CSE 417 Algorithms and Complexity

Richard Anderson
Lecture 22
Longest Common Subsequence
Longest Common Subsequence

• $C = c_1 \ldots c_g$ is a subsequence of $A = a_1 \ldots a_m$ if $C$ can be obtained by removing elements from $A$ (but retaining order)

• $LCS(A, B)$: A maximum length sequence that is a subsequence of both $A$ and $B$

occurranec  attacggct
occurrence   tacgacca
Determine the LCS of the following strings

BARTHOLEMEWSIMPSON

KRUSTYTHECLOWN
String Alignment Problem

- Align sequences with gaps
  
  CAT TGA AT
  
  CAGAT AGGA

- Charge $\delta_x$ if character $x$ is unmatched
- Charge $\gamma_{xy}$ if character $x$ is matched to character $y$

Note: the problem is often expressed as a minimization problem, with $\gamma_{xx} = 0$ and $\delta_x > 0$
LCS Optimization

• $A = a_1 a_2 \ldots a_m$
• $B = b_1 b_2 \ldots b_n$

• $\text{Opt}[j, k]$ is the length of $\text{LCS}(a_1 a_2 \ldots a_j, b_1 b_2 \ldots b_k)$
Optimization recurrence

If $a_j = b_k$, $\text{Opt}[j,k] = 1 + \text{Opt}[j-1,k-1]$

If $a_j \neq b_k$, $\text{Opt}[j,k] = \max(\text{Opt}[j-1,k], \text{Opt}[j,k-1])$
Give the Optimization Recurrence for the String Alignment Problem

• Charge $\delta_x$ if character $x$ is unmatched
• Charge $\gamma_{xy}$ if character $x$ is matched to character $y$

$$\text{Opt}[j, k] =$$

Let $a_j = x$ and $b_k = y$
Express as minimization
String edit with Typo Distance

• Find closest dictionary word to typed word
  • Dist('a', 's') = 1
  • Dist('a', 'u') = 6
• Capture the likelihood of mistyping characters
• Different distance model for T9 on basic mobile phone
Dynamic Programming Computation
Code to compute \( \text{Opt}[n, m] \)

\[
\begin{align*}
\text{for} & \quad (\text{int} \ i = 0; \ i < n; \ i++) \\
\text{for} & \quad (\text{int} \ j = 0; \ j < m; \ j++) \\
& \quad \text{if} \quad (A[i] == B[j]) \\
& \quad \quad \text{Opt}[i,j] = \text{Opt}[i-1, j-1] + 1; \\
& \quad \text{else if} \quad (\text{Opt}[i-1, j] \geq \text{Opt}[i, j-1]) \\
& \quad \quad \text{Opt}[i, j] := \text{Opt}[i-1, j]; \\
& \quad \text{else} \\
& \quad \quad \text{Opt}[i, j] := \text{Opt}[i, j-1];
\end{align*}
\]
Storing the path information

\begin{align*}
A[1..m], & \quad B[1..n] \\
\text{for } i := 1 \text{ to } m \quad & \quad \text{Opt}[i, 0] := 0; \\
\text{for } j := 1 \text{ to } n \quad & \quad \text{Opt}[0,j] := 0; \\
\text{Opt}[0,0] := 0; \quad & \\
\text{for } i := 1 \text{ to } m \\
\quad & \quad \text{for } j := 1 \text{ to } n \\
\quad & \quad \quad \text{if } A[i] = B[j] \quad \{ \quad \text{Opt}[i,j] := 1 + \text{Opt}[i-1,j-1]; \quad \text{Best}[i,j] := \text{Diag}; \quad \} \\
\quad & \quad \quad \text{else if } \text{Opt}[i-1,j] \geq \text{Opt}[i,j-1] \\
\quad & \quad \quad \quad \{ \quad \text{Opt}[i,j] := \text{Opt}[i-1,j], \quad \text{Best}[i,j] := \text{Left}; \quad \} \\
\quad & \quad \quad \text{else} \quad \{ \quad \text{Opt}[i,j] := \text{Opt}[i,j-1], \quad \text{Best}[i,j] := \text{Down}; \quad \}
\end{align*}
Reconstructing Path from Distances
How good is this algorithm?

• Is it feasible to compute the LCS of two strings of length 300,000 on a standard desktop PC? Why or why not.
public int ComputeLCS() {
    int n = str1.Length;
    int m = str2.Length;

    int[,] opt = new int[n + 1, m + 1];
    for (int i = 0; i <= n; i++)
        opt[i, 0] = 0;
    for (int j = 0; j <= m; j++)
        opt[0, j] = 0;

    for (int i = 1; i <= n; i++)
        for (int j = 1; j <= m; j++)
            if (str1[i - 1] == str2[j - 1])
                opt[i, j] = opt[i - 1, j - 1] + 1;
            else if (opt[i - 1, j] >= opt[i, j - 1])
                opt[i, j] = opt[i - 1, j];
            else
                opt[i, j] = opt[i, j - 1];

    return opt[n, m];
}
N = 17000

Runtime should be about 5 seconds*

* Personal PC, 6 years old
public int SpaceEfficientLCS() {
    int n = str1.Length;
    int m = str2.Length;
    int[] prevRow = new int[m + 1];
    int[] currRow = new int[m + 1];

    for (int j = 0; j <= m; j++)
      prevRow[j] = 0;

    for (int i = 1; i <= n; i++) {
      currRow[0] = 0;
      for (int j = 1; j <= m; j++) {
        if (str1[i - 1] == str2[j - 1])
          currRow[j] = prevRow[j - 1] + 1;
        else if (prevRow[j] >= currRow[j - 1])
          currRow[j] = prevRow[j];
        else
          currRow[j] = currRow[j - 1];
      }
      for (int j = 1; j <= m; j++)
        prevRow[j] = currRow[j];
    }
    return currRow[m];
}
N = 300000

N: 10000 Base 2 Length: 8096  Gamma: 0.8096  Runtime:00:00:01.86
N: 20000 Base 2 Length: 16231  Gamma: 0.81155  Runtime:00:00:07.45
N: 30000 Base 2 Length: 24317  Gamma: 0.8105667  Runtime:00:00:16.82
N: 40000 Base 2 Length: 32510  Gamma: 0.81275  Runtime:00:00:29.84
N: 50000 Base 2 Length: 40563  Gamma: 0.81126  Runtime:00:00:46.78
N: 60000 Base 2 Length: 48700  Gamma: 0.8116667  Runtime:00:01:08.06
N: 70000 Base 2 Length: 56824  Gamma: 0.8117715  Runtime:00:01:33.36

N: 300000 Base 2 Length: 243605  Gamma: 0.8120167  Runtime:00:28:07.32
Observations about the Algorithm

• The computation can be done in $O(m+n)$ space if we only need one column of the Opt values or Best Values

• The computation requires $O(nm)$ space if we store all of the string information
Computing LCS in $O(nm)$ time and $O(n+m)$ space

- Divide and conquer algorithm
- Recomputing values used to save space

- Section 6.7 of the text, but we will not have time to cover in detail (so you are not responsible for section 6.7)
Divide and Conquer Algorithm

- Where does the best path cross the middle column?

- For a fixed $i$, and for each $j$, compute the LCS that has $a_i$ matched with $b_j$
Algorithm Analysis

• $T(m,n) = T(m/2, j) + T(m/2, n-j) + cnm$
• Solution: $T(m,n) \leq 2cnm$