Exercise 2

Problem 5

Idea: 3 matrices M, I, D for (partial) alignments ending with a (mis)match, insertion or deletion, respectively

- M: $M_{0,0} = 0$, $M_{i,0} = \infty$, $M_{0,j} = \infty$

$$M_{i,j} = p(s_i, T_j) + \min \begin{cases} M_{i-1,j-1} \\ I_{i-1,j-1} \\ D_{i-1,j-1} \end{cases}$$

- I: $I_{0,0} = \infty$, $I_{i,0} = \infty$, $I_{0,j} = \rho + j \cdot \sigma$

$$I_{i,j} = \sigma + \begin{cases} M_{i,j-1} + \rho \\ I_{i,j-1} \\ D_{i,j-1} + \rho \end{cases}$$

- D: $D_{0,0} = \infty$, $D_{i,0} = \rho + i \cdot \sigma$, $D_{0,j} = \infty$

$$D_{i,j} = \sigma + \min \begin{cases} M_{i-1,j} + \rho \\ I_{i-1,j} + \rho \\ D_{i-1,j} \end{cases}$$

Algorithm: for 'step' (i,j) we calculate $M_{i,j}$, $D_{i,j}$, $I_{i,j}$

$\rightsquigarrow$ runtime/space $\approx 3\times$ as before

Alignment: backtracking through all three matrices

Starting at the entry $\min(M_{n,m}, I_{n,m}, D_{n,m})$

Originally proposed: Gotoh (1982), An improved algorithm for matching biological sequences

Problem 6

(a) (M1) p a metric $\Longrightarrow p(a,b) \geq 0$

$g > 0$

} It follows directly that alignment score is non-negative

(M2) $\left.\begin{array}{l} p(a,a) = 0 \\ g > 0 \end{array}\right\} \Rightarrow sim(w,w) = 0$

$sim(u,v) = 0 \Rightarrow$ no gaps $(g > 0)$, no substitutions (p a metric, $p(a,b) > 0$, $a \neq b$)

(M3) replacements/gaps are symmetric

(M4)

Idea: Can construct alignment for (x,z) from optimal alignments for (x,y) and (y,z).

Observation: Scores are computed columnwise.

$\rightsquigarrow$ Restrict attention to single alignment columns.

Plan: "Shuffle" the alignments of (x,y) and (x,z) into each other by inserting gap-only columns into them,

+

Case distinction on a sequence of triples

$$\begin{pmatrix} a \\ b \\ c \end{pmatrix} \qquad a,b,c \in \Sigma' = \Sigma \cup \{-\}$$

Note: • $\delta(-,-) = 0$

• $\begin{pmatrix} - \\ - \\ - \end{pmatrix}$ cannot occur

• $\begin{pmatrix} - \\ b \\ - \end{pmatrix}$, $b \neq -$: $\rightsquigarrow$ transform $\begin{pmatrix} - \\ - \end{pmatrix}$ (remove afterwards)

$$\underbrace{\delta(-,-)}_{\text{contribution to } sim(x,z)} = 0 \leq 2g = \underbrace{\delta(-,b) + \delta(b,-)}_{\text{contribution to } sim(x,y) + sim(y,z)}$$

• $\begin{pmatrix} - \\ - \\ c \end{pmatrix}$, $c \neq -$: $\longrightarrow \begin{pmatrix} - \\ c \end{pmatrix} \rightsquigarrow \delta(-,c) = g \leq 0 + g$

$$= \delta(-,-) + \delta(-,c)$$

$\hookrightarrow \begin{pmatrix} a \\ - \\ - \end{pmatrix}$ is similar by symmetry

$\cdot \begin{pmatrix} - \\ b \\ c \end{pmatrix} \rightsquigarrow \begin{pmatrix} - \\ c \end{pmatrix} \rightsquigarrow \delta(-,c) = g \leq g + p(b,c) = \delta(-,b)$
$+ \delta(b,c)$

$\hookrightarrow \begin{pmatrix} a \\ b \\ - \end{pmatrix}$ is similar by symmetry

$(*) \cdot \begin{pmatrix} a \\ b \\ c \end{pmatrix}, a,b,c \in \Sigma: \begin{pmatrix} a \\ c \end{pmatrix} \to \delta(a,c) = p(a,c) \overset{\Delta\text{-inequality}}{\leq} p(a,b) + p(b,c)$
$= \delta(a,b) + \delta(b,c)$

