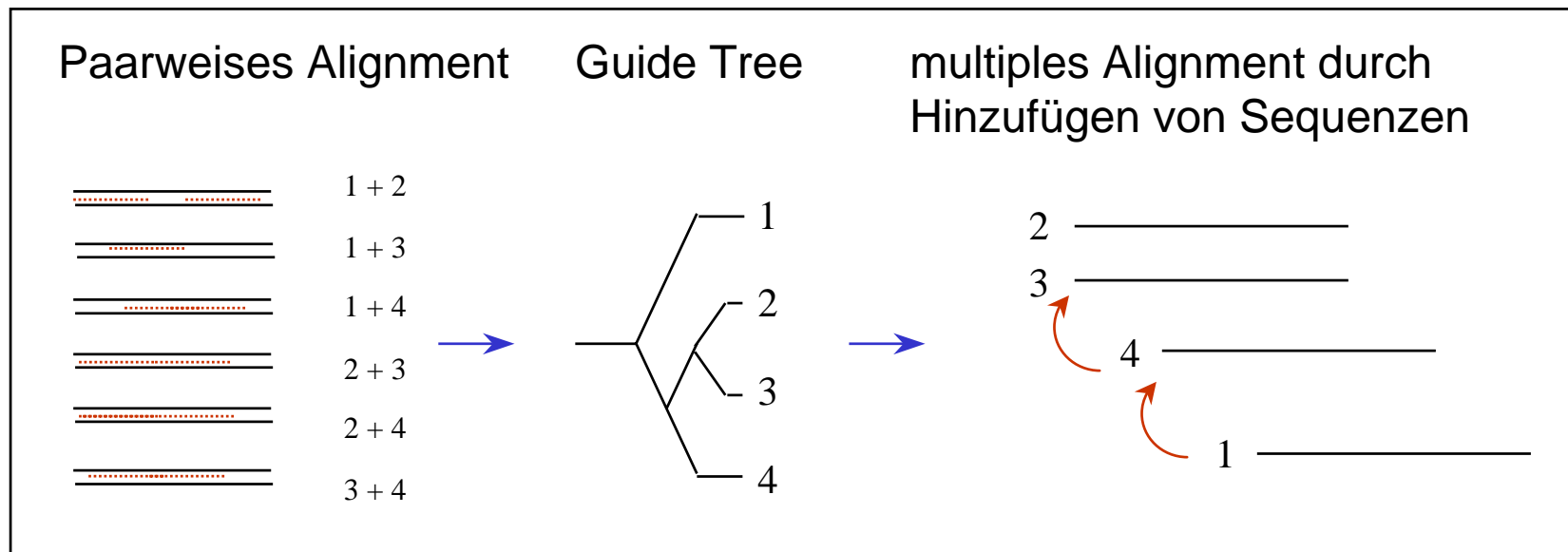


Methoden zum Multiplen-Sequenz-Alignment

- Multidimensionales dynamisches Programmieren
(MSA, Lipman 1988, DCA, Jens Stoye)
- **Progressive Alignments**
(Clustalw, Higgins 1996; PileUp, Genetics Computer Group (GCG))
- Lokale Alignments
(e.g. DiAlign, Morgenstern 1996; viele Andere)

Progressives Alignment (z.B. ClustalW)

Prinzip:



1

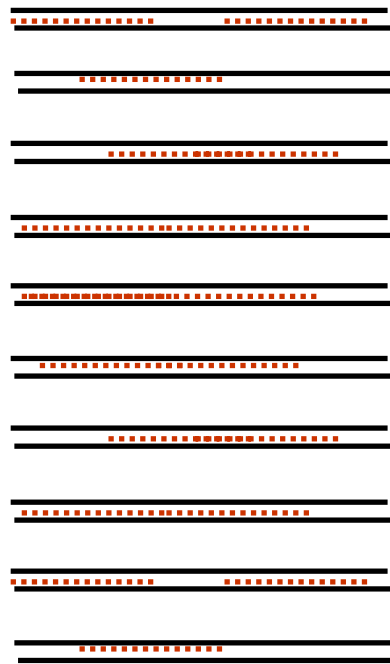


2

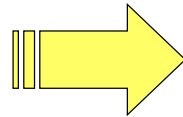


3

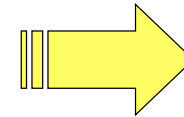
Pairweiser Vergleich aller Sequenzen



- 1 : 2
- 1 : 3
- 1 : 4
- 1 : 5
- 2 : 3
- 2 : 4
- 2 : 5
- 3 : 4
- 3 : 5
- 4 : 5



Ähnlichkeits-
Score von
jedem Paar,
der "Score"
(vorherige Vorlesung)



