Longest Common Subsequence (LCS)

C.S. Marron
cmarron@umbc.edu

CMSC 441 — Algorithms

Outline

Sequences and Subsequences

Definitions

- A *sequence* is an ordered multiset in which the elements are taken from an underlying set. The usual sequence notation uses subscripts, e.g.,

$$X = \langle x_1, x_2, \ldots, x_n \rangle$$

where x_i are elements of the underlying set.

- A *subsequence* of a sequence is an ordered multi-subset. Using subscript notation, if $X = \langle x_1, x_2, \ldots, x_n \rangle$ is a sequence, then a subsequence is

$$\langle x_{i_1}, x_{i_2}, \ldots, x_{i_k} \rangle$$

where $1 \leq i_1 < i_2 < \cdots < i_k \leq n$.

Common Subsequences

Definitions

- If X and Y are sequences, Z is a *common subsequence* if it is a subsequence of both X and Y.
- A common subsequence of X and Y with maximum length is a *longest common subsequences* (LCS).

An LCS is not necessarily unique, although its length is!

An Example

Set = {A, B, C, D}

- $X = \langle D, A, B, D, D, C, D, A \rangle$
- $Y = \langle C, B, A, B, D, C, A, D \rangle$
- $\langle D, A, B \rangle$ is a subsequence of X, but *not* of Y.
- $\langle D, A, D \rangle$ is a subsequence of both X and Y. It is a *common subsequence*, but it is not a longest common subsequence.
- $Z = \langle A, B, D, C, D \rangle$ is a *longest common subsequence* of X and Y.
- $Z' = \langle A, B, D, C, A \rangle$ is also an LCS of X and Y. The sequence is not unique, but the length (5) is.

An Example

- $X = \langle D, A, B, D, D, C, D, A \rangle$
- $Y = \langle C, B, A, B, D, C, A, D \rangle$
- $\langle D, A, B \rangle$ is a subsequence of X, but *not* of Y.
- $\langle D, A, D \rangle$ is a subsequence of both X and Y. It is a *common subsequence*, but it is not a longest common subsequence.
- $Z = \langle A, B, D, C, D \rangle$ is a *longest common subsequence* of X and Y.
- $Z' = \langle A, B, D, C, A \rangle$ is also an LCS of X and Y. The sequence is not unique, but the length (5) is.

An Example

- $X = \langle D, A, B, D, D, C, D, A \rangle$
- $Y = \langle C, B, A, B, D, C, A, D \rangle$
- $\langle D, A, B \rangle$ is a subsequence of X, but *not* of Y.
- $\langle D, A, D \rangle$ is a subsequence of both X and Y. It is a *common subsequence*, but it is not a longest common subsequence.
- $Z = \langle A, B, D, C, D \rangle$ is a *longest common subsequence* of X and Y.
- $Z' = \langle A, B, D, C, A \rangle$ is also an LCS of X and Y. The sequence is not unique, but the length (5) is.

An Example

- $X = \langle D, A, B, D, D, C, D, A \rangle$
- $Y = \langle C, B, A, B, D, C, A, D \rangle$
- $\langle D, A, B \rangle$ is a subsequence of X, but *not* of Y.
- $\langle D, A, D \rangle$ is a subsequence of both X and Y. It is a *common subsequence*, but it is not a longest common subsequence.
- $Z = \langle A, B, D, C, D \rangle$ is a *longest common subsequence* of X and Y.
- $Z' = \langle A, B, D, C, A \rangle$ is also an LCS of X and Y. The sequence is not unique, but the length (5) is.

An Example

- $X = \langle D, A, B, D, D, C, D, A \rangle$
- $Y = \langle C, B, A, B, D, C, A, D \rangle$
- $\langle D, A, B \rangle$ is a subsequence of X, but *not* of Y.
- $\langle D, A, D \rangle$ is a subsequence of both X and Y. It is a *common subsequence*, but it is not a longest common subsequence.
- $Z = \langle A, B, D, C, D \rangle$ is a *longest common subsequence* of X and Y. Length 5
- $Z' = \langle A, B, D, C, A \rangle$ is also an LCS of X and Y. The sequence is not unique, but the length (5) is.

Optimization: find LCS length.

Using the Example

X = <D,A, B, D, D, C, D, A>. Y = <C, B, A, B, D, C, A, D>. Z = <A, B, D, C, D>
Look at what happens considering the last characters of each sequence.

What can we say when the last character of X and the last character of Y are different?

x_8 is not used in Z. We can shorten X by one character and Z is still an LCS.

* Z is an LCS of X_7, Y

Consider X and Y_7 = <C, B, A, B, D, C, A>.
The last characters are the same
Z_1 = <A, B, D, C, A> is an LCS.

