Edit Distance

- Designed to give a numerical value to the similarity between two strings

- Used in speech recognition, spelling corrections, DNA analysis, etc.
Edit Distance: DNA Analysis

- Mutation in DNA is evolutionary
- DNA replication errors cause
  - Substitutions,
  - Insertions,
  - and Deletions…
  of nucleotides, leading to "edited" DNA texts

Edit Distance: Definition

- Introduced by Vladimir Levenshtein in 1966
- Similar to error detecting Hamming Code (1950)

- The Edit Distance between two strings is the minimum number of editing operations needed to transform one string into another

- Operations:
  - Insertion of a symbol
  - Deletion of a symbol
  - Substitution of one symbol for another.

Note: These may have different associated "costs".
Edit Distance: Example

How would you transform:

X: TGCATAT

to the string:

Y: ATCCGAT

Edit Distance: Construction

• Constructed in a similar manner to finding LCS using Dynamic Programming

• Assume X has length m, Y has length n
• Define \( d(i,j) = \text{edit distance of } X[1,\ldots,i] \text{ and } Y[1,\ldots,j] \)

• Compute \( d(i,j) \) for small \( i \) and \( j \) and use these values to construct larger \( d(i,j) \) until we reach \( d(m,n) \)
Recursive Solution

- Initialize matrix $d$:
  - $d(i,0) = i$, $d(0,j) = j$ // $i =$ row, $j =$ column

Compute values in a similar row-centric way using a nested-loop the values:

$$d[i,j] = \begin{cases} d[i-1,j-1] & \text{if } x_i = y_i \\ d[i-1,j] + 1 & \text{if } x_i \neq y_i \\ d[i,j-1] + 1 & \text{if } x_i \neq y_i \\ \min \{ \{d[i-1,j-1] + 1 & \text{if } x_i \neq y_i \} \} \end{cases}$$

Arrows point to the cell used to compute the current cell's value.

Example: Edit Distance Matrix

- $X = \text{ATCGTT}$
- $Y = \text{AGTTAC}$

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>G</th>
<th>T</th>
<th>T</th>
<th>A</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>A</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

http://www.let.rug.nl/~kleiweg/lev/