Distanz-Score
von jedem Paar

Berechnen des Distanz-Scores zweier Sequenzen

einfachste Methode: "Hamming Distanz"

Zählen der Mismatches:

Seq A =	T	A	T	T	C	G
Seq B =	T	G	C	T	G	T,

ergibt Distanz-Score = 4

Berechnen des Distanz-Scores zweier Sequenzen

etwas komplexer: Ähnlichkeitsscores ("normale" Scorewerte) werden in Distanz-Werte umgerechnet, z.B. (Feng & Doolittle 1996):

normalisieren:

$S_{\text{real}} =$ normaler Score-Wert von A und B

$S_{\text{ident}} =$ Mittelwert der Scores von A mit sich selbst und B mit sich selbst

$S_{\text{rand}} =$ Mittelwert der Scores von ~1000 randomisierter Sequenzen A und B

$$\Rightarrow S_{\text{norm}} = \frac{S_{\text{real}} - S_{\text{rand}}}{S_{\text{ident}} - S_{\text{rand}}} \Rightarrow \text{Distanz}_{AB} = -\log S_{\text{norm}}$$

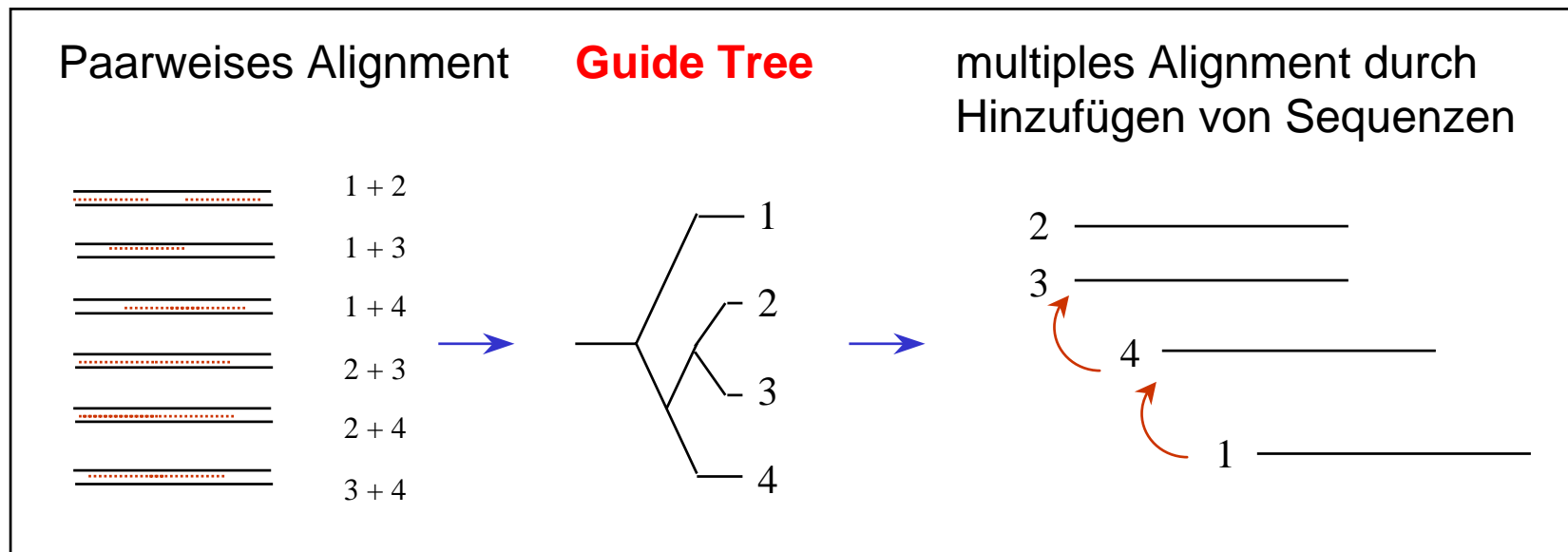
keine Ähnlichkeit: $S_{\text{norm}} = 0 \Rightarrow \text{Distanz} = \infty$

identisch: $S_{\text{norm}} = 1 \Rightarrow \text{Distanz} = 0$

Distanz-Matrix

Sequenz	1	2	3	4	5
1	<p><u>Distanz-Matrix:</u></p> <p>enthält Distanzen zu allen Sequenz- Paaren</p>				
2					
3					
4					
5					

