Longest Common Subsequence

- X = <D, A, B, D, D, C, A, C>
- Y = <C, B, A, B, D, C, A, D>
- LCS Z = <A, B, D, C, A>

- Z is a subsequence of both X and Y, and there is no longer sequence that is a subsequence of both X and Y.
- Algorithm to compute Z efficiently.

Another example:

- X = <A, B, C, B, D, A, B>
- Y = <B, D, C, A, B, A>
- LCS Z = <B, C, B, A>

Look at last character of X and Y; compare to last character of Z.

Is it possible for the last character of X to be part of the solution?
NO. If it were the, last character of Z would be "B".

Notation X_i = < first *i* characters in X>. E.g.

X_3 = < A, B, C>

Similarly, we can define Y_i, Z_i.

Z must be an LCS for X_6 = <A, B, C, B, D, A> and Y. If not, there must be a longer common subsequence Z'; but Z' would also be an LCS of X, Y. That's a contradiction, because started with Z being the LCS.

CASE 1: last character of X and Y are different, then one of them is not in the solution Z.

X_6 = <A, B, C, B, D, A> and Y = <B, D, C, A, B, A>

Now the last characters are the same, and they're same as the last character of Z.

Claim: Z_3 is an LCS for X_5 = <A, B, C, B, D> and Y_5 = <B, D, C, A, B>.

Z_3 = <B, C, B>. If this were not an LCS for X_5 and Y_5, then there would exist a common subsequence Z' that is longer than Z_3. Then if we appended "A" to Z' to create Z'', then Z'' would have be to an LCS for X_6, Y, but Z'' is longer than Z, which is a contradiction.

CASE 2: last character of X, Y, and Z are the same, then Z shortened by one character is an LCS for X and Y, both shortened by one character.

PROOF OF OPTIMAL SUBSTRUCTURE

Theorem: Let X = < x1, x2, ..., xm> and Y = <y1, y2, ..., yn>. Let Z be an LCS for X and Y, Z = <z1, z2, ..., zk>.

(1) if xm = yn, then zk = xm = yn, and Z_(k-1) is an LCS X_(m-1) and Y_(n-1).

(2) if xm != yn and zk != xm, then Z is an LCS for X_(m-1), Y.

(3) if xm != yn and zk != yn, then Z is an LCS for X, Y_(n-1).

Proof: (1) xm = yn. Then zk = xm = yn; otherwise, then xm, yn are not used in Z, and we could extend Z by one more character to produce a common subsequence of length k+1. Contradiction.

Now suppose Z_(k-1) is **not** an LCS for X_(m-1) and Y_(n-1). Then there must exist a common sequence Z' of X_(m-1) and Y_(n-1) that is longer than Z_(k-1). Construct Z'' = Z' append with < xm >, but then Z'' must be a common sequence of X and Y and Z'' is longer than Z, which is a contradiction.

[Aside]
X = <A, B, C, B, D, A> and Y = <B, D, C, A, B, A>. Z = <B, C, B, A>

(2) xm != yn, then if zk != xm Z is an LCS of X_(m-1), Y. If zk != xm, then Z must be a subsequence of X_(m-1); certainly it is also a subsequence of Y. Therefore, it is a common subsequence of X_(m-1) and Y.

Suppose Z is **not** an LCS for X_(m-1), Y. Then there exists a common subsequence Z' that is longer. Well, Z' must also be common subsequence of X and Y, but Z' is longer than Z. Contradiction. Therefore Z is an LCS for X_(m-1) and Y.

(3) xm != yn, then if zk != yn Z is an LCS of X, Y_(n-1).

Proof is similar to (2). QED.

1. If $x_m = y_n$, solve the sub problem LCS of X_{m-1}, Y_{n-1}; append x_m to the LCS for the subproblem.
2. If $x_m \neq y_n$ solve two subproblems
   LCS of X_{m-1}, Y
   LCS of X, Y_{n-1}
   Use whichever gives the longest sequence.

Let $c[i,j]$ be the length of an LCS for X_i and Y_j, $0 \le i \le m$, $0 \le j \le n$.

Note that $c[i,j] = 0$ if $i = 0$ or $j = 0$

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

Complete the c-table for the sequences ACGAAACGGC and GCGTCATC.

j

C:

		0	1	2	3	4	5	6	7	8
			G	C	G	T	C	A	T	C
0		0	0	0	0	0	0	0	0	0
1	A	0	0	0	0	0	0	1	1	1
2	C	0	0	1	1	1	1	1	1	2
3	G	0	1	1	2	2	2	2	2	2
4	A	0	1	1	2	2	2	3	3	3
5	A	0	1	1	2	2	2	3	3	3
6	A	0	1	1	2	2	2	3	3	3
7	C	0	1	2	2	2	3	3	3	4
8	G	0	1	2	3	3	3	3	3	4
9	G	0	1	2	3	3	3	3	3	4
10	C	0	1	2	3	3	4	4	4	4 *

i

LCS length for $\underline{X}$, $\underline{Y}$ is 4

C, A, G, C

LCS is ⟨C, G, A, C⟩