left to right, line by line.

- If row and column names match, increment adjascent top-left-diagonal cell by 1.
- Else take the maximum of top and left cells.

	∅	A	B	A	B	B
∅	0	0	0	0	0	0
A	0					
A	0					
B	0					
A	0					
B	0					

Procedure

Start from top-left corner. Move left to right, line by line.

- If row and column names match, increment adjascent top-left-diagonal cell by 1.
- Else take the maximum of top and left cells.

	∅	A	B	A	B	B
∅	0	0	0	0	0	0
A	0	1				
A	0					
B	0					
A	0					
B	0					

Procedure

Start from top-left corner. Move left to right, line by line.

- If row and column names match, increment adjascent top-left-diagonal cell by 1.
- Else take the maximum of top and left cells.

	∅	A	B	A	B	B
∅	0	0	0	0	0	0
A	0	1	1			
A	0					
B	0					
A	0					
B	0					

Procedure

Start from top-left corner. Move left to right, line by line.

- If row and column names match, increment adjascent top-left-diagonal cell by 1.
- Else take the maximum of top and left cells.

	∅	A	B	A	B	B
∅	0	0	0	0	0	0
A	0	1	1	1		
A	0					
B	0					
A	0					
B	0					

Procedure

Start from top-left corner. Move left to right, line by line.

- If row and column names match, increment adjascent top-left-diagonal cell by 1.
- Else take the maximum of top and left cells.

	∅	A	B	A	B	B
∅	0	0	0	0	0	0
A	0	1	1	1	1	
A	0					
B	0					
A	0					
B	0					

Procedure

Start from top-left corner. Move left to right, line by line.

- If row and column names match, increment adjascent top-left-diagonal cell by 1.
- Else take the maximum of top and left cells.

	∅	A	B	A	B	B
∅	0	0	0	0	0	0
A	0	1	1	1	1	1
A	0					
B	0					
A	0					
B	0					

Procedure

Start from top-left corner. Move left to right, line by line.

- If row and column names match, increment adjascent top-left-diagonal cell by 1.
- Else take the maximum of top and left cells.

	∅	A	B	A	B	B
∅	0	0	0	0	0	0
A	0	1	1	1	1	1
A	0	1				
B	0					
A	0					
B	0					

Procedure

Start from top-left corner. Move left to right, line by line.

- If row and column names match, increment adjascent top-left-diagonal cell by 1.
- Else take the maximum of top and left cells.

	∅	A	B	A	B	B
∅	0	0	0	0	0	0
A	0	1	1	1	1	1
A	0	1	1			
B	0					
A	0					
B	0					

Procedure

Start from top-left corner. Move left to right, line by line.

- If row and column names match, increment adjascent top-left-diagonal cell by 1.
- Else take the maximum of top and left cells.

	∅	A	B	A	B	B
∅	0	0	0	0	0	0
A	0	1	1	1	1	1
A	0	1	1	2		
B	0					
A	0					
B	0					

Procedure

Start from top-left corner. Move left to right, line by line.

- If row and column names match, increment adjascent top-left-diagonal cell by 1.
- Else take the maximum of top and left cells.

	∅	A	B	A	B	B
∅	0	0	0	0	0	0
A	0	1	1	1	1	1
A	0	1	1	2	2	2
B	0					
A	0					
B	0					

Procedure

Start from top-left corner. Move left to right, line by line.

- If row and column names match, increment adjascent top-left-diagonal cell by 1.
- Else take the maximum of top and left cells.

	∅	A	B	A	B	B
∅	0	0	0	0	0	0
A	0	1	1	1	1	1
A	0	1	1	2	2	2
B	0	1	2	2	3	3
A	0					
B	0					

Procedure

Start from top-left corner. Move left to right, line by line.

- If row and column names match, increment adjascent top-left-diagonal cell by 1.
- Else take the maximum of top and left cells.

	∅	A	B	A	B	B
∅	0	0	0	0	0	0
A	0	1	1	1	1	1
A	0	1	1	2	2	2
B	0	1	2	2	3	3
A	0	1	2	3	3	3
B	0					

Procedure

Start from top-left corner. Move left to right, line by line.

- If row and column names match, increment adjascent top-left-diagonal cell by 1.
- Else take the maximum of top and left cells.

	∅	A	B	A	B	B
∅	0	0	0	0	0	0
A	0	1	1	1	1	1
A	0	1	1	2	2	2
B	0	1	2	2	3	3
A	0	1	2	3	3	3
B	0	1	2	3	4	4

$\Longrightarrow$ **The length of the LCS is 4.**

2. How to find LCS ?

2.2 Step B: Crawling back up the table

Procedure

From the table, deduce LCS by starting from the bottom-right cell. Compare cell value with values of top and left cells.

- If cell value $\in$ {top cell value, left cell value}, move to the one with maximum value.
- Else, add character to LCS and move 1 cell diagonally top-left.