Progressives Alignment



1



2

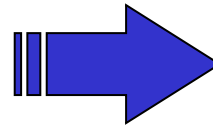


3

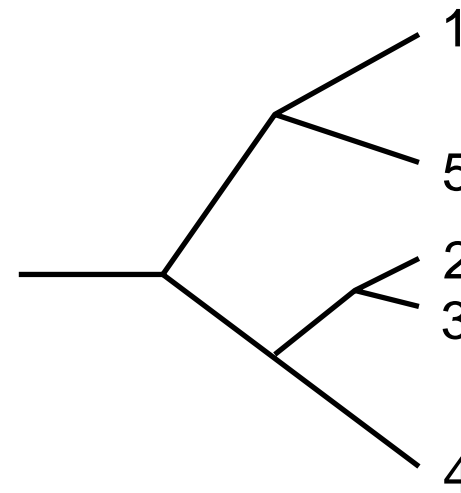
Konstruktion des Guide-Trees

kein phylogenetischer Baum !

Distanz-
Matrix



Guide Tree



es wird geclustert:

UPGMA (unweighted pair group method of arithmetic averages),

anderes Verfahren: Neighbour-Joining

Hierarchisches Clustern mit UPGMA

unweighed pairwise group method of arithmetic averages

Hierarchisches Clustern mit UPGMA

Prinzip:

- ermittle kleinsten Wert in der Distanz-Matrix
- Bilde einen Cluster (Gruppe) der entsprechenden Einträge (hier, $u = \{1,2\}$).
- berechne den Abstand dieses Clusters zu den restlichen Einträgen, durch Mitteln der Abstände der N geclusterten Einträge in dem Cluster zu den restlichen Einträgen

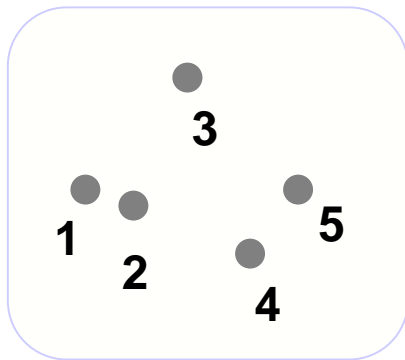
d_{ij}	1	2	3	4	5
1	0	2	6	9	7
2	2	0	5	7	7
3	6	5	0	5	4
4	9	7	5	0	3
5	7	7	4	3	0



