Outline

Last time:

- Review of P/NP and 2-approximate solution to TSP
- Dynamic Programming:
  - 0/1 Knapsack
  - Travelling Salesperson

Today:

- finish up Bitonic merge
- Longest Common Subsequence problem
Longest Common Subsequence (LCS)

Examples of LCS applications:

- approximate matching, e.g., spell checker
- file comparison (used in `diff`, `cvs`, cheater finder)
- searching for proteins with similar DNA sequences, etc.

For string `GDVEGTA`, `GVEA` is an example **subsequence** of the string.

Contrast:

- **substring**: elements in the substring must be consecutive in the string
- **subsequence**: not

Example: given 2 strings, `GCVEGTA` and `GVCEKST`, the **longest common subsequence** is `GVET`, note `CET` is also a common subsequence.

Sugih Jamin (jamin@eecs.umich.edu)
**LCS Definition**

More formally:

- given 2 sequences $a[1, \ldots, n]$ and $b[1, \ldots, m]$,
- find subsequences $a[i_1, \ldots, i_k]$ and $b[j_1, \ldots, j_k]$ where $i_l < i_{l+1}$ and $j_l < j_{l+1}$,
- such that $a[i_1] == b[j_1], a[i_2] == b[j_2], \ldots, a[i_k] == b[j_k]$
- for $k$ as large as possible

**Brute force approach:**

**Running time:**
Brute Force LCS

Given strings $a$ and $b$, of length $n$ and $m$ respectively:

- each element of $a$ is either in or not in a subsequence
- there are $2^n$ different subsequences
- each subsequence must be matched against $b$’s $m$ elements
- Running time: $O(m2^n)$
Dynamic Programming for LCS

Define $L[i, j] = k$ to be the length of the LCS of $a[1, \ldots, i]$ and $b[1, \ldots, j]$, where $1 \leq i \leq n$ and $1 \leq j \leq m$

Question: Can we find overlapping subproblems such that the optimal solution to the LCS problem consists of optimal solutions to the subproblems?

Would the optimal solution for $L[i, j]$ consists of optimal solutions for $L[i', j']$, where $i' \leq i$ and $j' \leq j$?

Use similar approach to solving for the 0/1 Knapsack problem: consider adding or not adding the last item to the solution of the subproblem
DP for LCS Example

Compute the LCS of $GTTCCCTAATA$ and $CGATAATTGAGA$ by filling in the table $L$:

<table>
<thead>
<tr>
<th></th>
<th>$b$</th>
<th>C</th>
<th>G</th>
<th>A</th>
<th>T</th>
<th>A</th>
<th>A</th>
<th>T</th>
<th>T</th>
<th>G</th>
<th>A</th>
<th>G</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a$</td>
<td>$i \setminus j$</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>G</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>T</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>T</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>C</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>C</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>T</td>
<td>6</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>A</td>
<td>7</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>A</td>
<td>8</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>A</td>
<td>9</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>A</td>
<td>10</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>6</td>
</tr>
</tbody>
</table>

Running time:

Sugih Jamin (jamin@eecs.umich.edu)
DP for LCS

Let $L[i, 0] = L[0, j] = 0$, $k$ the length of the LCS

Case 1: if $a[i] \neq b[j]$, one of them cannot be in the LCS of $L[i, j]$, so either $L[i, j] = L[i, j - 1]$ or $L[i, j] = L[i - 1, j]$

Since we want the longest common subsequence, $L[i, j] = \text{MAX}(L[i, j - 1], L[i - 1, j])$

Case 2: if $a[i] == b[j]$, then $L[i - 1, j - 1]$ must be $k - 1$

(Thus $L[i, j]$ is optimal and consists of optimal solutions to the subproblems)

Proof by contradiction: assume $L[i - 1, j - 1] \geq k$, then adding $a[i]$ to $a[1, \ldots, i - 1]$ and $b[j]$ to $b[1, \ldots, j - 1]$ makes $L[i, j] \geq k + 1 \Rightarrow$ contradiction

So:

$$L[i, j] = \begin{cases} 
0 & \text{if } i = 0 \text{ or } j = 0, \\
L[i - 1, j - 1] + 1 & \text{if } i, j > 0 \text{ and } a[i] == b[j], \\
\text{MAX}(L[i - 1, j], L[i, j - 1]) & \text{otherwise}
\end{cases} \quad (1)$$
LCS Extraction

Use stack to accumulate LCS in reversed order

Start from $L[n, m]$:

- if $L[i] == L[j]$, found a match, push $L[i]$, move to $L[i-1][j-1]$
- otherwise, move to the larger of $L[i-1][j]$ and $L[i][j-1]$, if $L[i-1][j] == L[i][j-1]$, always move horizontal or vertical

Alternative algorithm: again, start from $L[n, m]$:

- if $L[i, j] == L[i-1, j]$ move to $L[i-1, j]$
- else if $L[i, j] == L[i, j-1]$ move to $L[i, j-1]$
- (the above two steps can be swapped)
- else push $L[i]$, move to $L[i-1, j-1]$

Pop stack (if there are multiple common subsequences with the same length, each of the above extracts only one)
LCS Extraction Example

LCS: **CTAATA** or **GTTTAA** (or GTAATA, alt. alg.)

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>G</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>T</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>T</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>C</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>T</td>
<td>6</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>A</td>
<td>7</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>A</td>
<td>8</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>T</td>
<td>9</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>A</td>
<td>10</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>6</td>
</tr>
</tbody>
</table>

Running time:

Sugih Jamin (jamin@eecs.umich.edu)