Edit distance

• GAME > DAME > DOME > DOSE > LOSE
• Edit distance between two strings $S_1$ and $S_2$ is the minimum number of operations (insertion/deletion/replacement) that transform $S_1$ to $S_2$.
• Levenshtein distance 1966
• Transcript – string over alphabet I,D,R,M
• Example:
  ALA   SKA
  ALABAMA
  MMMIRRM
• Alignment vs edit transcript
Dynamic programming

• For two strings S1 and S2, D(i,j) is defined to be the edit distance of S1[1..i] and S2[1..j].
• D(n,m) is the solution.
• D(i,0)=i and D(0,j)=j
• D(i,j)=\min\{D(i-1,j)+1, D(i,j-1)+1, D(i-1,j-1)+t(i,j)\}
• t(i,j)=1 if S1(i) and S2(j) are not equal,
• t(i,j)=0 if S1(i) and S2(j) are equal.
Correctness of the recurrence

- Last symbol of transcript
  - I, then $D(i,j) = D(i, j-1) + 1$
  - D, then $D(i,j) = D(i-1, j) + 1$
  - R, then $D(i,j) = D(i-1, j-1) + 1$
  - M, then $D(i,j) = D(i-1, j-1)$
## Putting all together

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>G</th>
<th>C</th>
<th>T</th>
<th>T</th>
<th>A</th>
<th>C</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</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>
</tr>
<tr>
<td>A</td>
<td>1</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>
</tr>
<tr>
<td>C</td>
<td>2</td>
<td>1</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>
</tr>
<tr>
<td>G</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>C</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>T</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>A</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>T</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

$S_1$: ACGCTAT

$S_2$: ACGGCTTACT
## Traceback

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>G</th>
<th>C</th>
<th>T</th>
<th>T</th>
<th>A</th>
<th>C</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</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>
</tr>
<tr>
<td>A</td>
<td>1</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>
</tr>
<tr>
<td>C</td>
<td>2</td>
<td>1</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>
</tr>
<tr>
<td>G</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>C</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>T</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>A</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>T</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

\[S_1: \text{ACG CT A T}\]

\[S_2: \text{ACGGCTTACT}\]
Analysis

- i) Dynamic programming table can be computed in $O(mn)$ time.
- ii) Optimal edit transcript and optimal alignment can be found in $O(n+m)$ time if table has been computed.
- **Theorem.** Total time is $O(mn)$. 
Edit graph

- The weight of each edge is 1, except for the three zero-weight edges marked in figure.
- Optimal transcript is the shortest path from (0,0) to (n,m).
Weighted edit distance

• Weights (cost or score) may be different for transcript operations.
  • Insertion/deletion – d
  • Replacement – r
  • Match – e (usually small)

• Dynamic program still works in O(mn) time.
Alphabet-weight edit distance

- Scoring schemes:
  - PAM matrices
  - BLOSUM
- For example popular BLAST scores identities as +5 and mismatches as -4.