d_{ij}	u	3	4	5
u	0	5.5	8	7
3	5.5	0	5	4
4	8	5	0	3
5	7	4	3	0

- Wiederhole das solange, bis nur noch zwei Cluster übrig sind

UPGMA



d_{ij}	1	2	3	4	5
1	0	2	6	9	7
2	2	0	5	7	7
3	6	5	0	5	4
4	9	7	5	0	3
5	7	7	4	3	0

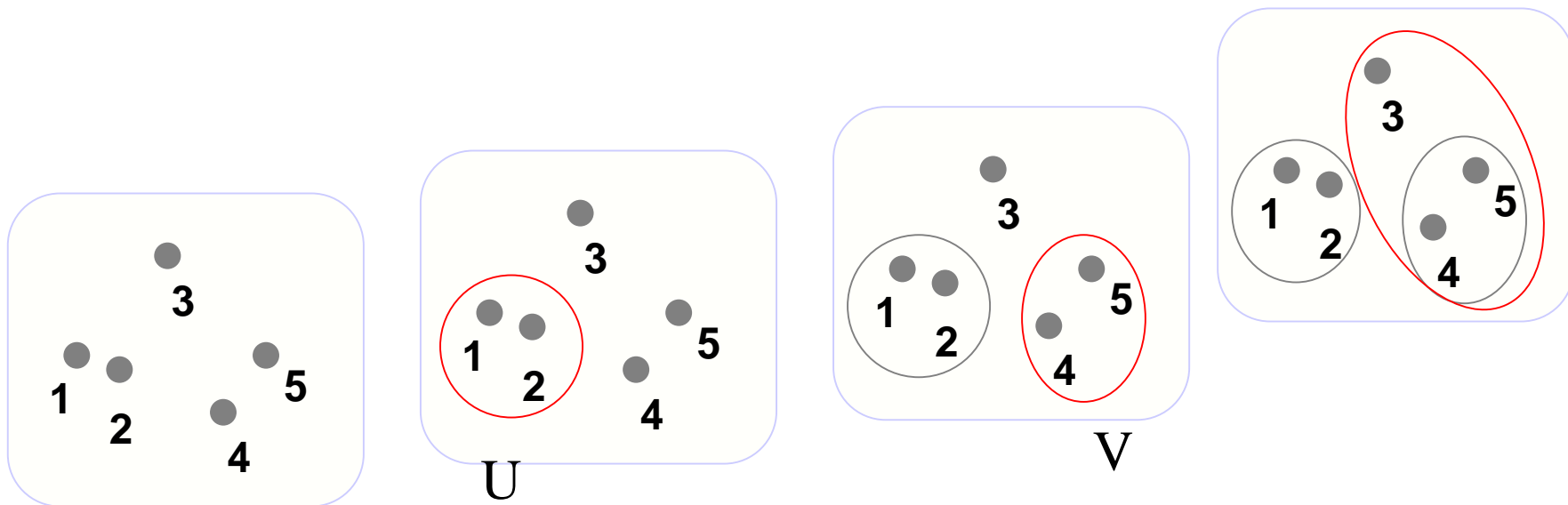
UPGMA

d_{ij}	1	2	3	4	5
1	0	2	6	9	7
2	2	0	5	7	7
3	6	5	0	5	4
4	9	7	5	0	3
5	7	7	4	3	0

d_{ij}	u	3	4	5
u	0	5.5	8	7
3	5.5	0	5	4
4	8	5	0	3
5	7	4	3	0

d_{ij}	u	3	v
u	0	5.5	7.5
3	5.5	0	4.5
v	7.5	4.5	0

d_{ij}	u	w
u	0	6.8
w	6.8	0



UPGMA

d_{ij}	1	2	3	4	5
1	0	2	6	9	7
2	2	0	5	7	7
3	6	5	0	5	4
4	9	7	5	0	3
5	7	7	4	3	0

d_{ij}	u	3	4	5
u	0	5.5	8	7
3	5.5	0	5	4
4	8	5	0	3
5	7	4	3	0

d_{ij}	u	3	v
u	0	5.5	7.5
3	5.5	0	4.5
v	7.5	4.5	0

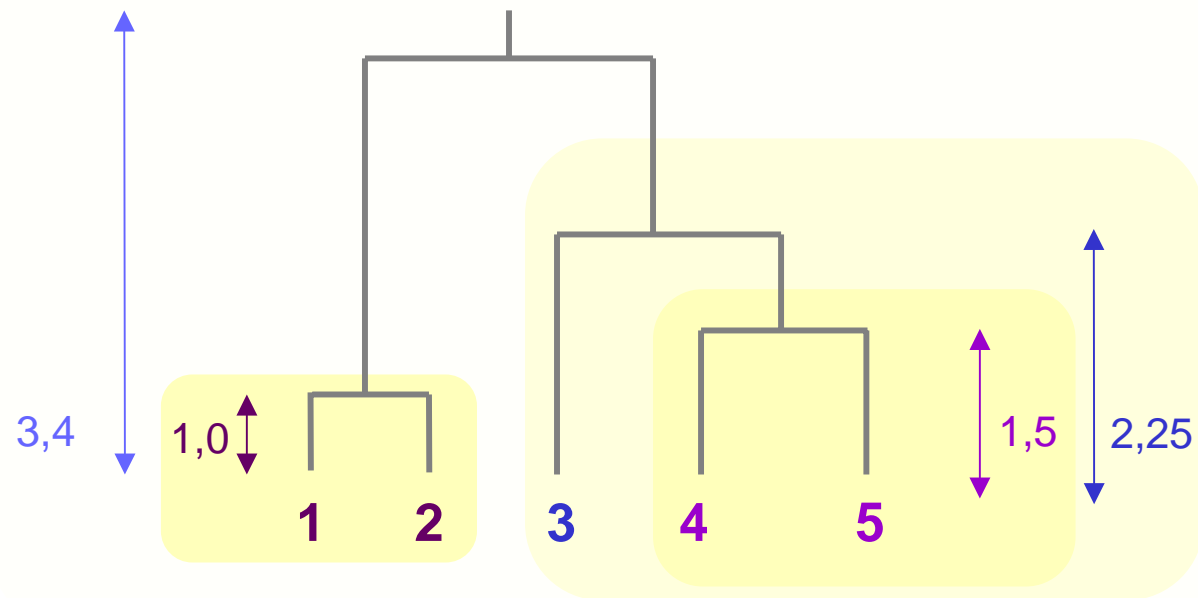
d_{ij}	u	w
u	0	6.8
w	6.8	0

Calculation of distances
between clusters:

