Multiple alignment by sequence annealing

Alignment Poset

σ_i^a ... *i-te* Element einer Sequenz $\sigma^a = \sigma_1^a, ..., \sigma_n^a$ der Länge *n*

$S = \{\sigma^1, ..., \sigma^k\}$ ... Menge der $n_1 + n_2 ... + n_k$ Sequenzbuchstaben der Sequenzen $\sigma^1, \sigma^2, ..., \sigma^k$ der Längen $n_1, n_2, ..., n_k$

$P = \{c_1, ..., c_m\}$ ... partiell geordnete Menge der Sequenzbuchstaben *S* zusammen mit der surjektiven Funktion $\varphi : S \to P$, so dass $\varphi(\sigma_i^a) < \varphi(\sigma_j^a)$, falls $i<j$

→ *partielles globales multiples Alignment*

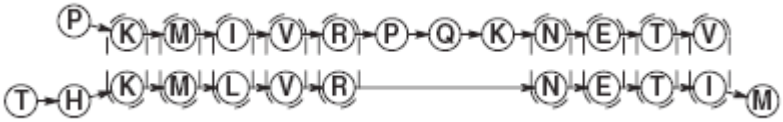
Die Elemente von *P* korrespondieren also zu den Spalten des multiplen Alignments und die partielle Ordnung gibt die Reihenfolge an, in der die Spalten auftreten müssen.

Beispiel:

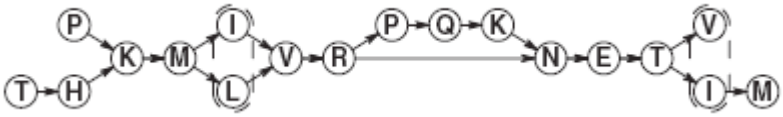
```
.  .  P  K  M  I  V  R  P  Q  K  N  E  T  V  .
T  H  .  K  M  L  V  R  .  .  .  N  E  T  I  M
```

2 Sequenzen

Sequenz 1 als gerichteter azyklischer Graph



Markierung der Alignment-Knoten



PO-MSA Repräsentation eines paarweisen Protein-Sequenz-Alignments

Scoring-Funktion für multiple Alignments

M ... Menge aller partiellen multiplen Alignments einer Menge Sequenzbuchstaben *S*

$f : M \to \mathbb{R}$ ... Funktion, die jedem multiplen Alignment einen „score“ zuordnet

f_D*(Developer score)* – equivalent zum sum-of-pairs score
Verhältnis der Anzahl der korrekt aligneden Paare im Alignment zur Anzahl der aligneden Paare im Referenzalignment

alignment metric accuracy (AMA) ... Verhältnis von Resten, die korrekt zu einem anderen Rest oder zu einer Gap aligned sind (Schwartz et al, 2006)
h^p – predicted Alignment
h^r – reference Alignment
$h^i, h^j \in A_{n1, n2, ..., nk}$ - MSAs mit k Sequenzen der Längen n_1 bis n_k

$$g(h^p, h^r) = 1 - \frac{d(h^p, h^r)}{(k-1)\sum_{i=1}^{k} n_i} \quad \text{mit} \quad d(h^i, h^j) = \sum_{s^1=1}^{k-1} \sum_{s^2>s^1}^{k} d(h^i_{s^1, s^2}, h^j_{s^1, s^2})$$

$$d(h^i, h^j) = 2|h^i_H| + |h^i_I| + |h^i_D| - 2|h^i_H \cap h^j_H| - |h^i_I \cap h^j_I| - |h^i_D \cap h^j_D|$$
$$= 2|h^j_H| + |h^j_I| + |h^j_D| - 2|h^i_H \cap h^j_H| - |h^i_I \cap h^j_I| - |h^i_D \cap h^j_D|$$
$$= n + m - 2|h^i_H \cap h^j_H| - |h^i_I \cap h^j_I| - |h^i_D \cap h^j_D|$$

$h_H = \{(i, j) : (\sigma^1_i \Diamond \sigma^2_j) \in h\}$ (*Align*)

$h_D = \{i : (\sigma^1_i \Diamond -) \in h\}$ (*Deletion*)

$h_I = \{j : (sigma^2_j \Diamond -) \in h\}$ (*Insertion*)

Sequence Annealing

Ziel: $\text{argmax}_{M \in \mathcal{M}}(f(M))$ finden (wobei M über alle partiellen multiplen Alignments geht)

$L = \sum_i n_i$ ... Länge aller Sequenzen

M_{null} ... Null globales multiples Alignment von k Sequenzen der *Längen* $n_1,...,n_k$ (disjunkte Vereinigung von k Ketten)

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

Wahrscheinlichkeiten:

