# Problemseminar

Multiple alignment by sequence annealing

Ariel S. Schwartz and Lior Pachter

# Gliederung

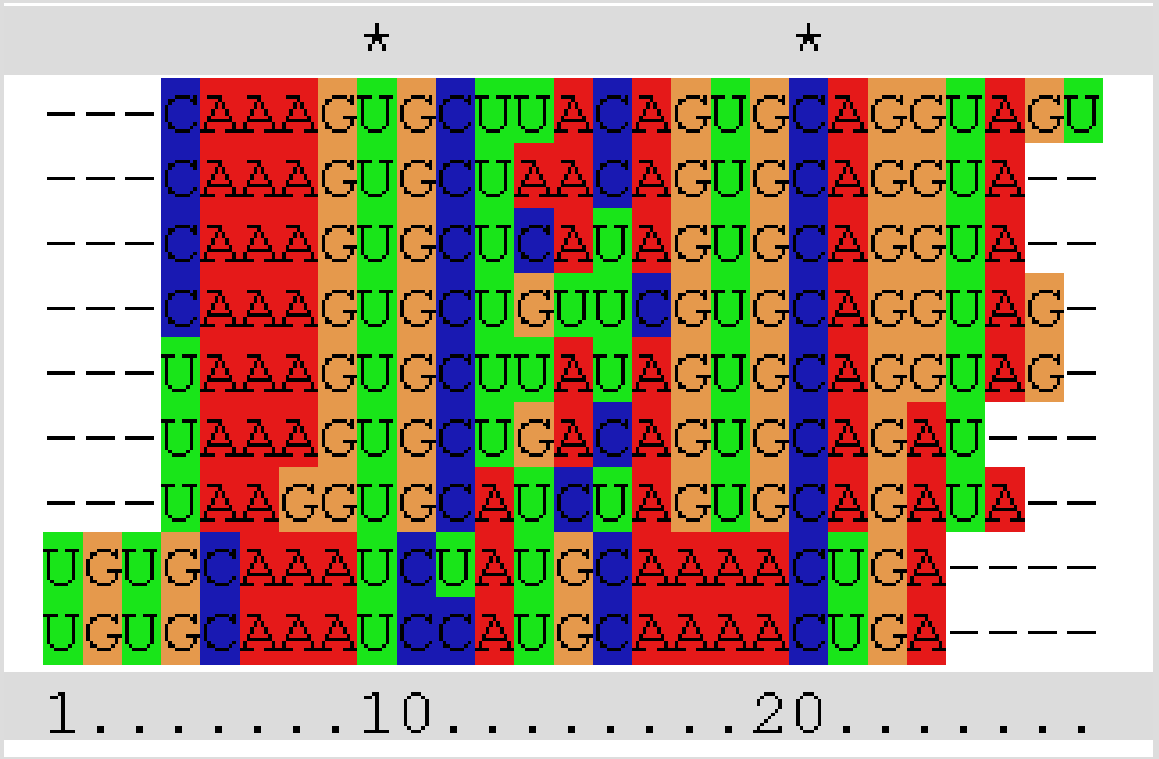
1. Einführung
2. Alignment Posets
3. Sequence Annealing
4. Results