Z_1 must end in A. Suppose it didn't:
then we could extend Z_1 by one character ('A').
Contradiction.

* Z_{k-1} must be an LCS of X_{m-1}, Y_{n-1}.

Optimal Substructure Theorem

Notation

If $X = \langle x_1, x_2, \ldots x_m \rangle$ is a sequence, then we define X_i, $1 \leq i \leq m$, as $X_i = \langle x_1, x_2, \ldots x_i \rangle$. The same notation is used for other sequences, e.g. Y_j.

$Z_k = \langle z_1, z_2, \ldots, z_k \rangle$.

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. Suppose that $Z = \langle z_1, z_2, \ldots, z_k \rangle$ is an 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 for X_{m-1} and Y_{n-1}.*
2. *if $x_m \neq y_n$ and $z_k \neq x_m$, then Z is an LCS for X_{m-1} and Y.*
3. *If $x_m = y_n$ and $z_k \neq y_n$, then Z is an LCS for X and Y_{n-1}.*

Optimal Substructure Theorem

Notation

If $X = \langle x_1, x_2, \ldots x_m \rangle$ is a sequence, then we define X_i, $1 \leq i \leq m$, as $X_i = \langle x_1, x_2, \ldots x_i \rangle$. The same notation is used for other sequences, e.g. Y_j.

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. Suppose that $Z = \langle z_1, z_2, \ldots, z_k \rangle$ is an 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 for X_{m-1} and Y_{n-1}.*
2. *if $x_m \neq y_n$ and $z_k \neq x_m$, then Z is an LCS for X_{m-1} and Y.*
3. *If $x_m = y_n$ and $z_k \neq y_n$, then Z is an LCS for X and Y_{n-1}.*

Optimal Substructure Theorem

Notation

If $X = \langle x_1, x_2, \ldots x_m \rangle$ is a sequence, then we define X_i, $1 \leq i \leq m$, as $X_i = \langle x_1, x_2, \ldots x_i \rangle$. The same notation is used for other sequences, e.g. Y_j.

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. Suppose that $Z = \langle z_1, z_2, \ldots, z_k \rangle$ is an 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 for X_{m-1} and Y_{n-1}.*
2. *if $x_m \neq y_n$ and $z_k \neq x_m$, then Z is an LCS for X_{m-1} and Y.*
3. *If $x_m \neq y_n$ and $z_k \neq y_n$, then Z is an LCS for X and Y_{n-1}.*

Proof of the Theorem

(1) $x_m = y_n$. Then $z_k = x_m = y_n$. If it were not, we could append x_m to Z_i to get a common subsequence of length $k+1$. Contradiction!

Need to show Z_{k-1} is LCS for X_{m-1}, Y_{n-1}.

Suppose not. Z_{k-1} is a common subsequence.

There must be a longer c.s. W of X_{m-1}, Y_{n-1}.

$W = \langle w_1, w_2, \ldots, w_k \rangle$

Then $W' = \langle w_1, w_2, \ldots, w_k, x_m \rangle$ is an LCS of X and Y. Contradiction. W' is longer than Z_i!

(2) $x_m \neq y_n$ and $z_k \neq x_m$. Show that Z_i is an LCS of X_{m-1}, Y. Z_i a c.s. of X_{m-1}, Y. If Z_i is not LCS then there is a c.s. W that is longer than Z_i. W is a c.s. of X, Y.

Contradiction. W is a C.S. of X, Y longer than Z_i!

Turn this into a solution...

- If x_m = y_n, solve the subproblem
  - LCS of X_(m-1), Y_(n-1) and append x_m to get LCS of X, Y.

- If x_m != y_n, solve the subproblems and use the larger solution
  - LCS of X_(m-1), Y
  - LCS of X, Y_(n-1)

Let C[i,j] be an m-by-n array. C[i,j] will hold the LCS length for X_i, Y_j, where 0 <= i <= m and 0 <= j <= n.

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

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

		G	C	G	T	C	A	T	C
	0	0	0	0	0	0	0	0	0
A	0	0	0	0	0	0	1	1	1
C	0	0	1	1	1	1	1	1	2
G	0	1	1	2	2	2	2	2	2
A	0	1	1	2	2	2	3	3	3
A	0	1	1	2	2	2	3	3	3
A	0	1	1	2	2	2	3	3	3
C	0	1	2	2	2	3	3	3	4
G	0	1	2	3	3	3	3	3	4
G	0	1	2	3	3	3	3	3	4
C	0	1	2	3	3	4	4	4	4

C array

i

j

Update m·n table. Each update is O(1). Solution is O(m·n)

C, G, C, G An LCS is ⟨G, C, G, C⟩.