1 Needleman wunsch

Alignment of sequences s1 and s2.

$$A(i,j) = opt \begin{cases} A(i-1,j-1) + \sigma(i,j) \\ A(i-1,j) + g \\ A(j,i-1) + g \end{cases} \tag{1}$$

mit den gapkosten g und

$$\sigma(i,j) = \begin{cases} \text{match-score if } s1(i) = s2(j) \\ \text{mismatch-score else (2)} \end{cases}$$

2 Semiglobales alignment

Semiglobales alignmnt: Kurze Sequenz in langer suchen
Endgaps der kurzen Sequenz kostenlos

3 Smith waterman alignment

Suckt lokale Übereinstimmungen

$$A(i,j) = \min \begin{cases} A(i-1,j-1) + \sigma(i,j) \\ A(i-1,j) + g \\ A(j,i-1) + g \\ 0 \end{cases} \tag{2}$$

3.1 Subotimale alignments

- Nach erstem SW backtrack
- Bei Backtrack besuchten Zellen $u$nveränderlich auf 0
- Matrix neu befüllen
- Backtrack gibt neues Ergebnis
- Iterieren

4 Affine Gapkosten (Gotoh algorithmus)

- Wenige, grössere Indels wahrscheinlicher als viele, kleinere
- Gapkosten längenabhängig
- Kosten für direkten Einbau in SW oder NW zu hoch
- Verwenden von 2 verschiedenen gapkosten:

- g_o gap open penalty
- g_e gap extension penalty

$$
\begin{aligned}
A(i,j) &= \max \begin{cases} A(i-1,j-1) + \sigma\left(S_1(i), S_2(j)\right) \text{ (mis)match} \\ R(i,j) \text{ gap} \\ L(i,j) \text{ gap} \end{cases} \\
R(i,j) &= \max \begin{cases} A(i-1,j) + g_o \\ R(i-1,j) + g_e \end{cases} \\
L(i,j) &= \max \begin{cases} A(i,j-1) + g_o \\ L(i,j-1) + g_e \end{cases}
\end{aligned}
$$

5 Scoring Matritzen

Für Proteine/Aminosäuren

5.1 PAM (percent accepted mutation)

- Aus gapfreien Alignments mit mind. 85% seq. identität
- 20x20 Matrix, mit der Anzahl der Mutationsevents
- Normalisiere auf eine Mutation pro 100 Basen entsteht $M1$
- Weitere Mn-Matritzen durch M_1^n

$$
\begin{aligned}
M(i,j) &= \frac{\lambda A(i,j)}{N f(j)} && \text{(-1)} \\
M(j,j) &= 1 - \sum_{1 \le i \le 20. i \ne j} M(i,j) && \text{(0)}
\end{aligned}
$$

$$
\begin{aligned}
\texttt{PAM}_n(i,j) &= \log \frac{f(j) M_n(i,j)}{f(i) f(j)} && (1) \\
&= \log \frac{M_1^n(i,j)}{f(i)} && (2)
\end{aligned}
$$

Je hoeher die Nummer, desto unterschiedlicher sind die Sequenzen

5.2 BLOSUM

- Aus Alignments von Proteinen: gap freie konservierte Blöcke
- Zählen der Mutationsraten in diesen Blöcken.

- Vorher: Clustern von Sequenzen mit Ähnlichkeit X in einem Cluster
- Clustersequenzen gewichtet wie eine Sequenz
- Alle Sequenzen zwischen den Blöcken paarweise vergleichen
- Häufigkeiten der Paare AS ermitteln
- Nicht innerhalb der Blöcke
- Dann: Log odds score:$s(i,j) = log_2(\frac{q(ij)}{2p_ip_j})$
- Je hoeher nummer, desto ähnlicher Sequenzen
- (Häufig: BLOSUM62)

6 MSA

- DP Lösung ist exponentiell in Anzahl der Sequenzen
- Heuristiken werden benötigt
- Sum of pairs score

Heuristiken finden schnell eine Lösung, diese muss aber nicht die Beste sein.

6.1 Progressives Aligment

- Aus PWA-Distanzen Leit-baum
- Entlang diese Baumes wird alignt
- Diverse Varianten, wie Sequenzen und Alignments hinzugefügt werden
  - Profile Alignment
  - Alignment über guide Sequenzen