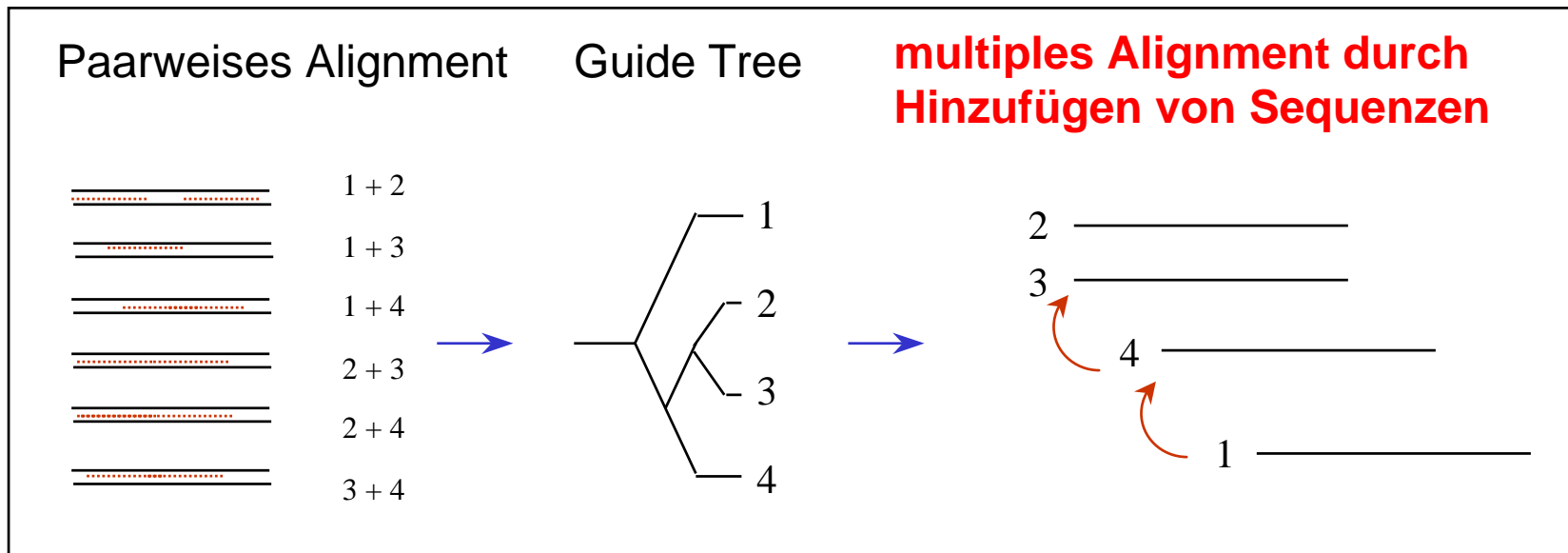
$$d_{u,v} = \frac{\sum_{i \in u} \sum_{j \in v} d_{ij}}{N}$$

$$N = |u| \times |v|$$

Der entstandene Guide-Tree



Progressives Alignment



1



2



3

Alignment-Profile

Wie ergeben sich diese Wahrscheinlichkeiten:
 $p_{AS} \sim$ Vorkommen in der Spalte

für das gesamte Profil ergibt sich dann die
folgende Formel:

$$P_i(a) = \frac{c_{ia}}{\sum_{a'} c_{ia'}} = \frac{c_{ia}}{N}$$

Wahrscheinlichkeit für
Aminosäure a in der i -
ten Spalte

Anzahl der
Zeilen

F
F
F
I
D
D
D

$$N = 7$$

$$c_F = 3$$

$$c_I = 1$$

$$c_D = 3$$

Anzahl der
Aminosäure a in
der i -ten Spalte

Wie wird eine Sequenz mit einem Profil aligniert?

