CS426: Introduction to Computational Biology

Section 2

Ends free alignement

- Why? Shotgun sequence assembly:
  - Large set of of partially overlapping subsequences that come from many copies of one original but unknown DNA sequence.
  - low global alignment score
  - high end – space free alignment score

Example

- Consider the sequences:
  - S = c a c t g t a c
  - T = g a c a c t t g
- Using the value 2 for match and –1 for indel/substitution the global alignment will have value 1:

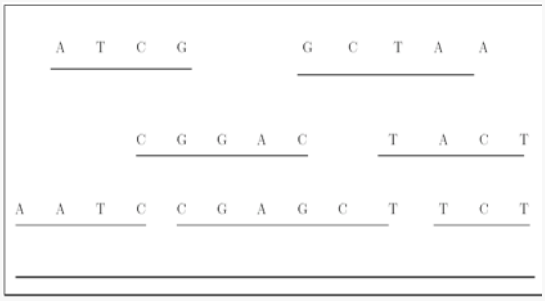
  S = c a c - - t - g t a c

  T = g a c a c t t g - - -
- The End-space free alignment will have value 9

  S = - - c a c - t g t a c

  T = g a c a c t t g - - -

Motivation



Case 1 suffix of S aligns with a prefix of T

- Find the best local alignment which includes both the last residue of S and the first residue of T.

Example

- `S = acc-gt--`
- `T = --cagtgc`

Algorithm

- We allow 0 weight to leading indel operations for string S;
- We fill in the values for V(i,j) using the recurrence relation from the global alignment algorithm
- We search for the maximal value in the last column thus allowing the S sequence to end before the other, with zero weight for all indel operations from there on. This value is the best value
- The aligned sequence is tracked from cell (0, 0) in the table until the end of S sequence (rightmost column) from there on, all indel operations until cell (*n,m*) are not counted in the total value (though they are present in the table).

Recursive definition

- Base condition: $V(0, j) = 0$

$$V(0,j) = \sum_{k=0}^{j} \sigma(-, T_k)$$

- Recursive relation:

$$V(i,j) = \max \begin{cases} V(i-1, j-1) + \sigma(S_i, T_j) \\ V(i-1, j) + \sigma(S_i, -) \\ V(i, j-1) + \sigma(-, T_j) \end{cases}$$

- Search for i* such that: $V(i^*, m) = \max_{1 \le i \le n, m} V(i, j)$
- The alignment score: $V(S, T) = V(i^*, m)$

Complexity

- Time Complexity: O(mn)
- Space Complexity : O(mn)/O(m+n) backtrack/no backtrack

Convex gap penalty model

- Each additional space in a gap contributes less to the gap weight than the previous space.
- This model is said to better describe biological behavior.
- Example: $Wg \log(q)$, where q is the length of the gap.
- The problem is solvable in $O(nm\log(m))$ time [1].