Contents

Consider two query amino acid sequences:

Gapless alignment compares two sequences:

$$Pr(S_i > x) \equiv \kappa exp(-\alpha x^{\gamma})$$

$Y_i \equiv I(S_i > x)$, where $I(S_i > x)$ denotes *Indicator Function*.
for ease set: $\gamma = 1$

Gumbel Distribution

Now Define

it is easy to show that,

which is the same as Gumbel distribution we know.

For random sequences, one can take $j = i$ without loss of generality. then we have:

$$S_{i,i} \equiv S(i) = max\{S(i-1) + s(i), 0\},$$

where the noise $s(i) \equiv s_{a_i,b_i}$.

Gapped Alignment

Suppose now deletion of length k are given weight W_k. We define matrix S with elements of $\{S_{i,j}\}$.
Preliminary values of S have the interpretation that $S_{i,j}$ is the maximum similarity of two segments *ending* in a_i and b_j , respectively. these values are obtained from the relationship:

$$S_{i,j} = max\{S_{i-1,j-1} + s(a_i, b_j), max_{k\geq 1}\{S_{i-k,j} - W_k\}, max_{l\geq 1}\{S_{i,j-l} - W_l\}, 0\}$$
for $1 \leq i \leq n$ and $1 \leq j \leq m$

Example

S matrix generated to the sequences $AAUGCCAUUGACGG$ and $CAGCCUCGCUUAG$. In this case a match $a_i = b_j$ produced an $s(a_i, b_j)$ value of unity while a mismatch produced a minus one-third. The delition weight used here was $W_k = 1 + 1/3 * k$.

In this simple example, the alignment obtained by:

$$\begin{array}{c} GCCAUUG \\ GCC-UCG \end{array}$$

Segment Identification

The pair of segments with maximum similarity is found by first locating the maximum elements of S. The other matrix elements leading to this maximum value are than sequentially determined with a traceback procedure ending with an element of S equal to zero.

Segment Identification

The pair of segments with maximum similarity is found by first locating the maximum elements of S. The other matrix elements leading to this maximum value are than sequentially determined with a traceback procedure ending with an element of S equal to zero.

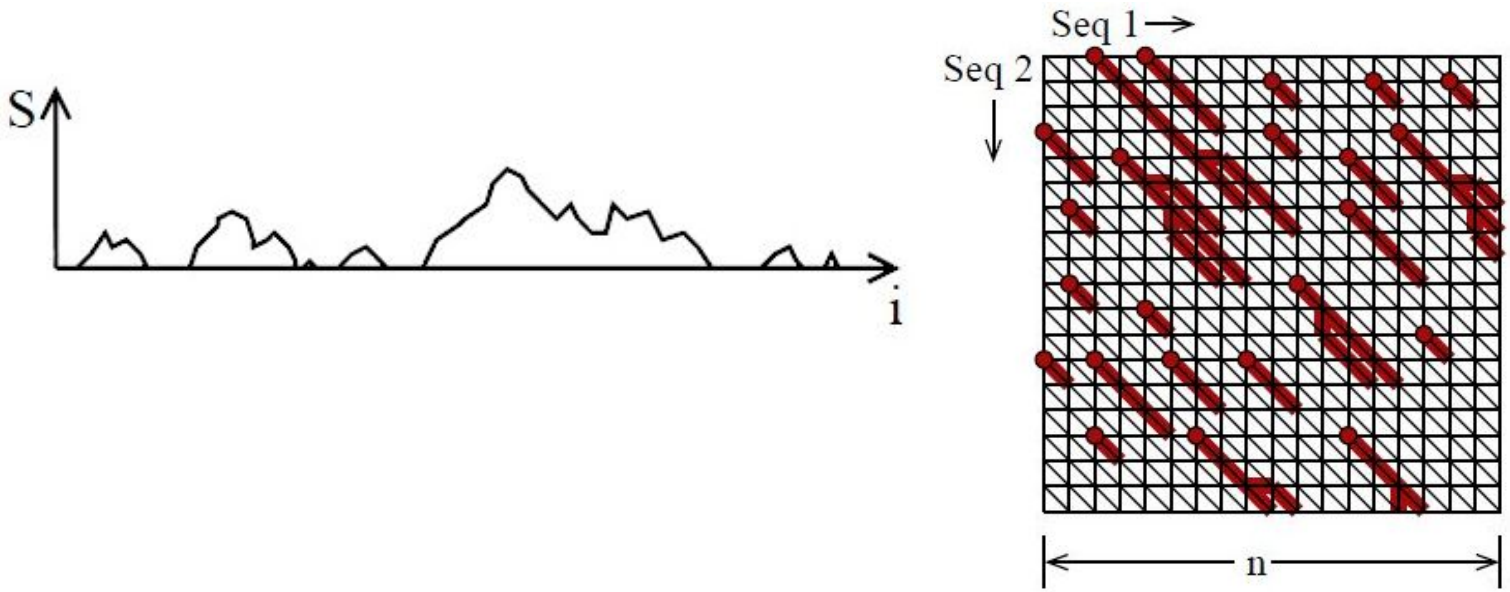
Other Segments

The pair of segments with the next best similarity is found by applying the traceback procedure to the second largest element of S not associated with the first traceback.

Introduction

The Smith-Waterman algorithm generates a score for each cell C in path graph, corresponding to the highest-scoring local alignment ending at C. This local alignment starts at a specific anchoring cell, and an *island* consists of all cells with identical anchor.

Figure: Left hand denoting gapless alignment and right hand denoting gapped alignment



P-value

and so

$$prob(S' = x') = prob(S' \leq x') - prob(S' \leq x' - 1)$$

If we have ρmn island scores and define each island score with S:

$$prob(S' = x') = prob(S' \leq x')^{\rho mn} - prob(S' \leq x' - 1)^{\rho mn}$$

Since each island score follows the geometric distribution we get:

$$prob(S' = x') = (1 - D\frac{p^{x'+1}}{1-p})^{\rho mn} - (1 - D\frac{p^{x'}}{1-p})^{\rho mn}$$

$$\approx exp(-Kmnp^{x'+1}) - exp(-Kmnp^{x'})$$

with $K = \frac{\rho D}{1-p}$.