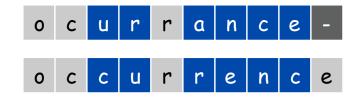
6.6 Sequence Alignment

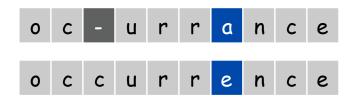
String Similarity

How similar are two strings?

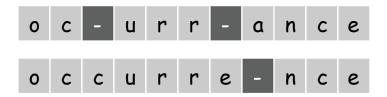
- ocurrance
- occurrence



5 mismatches, 1 gap



1 mismatch, 1 gap



0 mismatches, 3 gaps

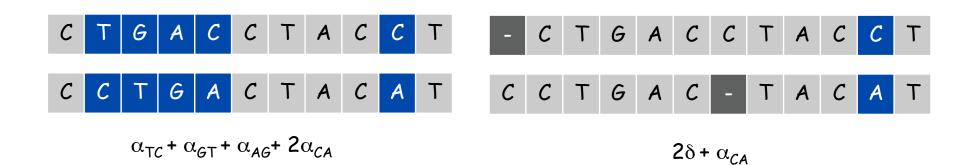
Edit Distance

Applications.

- Basis for Unix diff.
- Speech recognition.
- Computational biology.

Edit distance. [Levenshtein 1966, Needleman-Wunsch 1970]

- \blacksquare Gap penalty δ ; mismatch penalty $\alpha_{pq}.$
- Cost = sum of gap and mismatch penalties.



Sequence Alignment

Goal: Given two strings $X = x_1 x_2 ... x_m$ and $Y = y_1 y_2 ... y_n$ find alignment of minimum cost.

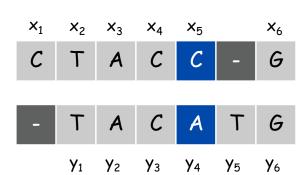
Def. An alignment M is a set of ordered pairs x_i - y_j such that each item occurs in at most one pair and no crossings.

Def. The pair $x_i - y_j$ and $x_{i'} - y_{j'}$ cross if i < i', but j > j'.

$$cost(M) = \underbrace{\sum_{(x_i, y_j) \in M} \alpha_{x_i y_j}}_{\text{mismatch}} + \underbrace{\sum_{i: x_i \text{ unmatched}} \delta + \sum_{j: y_j \text{ unmatched}} \delta}_{\text{gap}}$$

EX: CTACCG VS. TACATG.

Sol:
$$M = x_2 - y_1, x_3 - y_2, x_4 - y_3, x_5 - y_4, x_6 - y_6.$$



Sequence Alignment: Problem Structure

Def. OPT(i, j) = min cost of aligning strings $x_1 x_2 ... x_i$ and $y_1 y_2 ... y_j$.

- Case 1: OPT matches $x_i y_j$.
 - pay mismatch for x_i - y_j + min cost of aligning two strings $x_1 x_2 \dots x_{i-1}$ and $y_1 y_2 \dots y_{j-1}$
- Case 2a: OPT leaves x_i unmatched.
 - pay gap for x_i and min cost of aligning $x_1 x_2 \ldots x_{i-1}$ and $y_1 y_2 \ldots y_j$
- Case 2b: OPT leaves y_i unmatched.
 - pay gap for y_j and min cost of aligning $x_1 x_2 \dots x_i$ and $y_1 y_2 \dots y_{j-1}$

$$OPT(i, j) = \begin{cases} j\delta & \text{if } i = 0 \\ \alpha_{x_i y_j} + OPT(i-1, j-1) & \text{otherwise} \\ \delta + OPT(i, j-1) & \text{otherwise} \end{cases}$$

$$i\delta & \text{if } j = 0$$

Sequence Alignment: Algorithm

Analysis. $\Theta(mn)$ time and space.

English words or sentences: $m, n \le 10$.

Computational biology: m = n = 100,000. 10 billions ops OK, but 10GB array?