String Editing

String Editing

- Problem
 - Find The Edit Distance Between Two Strings

Edit Distance

Applications

- Approximate String Matching
- Spell checking
- Google finding similar word variations
- DNA sequence comparison
- Pattern Recognition

String Editing

- We are given 2 strings
- X=x1, x2,...., xn and Y=y1, y2,... ym where xi, 1<=i<=n and yj, 1<=j<=m
- Problem: Transform X→ Y
- · Edit operations:
 - Insert (I),
 - Delete (D),
 - Change (C): a symbol of X into another.

Cost (Transforming X→Y)

- Cost =Sum of costs of the individual operation in the sequence
- $D(xi) \rightarrow cost of deleting the symbol xi from X$
- $I(yj) \rightarrow cost of inserting the symbol yj into X$
- $C(xi, yj) \rightarrow cost of changing the symbol xi of X into yj$

Example

- X=x1, x2, x3, x4, x5=a, a, b, a, b
- Y=y1, y2, y3, y4 = b, a, b, ,b
- · Cost of deleting and inserting be 1
- · Cost of changing be 2
- 1<=i<=5
- 1<=j<=4
- cost (i, j): minimum cost of any edit sequence for transforming X=x1, x2, x3, x4, x5 into y1, y2, y3, y4

cost(i, j)

- 1. i=j=0, cost(i,j)=0
- j=0 and i>0→ transform X into Y by a sequence of deletes
 - cost(i,0)=cost(i-1,0) + D(xi)
- i=0 and j>0→ transform X into Y by a sequence of inserts
 - cost(0,j)=cost(0,j-1) + I(yi)

cost(i, j)

- 4. If i#0 and j#0, transforming X=x1, x2, x3, x4, x5 into y1, y2, y3, y4
- 1. Transform $x_1, x_2, \ldots, x_{i-1}$ into y_1, y_2, \ldots, y_j using a minimum-cost edit sequence and then delete x_i . The corresponding cost is $cost(i-1,j) + D(x_i)$.
- 2. Transform x_1,x_2,\ldots,x_{i-1} into y_1,y_2,\ldots,y_{j-1} using a minimum-cost edit sequence and then change the symbol x_i to y_j . The associated cost is $cost(i-1,j-1)+C(x_i,y_j)$.
- 3. Transform x_1, x_2, \ldots, x_i into $y_1, y_2, \ldots, y_{j-1}$ using a minimum-cost edit sequence and then insert y_j . This corresponds to a cost of $cost(i, j-1) + I(y_j)$.

cost(i, j)

$$cost(i,j) = \left\{ \begin{array}{ll} 0 & i=j=0 \\ cost(i-1,0) + D(x_i) & j=0, \ i>0 \\ cost(0,j-1) + I(y_j) & i=0, \ j>0 \\ cost'(i,j) & i>0, \ j>0 \end{array} \right.$$

 $\begin{array}{ll} \text{where } cost'(i,j) = \min \left\{ \begin{array}{ll} cost(i-1,j) + D(x_i), \\ cost(i-1,j-1) + C(x_i,y_j), \\ cost(i,j-1) + I(y_j) \end{array} \right\} \end{array}$

Example

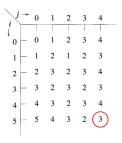
- X=x1, x2, x3, x4, x5=a, a, b, a, b
- Y=y1, y2, y3, y4= b, a, b, ,b
- · Cost of deleting and inserting be 1
- · Cost of changing be 2
- 1<=i<=5
- 1<=j<=4

Cost table

cost(1,1), cost(1,2)

$$\begin{array}{lll} cost(1,1) & = & \min \left\{ cost(0,1) + D(x_1), cost(0,0) + C(x_1,y_1), cost(1,0) + I(y_1) \right\} \\ & = & \min \left\{ 2,2,2 \right\} = 2 \\ \\ cost(1,2) & = & \min \left\{ cost(0,2) + D(x_1), cost(0,1) + C(x_1,y_2), cost(1,1) + I(y_2) \right\} \\ & = & \min \left\{ 3,1,3 \right\} = 1 \end{array}$$

Cost table



cost(5,4)

- X=x1, x2, x3, x4, x5=a, a, b, a, b
- Y=y1, y2, y3, y4 = b, a, b, b
- cost(5,4)=3.
- · Possible minimum cost edit sequence
 - 1. delete x1, delete x2, and insert y4
 - 2. Change x1→ y1 and delete x4

Edit Distance

- How many edits are needed to exactly match the Target with the Pattern
- Target: TCGACGT CAPattern: T GACGTGC
- Three:
 - By Deleting C and A from the target, and by Deleting G from the Pattern