- probabilistisches Modell muss gegeben sein
- *Match posterior probability:* $P(\sigma^1_i \Diamond \sigma^2_j | \sigma^1, \sigma^2, \theta)$
- *Gap posterior probability:* $P(\sigma^1_i \Diamond - | \sigma^1, \sigma^2, \theta)$
  $P(- \Diamond \sigma^2_j | \sigma^1, \sigma^2, \theta)$
- Gap-Faktor: G_f
- Resultierende Zielfunktion *f:*

$$f^{G_f}(M) = \sum_{\sigma^a, \sigma^b \, mit \, a \neq b} \Big(\sum_{\{(j,k) | \varphi^M(\sigma^a_j) = \varphi^M(\sigma^b_k)\}} P(\sigma^a_j \Diamond \sigma^b_k | \sigma^a, \sigma^b, \theta) + G_f \cdot \sum_{\{j | \forall \sigma^b_k \varphi^M(\sigma^a_j) \neq \varphi^M(\sigma^b_k)\}} P(\sigma^a_i \Diamond - | \sigma^a, \sigma^b, \theta) + G_f \cdot \sum_{\{k | \forall \sigma^a_j \varphi^M(\sigma^a_j) \neq \varphi^M(\sigma^b_k)\}} P(- \Diamond \sigma^b_k | \sigma^a, \sigma^b, \theta) \Big)$$

- G_f=0: f^0 berechnet f_D-score, G_f=0.5: $f^{0.5}$ berechnet AMA-score
- allgemein: 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
  - bei positiven Gewichten soll im Falle des Alignments dieses Paares das entstehende multiple Alignment einen höheren score haben als vorher: $f(M_{i-1}) \geq f(M_i)$
  - Gewichtsfunktionen werden aus f berechnet
    - statisch: einmal am Anfang des Algorithmus' berechnet
    - dynamisch: für jedes Kandidatenpaar p neu berechnet
  - das Paar mit dem höchsten Gewicht wird jeweils gemerged
  - $P_{match}(c_k, c_l)$ - Wahrscheinlichkeit, dass c_k und c_l matchen
    - $P_{match} = \sum_{\{\sigma^a_i \in \varphi^{-1}(c_k)\}} \sum_{\{\sigma^b_j \in \varphi^{-1}(c_l)\}} P(\sigma^a_i \Diamond \sigma^b_j | \sigma^a, \sigma^b, \theta)$
  - $P_{gap}(c_k, c_l)$ – Wahrscheinlichkeit, dass c_k und c_l eine Lücke bilden
    - $P_{gap} = \sum_{\{\sigma^a_i \in \varphi^{-1}(c_k)\}} P(\sigma^a_i \Diamond - | \sigma^a, \sigma^b, \theta) + \sum_{\{\sigma^b_j \in \varphi^{-1}(c_l)\}} P(- \Diamond \sigma^b_j | \sigma^a, \sigma^b, \theta)$

- Gewichtsfunktion: maxstep
  - $w_{maxstep}^{G_f}(c_k, c_l) = \frac{P_{match} - G_f \cdot P_{gap}}{|\varphi^{-1}(c_k)\varphi^{-1}(c_l)|}$, $falls\ c_k \neq c_l, sonst -\infty$
  - Anstieg in der Scoring-Funktion f^{G_f} soll maximal sein

- Gewichtsfunktion: *tgf*
  - $w_{tgf}^{G_f}(c_k, c_l) = \frac{P_{match}}{P_{gap}} - G_f$, $falls\ c_k \neq c_l, sonst -\infty$
  - $temp = w_{tgf}^{G_f}(p_{max}) + G_f$
  - scoring-Funktion $f^{temp}(M)$ kann neu berechnet werden
  - Gewichte werden somit dynamisch neu berechnet

Algorithmus:

1. $M_L \leftarrow M_{null}$
2. $i \leftarrow L$
3. ***while*** $\exists c_k^{M_i}, c_l^{M_i}$ *, so dass* $c_k^{M_i}$ *und* $c_l^{M_i}$ *zu M' gemerged wird und* $f(M') \geqslant f(M_i)$ ***do***
4. $M_{i-1} \leftarrow M'$
5. $i \leftarrow i\text{-}1$
6. ***end while***

vereinfacht:
„Solange es in dem aktuellen partiellen multiplen Alignment noch ein Buchstabenpaar gibt, dass den score des Alignments erhöhen würde, merge diese.“

Programm: http://baboon.math.berkeley.edu/amap/

Quellen

A.S. Schwartz and L. Pachter, *Multiple Alignment by Sequence Annealing*, Bioinformatics 23 (2007), e24–e29.

A. S. Schwartz, E. W. Myers and L. Pachter, *Alignment Metric Accuracy*, arXiv:q-bio/0510052