$\cdot \begin{pmatrix} a \\ - \\ c \end{pmatrix}, a,c \in \Sigma: \rightsquigarrow \begin{pmatrix} a & - \\ - & c \end{pmatrix}$

$\rightsquigarrow \delta(a,-) + \delta(-,c) = 2g$

(b) Example: $\Sigma = \{a,b,c\}$

$g = \varepsilon < 1$

$$p = \begin{pmatrix} 0 & 1 & 3 \\ 1 & 0 & 1 \\ 3 & 1 & 0 \end{pmatrix}$$

(M1)-(M3) clear

(M4): $p(a,b) + p(b,c) = 2 < 3 = p(a,c)$ ↯ Δ-inequality

BUT: sim_δ is still a metric

$\hookrightarrow$ proof: Same as in (a), except for $(*)$

@ $(*)$: $\begin{pmatrix} a \\ b \\ c \end{pmatrix}, a,b,c \in \Sigma$

1) $b = a \lor b = c$

$$\begin{pmatrix} a \\ b \\ c \end{pmatrix} \rightsquigarrow \begin{pmatrix} a \\ c \end{pmatrix}$$

$$\rightsquigarrow \delta(a,c) \underset{\uparrow}{\overset{p \geq 0}{\leq}} \delta(a,b) + \delta(b,c)$$

$$\delta(a,c) \in \{\delta(a,b), \delta(b,c)\}$$

2) otherwise

$$\begin{pmatrix} a \\ b \\ c \end{pmatrix} \rightsquigarrow \begin{pmatrix} a & - \\ - & c \end{pmatrix}$$

$$\rightsquigarrow \delta(a,-) + \delta(-,c) = 2g \underset{\substack{\uparrow \\ \text{choice of} \\ g}}{<} 2 \underset{\substack{\uparrow \\ a \neq b \text{ and } b \neq c, \\ \text{choice of } p}}{\leq} \delta(a,b) + \delta(b,c)$$

(c) (M1) – (M3) remain valid

<u>(M4)</u>: similar, but new group columns such that consecutive inserts resp. deletes are considered as single, atomic operations

↳ new symbol $\vdash$ ≙ start of a gap

$$\begin{pmatrix} a \\ \vdash \end{pmatrix} \rightsquigarrow \text{cost } \rho + \sigma$$

$$\begin{pmatrix} a \\ - \end{pmatrix} \rightsquigarrow \text{cost } \sigma$$

Again start with optimal alignments for (x,y) and (y,z) and insert gap-only columns such that the y-lines coincide.

$\rightsquigarrow \begin{pmatrix} a \\ b \\ c \end{pmatrix}$, $a, b, c \in \Sigma'' = \{\vdash, -\} \cup \Sigma$

The following cases are handled by simply taking $\begin{pmatrix} a \\ c \end{pmatrix}$:

$\begin{pmatrix} a \\ b \\ a \end{pmatrix}, \begin{pmatrix} a \\ b \\ \vdash \end{pmatrix}, \begin{pmatrix} \vdash \\ b \\ c \end{pmatrix}, \begin{pmatrix} a \\ b \\ - \end{pmatrix}, \begin{pmatrix} - \\ b \\ c \end{pmatrix}, \begin{pmatrix} \vdash \\ \vdash \\ c \end{pmatrix}, \begin{pmatrix} a \\ \vdash \\ \vdash \end{pmatrix}, \begin{pmatrix} - \\ - \\ c \end{pmatrix}, \begin{pmatrix} a \\ - \\ - \end{pmatrix},$

$\begin{pmatrix} \vdash \\ b \\ \vdash \end{pmatrix}, \begin{pmatrix} - \\ b \\ - \end{pmatrix}, \begin{pmatrix} \vdash \\ b \\ - \end{pmatrix}, \begin{pmatrix} - \\ b \\ \vdash \end{pmatrix}$

Block of the form $\begin{pmatrix} u \\ - \\ w \end{pmatrix}$; u, w subwords of x, y where parts of u are deleted, parts of w are inserted.

↳ example: $\begin{pmatrix} a & a & \vdash & a \\ \vdash & - & - & - \\ \vdash & c & c & \vdash \end{pmatrix}$

$\begin{pmatrix} u \\ - \\ w \end{pmatrix} \rightsquigarrow \begin{pmatrix} u \\ \vdash \end{pmatrix}\begin{pmatrix} \vdash \\ w \end{pmatrix}$ simply check by old cases

See also: Waterman, Smith, Beyer (1976), Some biological metrics