=> dynamisches Programmieren mit Sequenz und Profil ("Mischsequenz")

(1) Berechne Score-Matrix $s(i,j)$ für diese Sequenz mit diesem Profil:

Profil-Position \swarrow Sequenz-Position \swarrow

$$s(i, j) = \sum_a P_i(a) \times \text{blosum}(a, b)$$

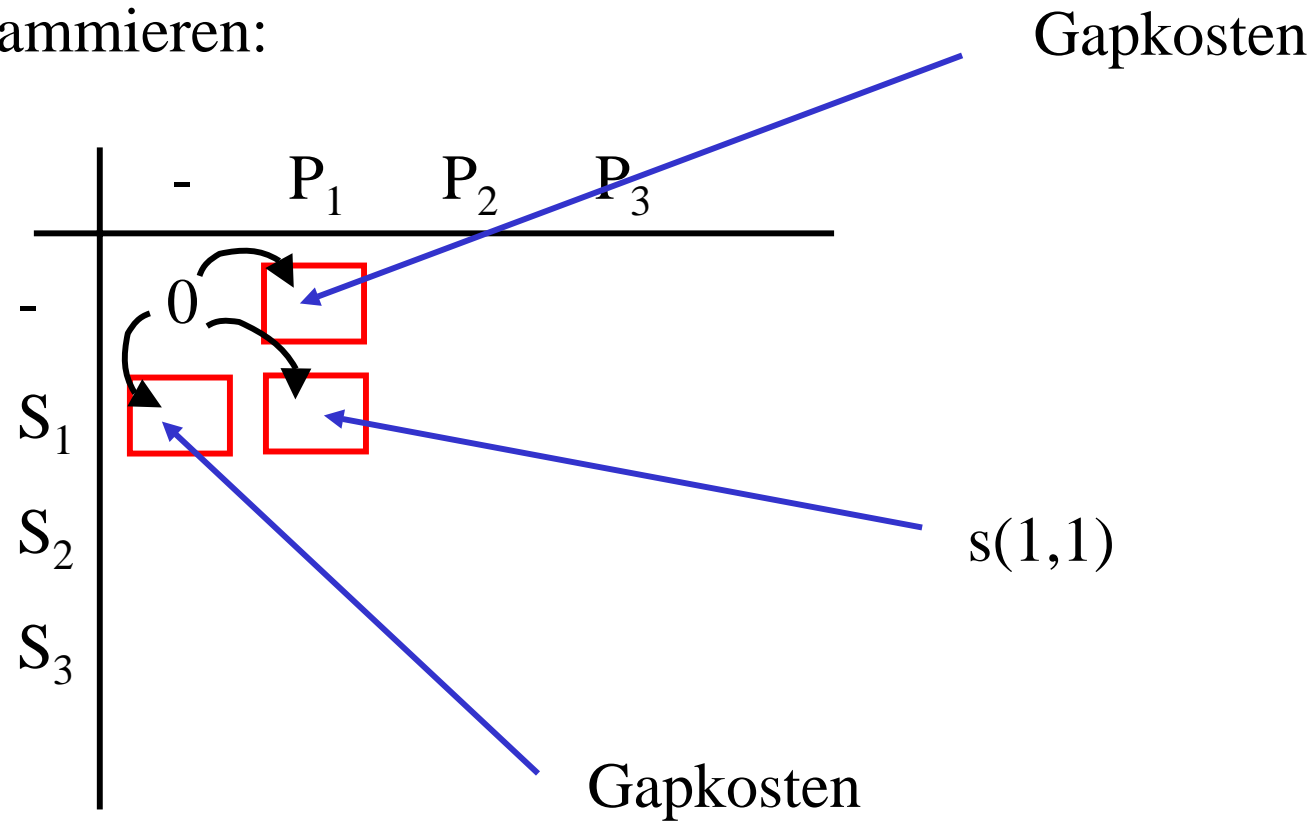
$P_i(a)$ ist die Wahrscheinlichkeit, dass Aminosäure a in der i -ten Spalte auftaucht (im Profil).

b ist die Aminosäure, die in der Sequenz an der j -ten Stelle kommt

(2) dynamisches Programmieren, Sequenz gegen Mischsequenz, mit Score-Matrix aus (1) und normalen Gapkosten

Wie wird eine Sequenz mit einem Profil aligniert?

Dynamisches
Programmieren:



Wie wird ein Profil mit einem Profil aligniert?

