Idee: Drei Matrizen:

Raphael
r_reitzi@cs....

- M — enden in (Mis)Match,
- I — enden in Insertion
- D — —"— Deletion

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

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

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

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

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

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

Algorithmus: wie bisher, aber berechne

im Schritt (i,j) $M_{i,j}, I_{i,j}, D_{i,j}$

→ Laufzeit = Faktor 3

Ergebnis: $\min\{M_{m,n}, I_{m,n}, D_{m,n}\}$

Erstmals beschrieben von

Gotoh: An improved algorithm for matching biological sequences. (1982)

Alignment wie gehabt durch Backtrac. (durch drei Tabellen)

6| a) (M1), (M2), (M3) klar.

ad (M4):

Gegeben optimale Alignments für (x,y), (y,z).

Wir konstruieren Alignment für (x,z) mit Kosten $\leq Sim(x,y) + Sim(y,z)$.

(gzz; der optimale Alignment ist höchstens besser.)

Beobachtung: Kosten werden spaltenweise berechnet.

↝ Idee: Alignments ineinander shufflen und "zusammenschieben", bis die Worte passen.

1) Füge in Alignments $(x,y)/(y,z)$ so wenig wie möglich Gap-Spalten ein, bis die y-Zeilen zusammenpassen.

wichtig: $\delta(-,-) = 0$

2) Kombinieren zu Spalten der Form

$$\begin{matrix} x \\ y \\ z \end{matrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} \in (\Sigma \cup \{-\})^3$$

wichtig: keine Spalte $\begin{pmatrix} - \\ - \\ - \end{pmatrix}$

3) Spaltenweise Δ-Ungl. nachweisen, also dass

„ $\delta(a,c) \leq \delta(a,b) + \delta(b,c)$ "

- $\begin{pmatrix} \bar{b} \\ - \end{pmatrix} \rightsquigarrow b \neq -$ $\rightsquigarrow \begin{pmatrix} - \\ - \end{pmatrix}$ (und dann weg)

$\rightsquigarrow \underbrace{\delta(-,-) = 0}_{\text{Beitrag zu } \mathrm{Sim}(x,z)} \leq 2g = \delta(-,b) \underbrace{+\delta(b,-)}_{\text{Beitrag zu } \mathrm{sim}(x,y)+\mathrm{sim}(y,z)}$

- $\begin{pmatrix} \bar{-} \\ c \end{pmatrix} \rightsquigarrow \begin{pmatrix} - \\ c \end{pmatrix}$ ($\begin{pmatrix} g \\ - \end{pmatrix}$ symmetrisch)

$\rightsquigarrow \delta(-,c) = g \leq 0 + g = \delta(-,-) + \delta(-,c)$

- $\begin{pmatrix} \bar{b} \\ c \end{pmatrix} \rightsquigarrow \begin{pmatrix} - \\ c \end{pmatrix}$ ($\begin{pmatrix} a \\ b \end{pmatrix}$ symmetrisch)

$\rightsquigarrow \delta(-,c) = g \overset{p \geq 0}{\leq} g + p(b,c) = \delta(-,b) + \delta(b,c)$

(*) - $\begin{pmatrix} a \\ b \\ c \end{pmatrix} \rightsquigarrow \begin{pmatrix} a \\ c \end{pmatrix}$

$\rightsquigarrow \delta(a,c) = p(a,c) \overset{\Delta\text{-Ungl. für } p}{\leq} p(a,b) + p(b,c)$

$= \delta(a,b) + \delta(b,c)$

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

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

Kosten vor- wie nachher. !!

b) Vorschlag: Nehme "Fastmetrik" □

(↝ Beweis wie in a) + ε)

$\Sigma = \{a,b,c\}$

$0 < g < 1$

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

Klar: (M1-3) erfüllt.

Aber:

$p(a,b) + p(b,c) = 2$

$p(a,c) = 3$

↝ Δ-Ungl. verletzt.

Beweis, dass Metrik, wie in a) außer für (*); sonst brauchten wir Δ-Ungl. nicht.

ad (*)

$\binom{A}{B}{C}$ $A,B,C \neq -$ $\rightsquigarrow$ $\begin{cases} \binom{A}{-}\binom{-}{C} & \text{, sonst} \quad i) \\ \binom{A}{C} & \text{, } B \in \{A,C\} \quad ii) \end{cases}$

$\rightsquigarrow$ i) $\delta(A,-) + \delta(-,C)$

$= 2g < 2 \overset{i)}{\leq} \delta(A,B) + \delta(B,C)$

ii) $\delta(A,C) \overset{p \geq 0}{\leq} \delta(A,B) + \delta(B,C)$

$\in \{\delta(A,B), \delta(B,C)\}$

□

c) (MA-3) — s.o.

(MA) ähnlich, können aber nicht alles beweisen.

Neues Symbol: $\vdash$ - Start eines Gap

$\rightsquigarrow$ Schreiben $\binom{a}{+}$ → cost $p+g$ für die ersten (linksten) Gaps, $\binom{a}{-}$ sonst. (insertions analog.)

↳ cost g

⚠ Werden $\vdash$ zu $-$ und umgekehrt?

- Behandle viele Fälle wie oben; genauer: $\rightsquigarrow \binom{a}{c}$ (Δ-Ungl. gilt)

für:

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

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

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

- Übrig bleiben Blöcke der Form

z.B.
$\begin{pmatrix} a & a & \vdash & a \\ - & - & - & - \\ \vdash & c & c & \vdash \end{pmatrix}$

$$\begin{pmatrix} u \\ - \\ w \end{pmatrix}, \quad u, w$$

$$\rightsquigarrow \begin{pmatrix} u & \vdash \\ \vdash & w \end{pmatrix}$$

, dann passt es.

□

Siehe: Waterman, Smith, Beyer (1976)
Some biological sequence metrics.

d) (M2) geht kaputt:

- für semi-global (x, xy) bzw. (x, yx)

- für local (xyz, y)

haben Score 0, aber die Worte sind i.A. nicht gleich.

(Ann: $\delta(a,a) = 0$ f.a. a; sonst ist aber $\mathrm{sim}(x,x) \neq 0 \rightsquigarrow \not\delta(\mathbb{Z})$)