Hello Class

STATISTICAL FOUNDATIONS

CS Foundations

Biology Foundations

Heuristic interpretation of the "alignment score"

\[ X = \text{AACGTAAT} \]
\[ Y = \text{ACGCGTGT} \]

scoring scheme

\[ \alpha \cdot \begin{cases} 1 & \text{match} \\ -1 & \text{gap (incl.)} \\ -2 & \text{mismatch} \end{cases} \]

\[ p = \text{prob. of a match} \]
\[ q = \text{prob. of a mismatch} \]
\[ r = \text{prob. of a single gap} \]

\[ \gamma = \text{probability of the alignment } A \]

\[ P = p^4 q^2 r^2 \]
Take the log likelihood
\[ s' = \log_2 \frac{P_y}{P_x} = \log_2 \frac{q^2}{p^2} = \]
\[ = 4 \log p + 2 \log q + 2 \log r \]

Consider
\[ s = s' - 8 \log k \quad k \text{ is picked such that } \log \left( \frac{p}{k} \right) = 1 \]

\[ s = s' - 8 \log k = \]
\[ 4 \log p + 2 \log q + 2 \log r - \]
\[ -4 \log k - 2 \log k - 2 \log k = \]
\[ = 4 \log \left( \frac{p}{k} \right) + 2 \log \left( \frac{q}{k} \right) + 2 \log \left( \frac{r}{k} \right) = \]
\[ = 1 - \frac{2u}{n} - \frac{2z}{n} = \]
\[ = 4 - 2u - 2z \geq \text{ MAX} \]

 Alignment of max score
\[ \equiv \text{maximizing the likelihood of the alignment} \]
GLOBAL ALIGNMENT

The Needleman-Wunsch Alg.

Edit Graph

Input: X, Y sequences
8 scoring scheme/matrix

Output: Alignment of max score

Edit Graph

\[ G_{\text{vertices}} = (M+1) \times (N+1) \]

\[ |X| = N \quad |Y| = M \]

represented as a rectangular grid = Matrix

\[ (i,j) \text{ coordinates} \]

\[ X = x_1 x_2 \ldots x_M \]

\[ Y = y_1 y_2 \ldots y_N \]

\[ x_{i,j} \in \Sigma \]

Edges

\[ 1 \leq i \leq M \]

\[ 0 \leq j \leq N \]

\[ (i-1,j) \rightarrow (i,j) \]

\[ (i,j) \rightarrow (i,j+1) \]
CONCLUSION: ONLY at most 3 DIRECT edges are coming into (i,j)

There is a one-to-one correspondence between paths between BEGIN and END nodes in the edit graph and alignments of the
two sequences

\[ n \]

\[ M = N \]

\[ \text{BEGIN} \]

\[ \text{END} \]

\[ c_i \text{ in } N \]

\[ d_i \]

\[ \geq 2^n \]

There are exponentially many alignment for two sequences of size \( N \)

\[ A \ C \ G \ T \ T \ T \]

\[ \text{BEGIN} \]

\[ (i) \]

\[ s(i,j) \]

\[ (i,j) \]
Cost of the optimal alignment from BEGIN to \((i,j)\)

\[ S = \text{matrix, DP matrix} \]

\[ \text{DP} = \text{Dynamic Programming} \]

\[ S(i,j) = \text{the score of the max score path from BEGIN to (i,j) ending in } S. \]

\[ S(i,j) = \max \left( S(i-1,j-1) + \delta(x_i, y_j) \right) \]
Global Alignment Algorithm

\[ S[0,0] = 0 \]

\[ \text{FOR} \ i = 1 \ \text{to} \ M \ \text{DO} \]

\[ S[i,0] = S[i-1,0] + \delta(z_{i,\cdot}) \]

\[ \text{FOR} \ j = 1 \ \text{to} \ N \ \text{DO} \]

\[ S[0,j] = S[0,j-1] + \delta(\cdot,y_{\cdot,j}) \]

\[ \text{FOR} \ i = 1 \ \text{to} \ M \ \text{DO} \]

\[ S[i,j] = \max \left\{ S[i-1,j-1] + \delta(z_{i,j}), \right. \]

\[ S[i-1,j] + \delta(z_{i,j}), \]

\[ S[i,j-1] + \delta(\cdot, y_{\cdot,j}) \]}

\[ \text{PRINT } (\text{MAX SCOR} = \text{S}[M,N]) \]

Phase 1: Compute max score for the alignment
Phase 2: Traceback: construct the actual optimal (max score) alignment

3 big differences between GLOBAL & LOCAL

1. initialization
2. "0" in the recurrence
3. scoring matrix vs. in a drug