=> dynamisches Programmieren mit zwei Profilen

(1) berechnen der Score-Matrix $s(i,j)$:

$$s(i, j) = \sum_a \left[P_i(a) \sum_b \left[P_j(b) \times \text{blosum}(a, b) \right] \right]$$

$P_i(a)$ ist die Wahrscheinlichkeit, dass Aminosäure a in der i -ten Spalte im 2. Profil auftaucht,

$P_j(b)$ ist die Wahrscheinlichkeit, dass Aminosäure b in der j -ten Spalte im 2. Profil auftaucht

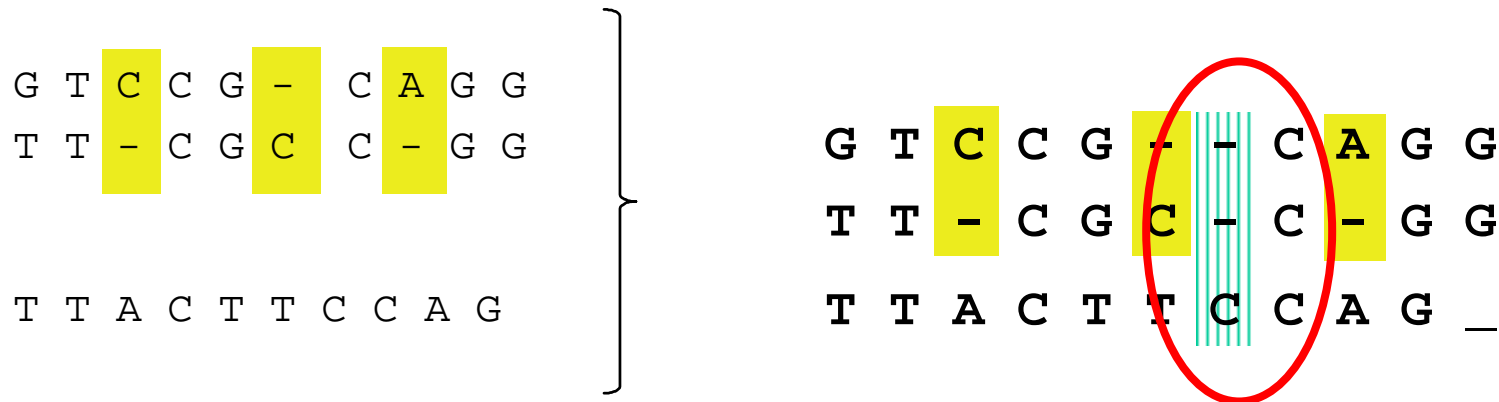
(2) dynamisches Programmieren mit dieser Score-Matrix

Spalten: einmal aligniert => nie mehr geändert

G	T	C	C	G	-	C	A	G	G	
T	T	-	C	G	C	C	-	G	G	
T	T	A	C	T	T	C	C	A	G	G

G	T	C	C	G	-	-	C	A	G	G
T	T	-	C	G	C	-	C	-	G	G
T	T	A	C	T	T	C	C	A	G	G

Spalten: einmal aligniert => nie mehr geändert



... und immer wieder neue Gaps dazu ...

Spalten: einmal aligniert => nie mehr geändert

G	T	C	C	G	-	-	C	A	G	G
T	T	-	C	G	C	-	C	-	G	G
T	T	A	C	T	T	C	C	A	G	-
}										
A	T	C	T	-	-	C	A	A	T	
C	T	G	T	C	C	C	T	A	G	

G	T	C	C	G	-	-	C	A	G	G
T	T	-	C	G	C	-	C	-	G	G
T	T	A	C	T	T	C	C	A	G	-
A	T	C	-	T	-	-	C	A	A	T
C	T	G	-	T	C	C	C	T	A	G

Gap-Kosten

- Problem: wenn man Gaps wie im paarweisen Alignment behandeln würde, würden es zuviele!

=> Lösung in Clustalw (Thompson, Higgins & Gibson 1994):

1. Sowohl Gap-Öffnen, als auch Gap-Erweiterung verteuert sich, wenn in der Spalte selbst keine Gaps sind, dafür aber in den Spalten daneben => zwingt Gaps in den gleichen Spalten aufzutreten

2. Gap-Kosten werden mit einem Modifizierer multipliziert, z.B. ergeben sich damit höhere Kosten in Spalten mit hydrophoben Residuen (im Protein geborgen, dürfen sich nicht groß ändern), als in Spalten mit hydrophilen oder flexiblen Residuen.

