The Longest Common Subsequence problem

K. Subramani
LCSEE,
West Virginia University,
Morgantown, WV
{ksmani@csee.wvu.edu}

1 Problem Definition

Let $\mathbf{X} =< x_1, x_2, \ldots x_m >$ and $\mathbf{Y} =< y_1, y_2, \ldots, y_n >$ denote two sequences on a fixed alphabet Σ (You may assume that $\Sigma = \{0, 1\}$, if it helps). Devise an efficient algorithm to determine the length of the Longest Common Subsequence between $\mathbf{X}$ and $\mathbf{Y}$. Note that a subsequence of a sequence need not be contiguous.

2 Solution

Let $\mathbf{Z} =< z_1, z_2, \ldots, z_r >$ denote the Longest Common Subsequence (LCS) between $\mathbf{X}$ and $\mathbf{Y}$. We use the following notational scheme:

(a) $\mathbf{X_i}$ denotes the sequence $< x_1, x_2, \ldots, x_i >$.

(b) $\mathbf{LC}(\mathbf{X},\ \mathbf{Y})$ denotes the LCS between $\mathbf{X}$ and $\mathbf{Y}$.

(c) $f[i][j]$ denotes the length of the LCS between $\mathbf{X_i}$ and $\mathbf{Y_j}$, i.e., $f[i][j] = |\mathbf{LC}(\mathbf{X_i},\ \mathbf{Y_j})|$.

Consider the following cases:

(i) $x_m = y_n$ - In this case, we must have $z_r = x_m$ and further, $\mathbf{Z_{r-1}} = \mathbf{LC}(\mathbf{X_{m-1}}, \mathbf{Y_{n-1}})$. Observe that if $z_r \neq x_m, y_n$, then the length of the LCS can be increased by at least 1, by appending x_m to $\mathbf{Z}$, thereby contradicting the optimality of $\mathbf{Z}$. Given that $z_r = x_m$, it is easy to see that $\mathbf{Z_{r-1}} = \mathbf{LC}(\mathbf{X_{m-1}},\ \mathbf{Y_{n-1}})$. If this were not the case, then we can determine $\mathbf{LC}(\mathbf{X_{m-1}},\ \mathbf{Y_{n-1}})$ and append x_m to it, thereby contradicting the optimality of $\mathbf{Z}$.

(ii) $x_m \neq y_n$ and $z_r \neq x_m$- We must have $\mathbf{Z} = \mathbf{LC}(X_{m-1}, Y_n)$. If this were not the case, then there is a subsequence between $\mathbf{X}$ and $\mathbf{Y}$ of length greater than k. However, this subsequence is also a subsequence between $\mathbf{X_{m-1}}$ and Y, thereby contradicting the optimality of $\mathbf{Z}$.

(iii) $x_m \neq y_n$ and $z_r \neq y_n$ - We must have $\mathbf{Z} = \mathbf{LC}(X_m, Y_{n-1})$. Symmetric to the case above.

From the above discussion, the following recurrence relation suggests itself for the length of the LCS beween $\mathbf{X_i}$ and $\mathbf{Y_j}$:

$$\begin{aligned} f[i][j] &= 0, \text{ if } \mathrm{i} = 0 \text{ or } \mathrm{j} = 0 \\ &= \max(f[i-1][j], f[i][j-1]), \text{ if } \mathrm{x_i} \neq \mathrm{y_j} \text{ and } \mathrm{i, j} \geq 1 \\ &= 1 + f[i-1][j-1], \text{ if } \mathrm{x_i} = \mathrm{y_j} \text{ and } \mathrm{i, j} \geq 1 \end{aligned}$$

This recurrence relation can be easily converted into an algorithm that computes $f[m][n]$ in bottom-up fashion and runs in time $O(m \cdot n)$.