	∅	A	B	A	B	B
∅	0	0	0	0	0	0
A	0	1	1	1	1	1
A	0	1	1	2	2	2
B	0	1	2	2	3	3
A	0	1	2	3	3	3
B	0	1	2	3	4	4

Procedure

From the table, deduce LCS by starting from the bottom-right cell. Compare cell value with values of top and left cells.

- If cell value $\in$ {top cell value, left cell value}, move to the one with maximum value.
- Else, add character to LCS and move 1 cell diagonally top-left.

	∅	A	B	A	B	B
∅	0	0	0	0	0	0
A	0	1	1	1	1	1
A	0	1	1	2	2	2
B	0	1	2	2	3	3
A	0	1	2	3	3	3
B	0	1	2	3	4	4

LCS : _ _ _ _

Procedure

From the table, deduce LCS by starting from the bottom-right cell. Compare cell value with values of top and left cells.

- If cell value $\in$ {top cell value, left cell value}, move to the one with maximum value.
- Else, add character to LCS and move 1 cell diagonally top-left.

	∅	A	B	A	B	B
∅	0	0	0	0	0	0
A	0	1	1	1	1	1
A	0	1	1	2	2	2
B	0	1	2	2	3	3
A	0	1	2	3	3	3
B	0	1	2	3	4	4

LCS : _ _ _ _

Procedure

From the table, deduce LCS by starting from the bottom-right cell. Compare cell value with values of top and left cells.

- If cell value $\in$ {top cell value, left cell value}, move to the one with maximum value.
- Else, add character to LCS and move 1 cell diagonally top-left.

	∅	A	B	A	B	B
∅	0	0	0	0	0	0
A	0	1	1	1	1	1
A	0	1	1	2	2	2
B	0	1	2	2	3	3
A	0	1	2	3	3	3
B	0	1	2	3	**4**	4

LCS : _ _ _ **B**

Procedure

From the table, deduce LCS by starting from the bottom-right cell. Compare cell value with values of top and left cells.

- If cell value $\in$ {top cell value, left cell value}, move to the one with maximum value.
- Else, add character to LCS and move 1 cell diagonally top-left.

	∅	A	B	A	B	B
∅	0	0	0	0	0	0
A	0	1	1	1	1	1
A	0	1	1	2	2	2
B	0	1	2	2	3	3
A	0	1	2	**3**	3	3
B	0	1	2	3	**4**	4

LCS : _ _ **A** **B**

Procedure

From the table, deduce LCS by starting from the bottom-right cell. Compare cell value with values of top and left cells.

- If cell value $\in$ {top cell value, left cell value}, move to the one with maximum value.
- Else, add character to LCS and move 1 cell diagonally top-left.

	∅	A	B	A	B	B
∅	0	0	0	0	0	0
A	0	1	1	1	1	1
A	0	1	1	2	2	2
B	0	1	**2**	2	3	3
A	0	1	2	**3**	3	3
B	0	1	2	3	**4**	4

LCS : _ _ **B** **A** **B**

Procedure

From the table, deduce LCS by starting from the bottom-right cell. Compare cell value with values of top and left cells.

- If cell value $\in$ {top cell value, left cell value}, move to the one with maximum value.
- Else, add character to LCS and move 1 cell diagonally top-left.

	∅	A	B	A	B	B
∅	0	0	0	0	0	0
A	0	1	1	1	1	1
A	0	1	1	2	2	2
B	0	1	**2**	2	3	3
A	0	1	2	**3**	3	3
B	0	1	2	3	**4**	4

LCS : **A B A B**

Procedure

From the table, deduce LCS by starting from the bottom-right cell. Compare cell value with values of top and left cells.

- If cell value $\in$ {top cell value, left cell value}, move to the one with maximum value.
- Else, add character to LCS and move 1 cell diagonally top-left.

	∅	A	B	A	B	B
∅	0	0	0	0	0	0
A	0	**1**	1	1	1	1
A	0	1	1	2	2	2
B	0	1	**2**	2	3	3
A	0	1	2	**3**	3	3
B	0	1	2	3	**4**	4

LCS : **A B A B**

Procedure

From the table, deduce LCS by starting from the bottom-right cell. Compare cell value with values of top and left cells.

- If cell value $\in$ {top cell value, left cell value}, move to the one with maximum value.
- Else, add character to LCS and move 1 cell diagonally top-left.

	∅	A	B	A	B	B
∅	0	0	0	0	0	0
A	0	**1**	1	1	1	1
A	0	1	1	2	2	2
B	0	1	**2**	2	3	3
A	0	1	2	**3**	3	3
B	0	1	2	3	**4**	4

LCS : **A B A B**

3. Data analysis of LCS results

3. Data analysis of LCS results

3.1 Average LCS length

3. Data analysis of LCS results

3.2 Normal fit

Thank you