Longest Common Subsequence

A subsequence of a string $S$, is a set of characters that appear in left-to-right order, but not necessarily consecutively.

Example

$$ACTTGC$$

- $ACT$, $ATTC$, $T$, $ACTTGC$ are all subsequences.
- $TTA$ is not a subsequence

A common subsequence of two strings is a subsequence that appears in both strings. A longest common subsequence is a common subsequence of maximal length.

Example

$$S_1 = AAACCGTGAGTTATTGCCTAGAA$$
$$S_2 = CACCCCTAAGGTACCTTTGGTTC$$
Example

\[ S_1 = AAACCGTGAGTTATTCTGTCTAGAA \]
\[ S_2 = CACCCCTAAGGTACCTTTGTTTC \]

LCS is

\[ ACCTAGTACTTTTGG \]

Has applications in many areas including biology.
Algorithm 1

Enumerate all subsequences of $S_1$, and check if they are subsequences of $S_2$.

Questions:
- How do we implement this?
- How long does it take?
Optimal Substructure

**Theorem** Let $X = \langle x_1, x_2, \ldots, x_m \rangle$ and $Y = \langle y_1, y_2, \ldots, y_n \rangle$ be sequences, and let $Z = \langle z_1, z_2, \ldots, z_k \rangle$ be any LCS of $X$ and $Y$.

1. If $x_m = y_n$, then $z_k = x_m = y_n$ and $Z_{k-1}$ is an LCS of $X_{m-1}$ and $Y_{n-1}$.
2. If $x_m \neq y_n$, then $z_k \neq x_m$ implies that $Z$ is an LCS of $X_{m-1}$ and $Y$.
3. If $x_m \neq y_n$, then $z_k \neq y_n$ implies that $Z$ is an LCS of $X$ and $Y_{n-1}$. 
Proof

Let $X = \langle x_1, x_2, \ldots, x_m \rangle$ and $Y = \langle y_1, y_2, \ldots, y_n \rangle$ be sequences, and let $Z = \langle z_1, z_2, \ldots, z_k \rangle$ be any LCS of $X$ and $Y$.

1. If $x_m = y_n$, then $z_k = x_m = y_n$ and $Z_{k-1}$ is an LCS of $X_{m-1}$ and $Y_{n-1}$.

2. If $x_m \neq y_n$, then $z_k \neq x_m$ implies that $Z$ is an LCS of $X_{m-1}$ and $Y$.

3. If $x_m \neq y_n$, then $z_k \neq y_n$ implies that $Z$ is an LCS of $X$ and $Y_{n-1}$.

Proof

1. If $z_k \neq x_m$, then we could append $x_m = y_n$ to $Z$ to obtain a common subsequence of $X$ and $Y$ of length $k + 1$, contradicting the supposition that $Z$ is a longest common subsequence of $X$ and $Y$. Thus, we must have $z_k = x_m = y_n$. Now, the prefix $Z_{k-1}$ is a length-$(k-1)$ common subsequence of $X_{m-1}$ and $Y_{n-1}$. We wish to show that it is an LCS. Suppose for the purpose of contradiction that there is a common subsequence $W$ of $X_{m-1}$ and $Y_{n-1}$ with length greater than $k - 1$. Then, appending $x_m = y_n$ to $W$ produces a common subsequence of $X$ and $Y$ whose length is greater than $k$, which is a contradiction.

2. If $z_k \neq x_m$, then $Z$ is a common subsequence of $X_{m-1}$ and $Y$. If there were a common subsequence $W$ of $X_{m-1}$ and $Y$ with length greater than $k$, then $W$ would also be a common subsequence of $X_m$ and $Y$, contradicting the assumption that $Z$ is an LCS of $X$ and $Y$. 
3. The proof is symmetric to the previous case.
Recursion for length

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

\[
LCS - Length(X, Y)
\]

1. \[ m \leftarrow \text{length}[X] \]
2. \[ n \leftarrow \text{length}[Y] \]
3. \[ \text{for } i \leftarrow 1 \text{ to } m \]
   \[ \text{do } c[i, 0] \leftarrow 0 \]
4. \[ \text{for } j \leftarrow 0 \text{ to } n \]
   \[ \text{do } c[0, j] \leftarrow 0 \]
5. \[ \text{for } i \leftarrow 1 \text{ to } m \]
   \[ \text{do } \text{for } j \leftarrow 1 \text{ to } n \]
   \[ \text{do if } x_i = y_j \]
   \[ \text{then } c[i, j] \leftarrow c[i - 1, j - 1] + 1 \]
   \[ b[i, j] \leftarrow "\" \]
   \[ \text{else if } c[i - 1, j] \geq c[i, j - 1] \]
   \[ \text{then } c[i, j] \leftarrow c[i - 1, j] \]
   \[ b[i, j] \leftarrow "↑" \]
   \[ \text{else } c[i, j] \leftarrow c[i, j - 1] \]
   \[ b[i, j] \leftarrow "←" \]
6. \[ \text{return } c \text{ and } b \]