3. Gap-Öffnen-Kosten sind an Stellen reduziert, die von fünf oder mehr hydrophilen Residuen umgeben sind

4. Gaps treten bevorzugt neben einigen bestimmten Aminosäuren auf, begünstige diese Gaps

Bemerkung. 3. und 4. hängt mit der Beobachtung zusammen, dass Gaps vermehrt zwischen Sekundärstruktur auftritt

Sequenz- Gewichtung

- Homologe Sequenzen sind nicht unabhängig. Sie stammen in unterschiedlichem Maße von gleichen Vorfahren oder voneinander ab.

=> einige Paare sind enger miteinander verwandt als andere

- Hat man viele eng miteinander verwandte Sequenzen und ein paar, die weniger mit diesen verwandt sind, kann ein Ungleichgewicht beim Erstellen des Profils entstehen

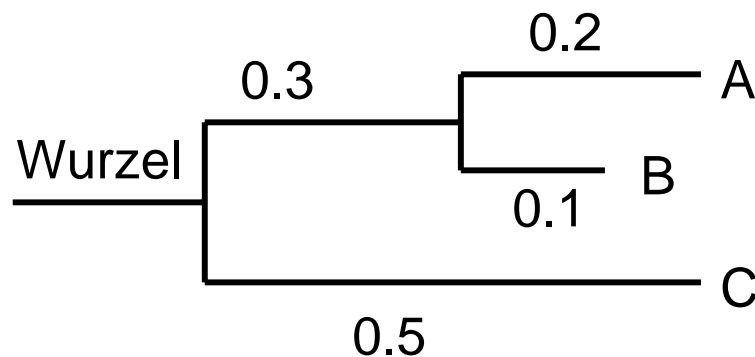
=> vermindere den Einfluss der eng Verwandten!

Sequenz- Gewichtung

Beispiel für 3 Sequenzen A, B, C:

- ungewichtet: für alle Sequenzen ist Gewicht w gleich (z.B 1 oder $1/3$)

- gewichtet:



$$w_A = 0.2 + 0.3/2 = 0.35$$

$$w_B = 0.1 + 0.3/2 = 0.25$$

$$w_C = 0.5$$

w = Distanz von der Wurzel, die sich geteilt wird, wenn Nachbarn auftauchen

-nach Normalisieren (Summe = 1): $w_A = 0.33$

$$w_B = 0.22$$

$$w_C = 0.45$$

$$w_A = 0.33$$

anstatt $w_B = 0.33$

$$w_C = 0.33$$

Sequenz- Gewichtung (bei CLUSTALW)

Beispiel für 4 Sequenzen A, B, C, D:

Spalte in Alignment 1

Sequenz A (Gewicht w_A)

Sequenz B (Gewicht w_B)

-----K-----

-----I-----

Score für Match dieser beiden Spalten in einem MSA:

Spalte in Alignment 2

Sequenz C (Gewicht w_C)

Sequenz D (Gewicht w_D)

-----L-----

-----V-----

$w_A \times w_C \times \text{score}(K,L) +$
 $w_A \times w_D \times \text{score}(K,V) +$
 $w_B \times w_C \times \text{score}(I,L) +$
 $w_B \times w_D \times \text{score}(I,V) / 4$

Einige Bemerkungen zur Software

Clustalw

Clustal: 1988, Higgins & Sharp

Clustalw: verbesserte Version, Sequenzen können gewichtet werden

Clustalx: wie Clustalw, enthält zusätzlich grafische Oberfläche

paarweises Alignment mit dynamischem Programmieren (verbesserte

Mmethod, von Myers & Miller 1988)

Guide-Tree: Neighbour-Joining

Pileup

im GCG Paket enthalten

paarweises Alignment: Needleman-Wunsch

Guide-Tree: UPGMA, Average-Linkage

Probleme mit Methoden des progressiven Alignments

- hängen stark vom Erfolg des paarweisen Alignments und der startenden zwei Sequenzen ab
 - => brauchen zwei nah verwandten Sequenzen zum Start (sehr verwandt)
 - => **alle Sequenze-Paare müssen paarweise alignierbar sein!! (verwandt)**
- wenn das nicht der Fall ist: versuche lokale Alignment-Methoden oder statistische Ansätze (z.B. HMM)

Methoden zum Multiplen-Sequenz-Alignment

- Multidimensionales dynamisches Programmieren
(MSA, Lipman 1988, DCA, Jens Stoye)
- Progressive Alignments
(Clustalw, Higgins 1996; PileUp, Genetics Computer Group (GCG))
- **Lokale Alignments**
(e.g. DiAlign, Morgenstern 1996; viele Andere)