## Progressives Alignment

- Multiples Alignment durch paarweises Alignment

# 1. Einführung

## 2. Partial order graphs



# Partial order graphs

# Alignment-Problem:

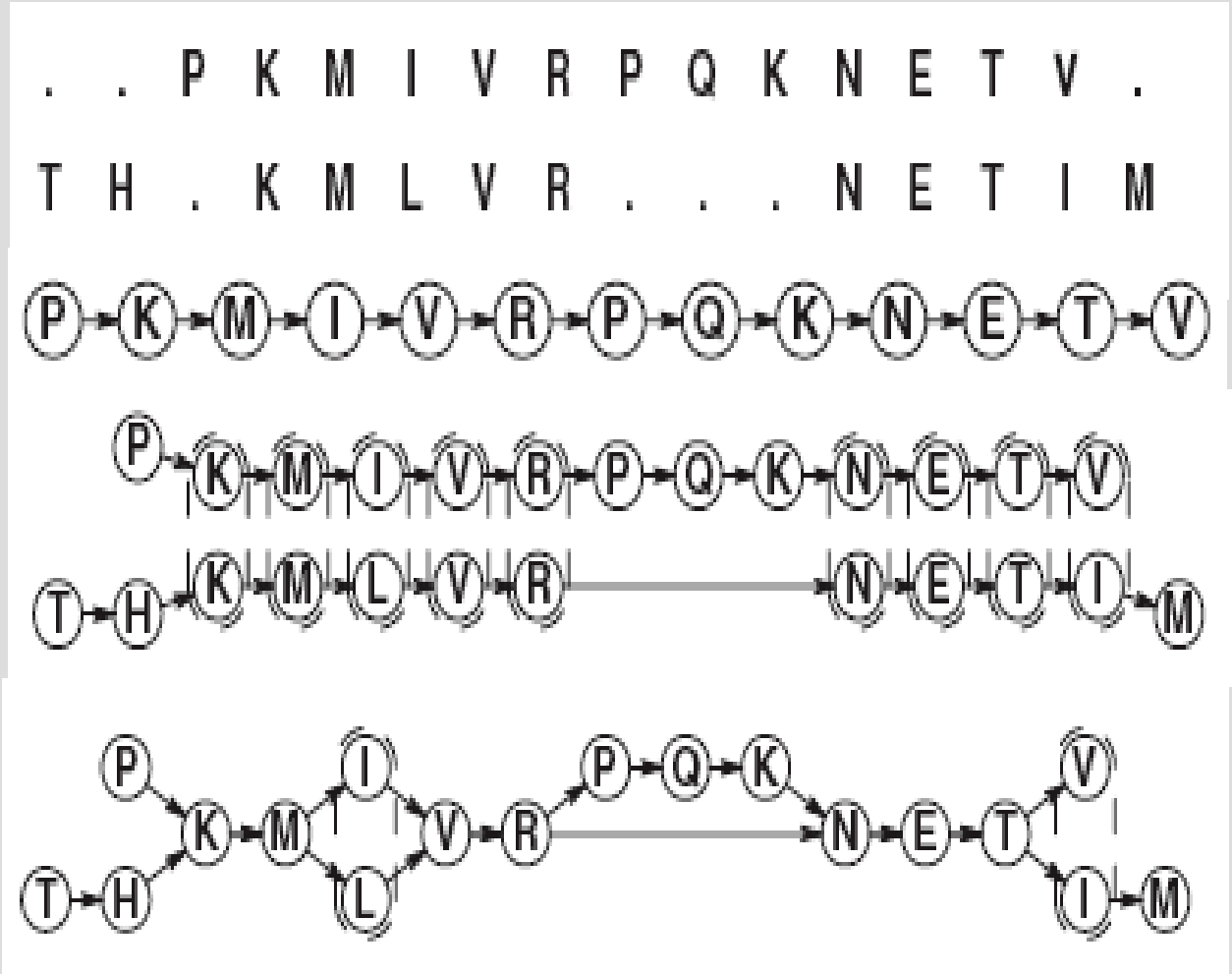
```
s1:          .....ACATGTCGAT.....AGGTG
s2:          TGCAC.....TCGATACATAAGGTG

Consensus1: TGCACACATGTCGATACATAAGGTG

Consensus2: ACATGTGCACTCGATACATAAGGTG
...
```

Neue Forderungen:

- Alignment durch paarweises dynamisches Alignment möglich (wie bisher auch bei Konsensussequenz)
- Kein Informationsverlust gegenüber dem MSA (Multiple Sequence Alignment)

- Welche Sequenzpositionen sind zueinander aligned?
- Ordnung dieser Positionen in den Sequenzen?

→ Partial Order Graph


. . P K M I V R P Q K N E T V .
T H . K M L V R . . . N E T I M

- Welche Sequenzpositionen sind zueinander aligned? ✔
- Ordnung dieser Positionen in den Sequenzen? ✔

# Sequence annealing

# Ausgangsposition:

- *k* Sequenzen der Länge $n_1$, $n_2$, ... , $n_k$
- $\sigma_i^a$ - i-te Position in der Sequenz a
- $L = \sum_{(i)} n_i$

- $M_{\text{null}}$ – Null-Alignment
- Länge $L$
- Bsp:

```
ABCD  ----  ----
----  EFGH  ----
----  ----  IJKL
```

# Idee

- Alle möglichen Paare bekommen Gewichte (posterior Wahrscheinlichkeiten für matches und gaps)
- Positionen, die am wahrscheinlichsten homolog sind, werden zuerst aligned

# Zielfunktion

- Exaktheit des multiplen Sequence Alignment muss bestimmt werden
- *Developer score ($f_D$) (~sum-of-pair-score)*
- *Alignment metric accuracy (AMA)*
- Zielfunktion, um sowohl $f_D$ als auch *AMA* zu optimieren

- probabilistisches Modell muss gegeben sein

- *Match posterior probability*

$$P(\sigma_i^1 \Diamond \sigma_j^2 | \sigma^1, \sigma^2, \theta)$$

- *Gap posterior probability*

$$P(\sigma_i^1 \Diamond - | \sigma^1, \sigma^2, \theta)$$

$$P(- \Diamond \sigma_j^2 | \sigma^1, \sigma^2, \theta)$$

- Gap-Faktor: $G_f$

# Resultierende Zielfunktion *f*

$$f^{G_f}(M) = \sum_{\sigma^a, \sigma^b \, mit \, a \neq b} \Bigg( \sum_{\{(j,k) | \varphi^M(\sigma_j^a) = \varphi^M(\sigma_k^b)\}} P(\sigma_j^a \Diamond \sigma_k^b | \sigma^a, \sigma^b, \theta)$$

$$+ G_f \cdot \sum_{\{j | \forall \sigma_k^b \varphi^M(\sigma_j^a) \neq \varphi^M(\sigma_k^b)\}} P(\sigma_i^a \Diamond \dashv | \sigma^a, \sigma^b, \theta)$$

$$+ G_f \cdot \sum_{\{k | \forall \sigma_j^a \varphi^M(\sigma_j^a) \neq \varphi^M(\sigma_k^b)\}} P(\vdash \Diamond \sigma_k^b | \sigma^a, \sigma^b, \theta) \Bigg)$$

$G_f=0$: $f^0$ berechnet $f_D$-*score*

$G_f=0.5$: $f^{0.5}$ berechnet *AMA-score*

je größer $G_f$, umso höher die Spezifität

je kleiner $G_f$, umso höher die Sensitivität

Gewichtsfunktion – $w(p)$

- jedem Paar $p=(c_k, c_i)$ soll ein Gewicht zugeordnet werden
- $f(M_{i-1}) \geqslant f(M_i)$ bei positiven Gewichten

- Aus *f* werden die Gewichtsfunktionen berechnet
  - statische vs. dynamische Version

- $P_{match}(c_k, c_l)$ - Wahrscheinlichkeit, dass $c_k$ und $c_l$ matchen
- $P_{gap}(c_k, c_l)$ – Wahrscheinlichkeit, dass $c_k$ und $c_l$ eine Lücke bilden
- Gap-Faktor $G_f$

1.Gewichtsfunktion: maxstep:

$$w_{maxstep}^{G_f}(c_k, c_l) = \frac{P_{match} - G_f \cdot P_{gap}}{|\varphi^{-1}(c_k)\varphi^{-1}(c_l)|}, falls\, c_k \neq c_l, sonst -\infty$$

# 2. Gewichtsfunktion: tgf

$$w_{tgf}^{G_f}(c_k, c_l) = \frac{P_{match}}{P_{gap}} - G_f, \; falls \; c_k \neq c_l, \; sonst \; -\infty$$

$$temp = w_{tgf}^{G_f}(p_{max}) + G_f$$

$$f^{temp}(M)$$

→dynamische Gewichte

# Algorithmus

1) $M_i = M_{Null}$
2) **while** ∃ Paar $p$ mit $w(p)>0$
   ***do*** Merge $p_{max}$ zu $M_{i-1}$
   $i$--
3) **end while**

## 4. Results

- AMAP 2.0 - http://bio.math.berkeley.edu/amap/
- Laufzeit vergleichbar mit anderen Programmen (Dialign-T, MUSCLE, T-Coffee)
- Methoden vermutlich auch für DNA anwendbar

## 4. Results

- Benchmark-Test auf Proteinsequenzen
- 3 verschiedene scores wurden berechnet (Spezifität, Sensibilität, Mix)
- In fast allen Tests bestes Programm
- Zu jedem Zeitpunkt des Algorithmus' kann ein Alignment ausgegeben werden(angefangen von sehr spezifischen Alignments mit geringer Sensibilität bis hin zu sehr sensiblen Alignments mit geringerer Spezifität

# Ende