

WS2018/2019

Genomforschung und Sequenzanalyse

- Einführung in Methoden der Bioinformatik-

Thomas Hankeln

HsaHBA	G	A	E	A	L	E	R	M	F	L	S	F	P	T	T	K	T	Y	F
HsaHBZ	G	T	E	T	L	E	R	L	F	L	S	H	P	Q	T	K	T	Y	F
HsaHBE	G	G	E	A	L	G	R	L	L	V	V	P	W	T	Q	R	F	F	F
HsaHBG	G	G	E	T	L	G	R	L	L	V	V	P	W	T	Q	R	F	F	F
HsaHBB	G	G	E	A	L	G	R	L	L	V	V	P	W	T	Q	R	F	F	F
HsaHBD	G	G	E	A	L	G	R	L	L	V	V	P	W	T	Q	R	F	F	F
HsaMB	G	Q	E	V	L	I	R	L	F	K	G	H	P	E	T	L	E	K	E

Multiples Sequenzalignment

Alignment von 2 Sequenzen

Gegeben: SeqA N A F L S
 SeqB N A F S

Gesucht: SeqA N A F L S
 SeqB N A F – S

Multiples Sequenz-Alignment

Gegeben:

SeqA	N	A	F	L	S	
SeqB	N	A	F	S		
SeqC	N	A	K	Y	L	S
SeqD	N	A	Y	L	S	

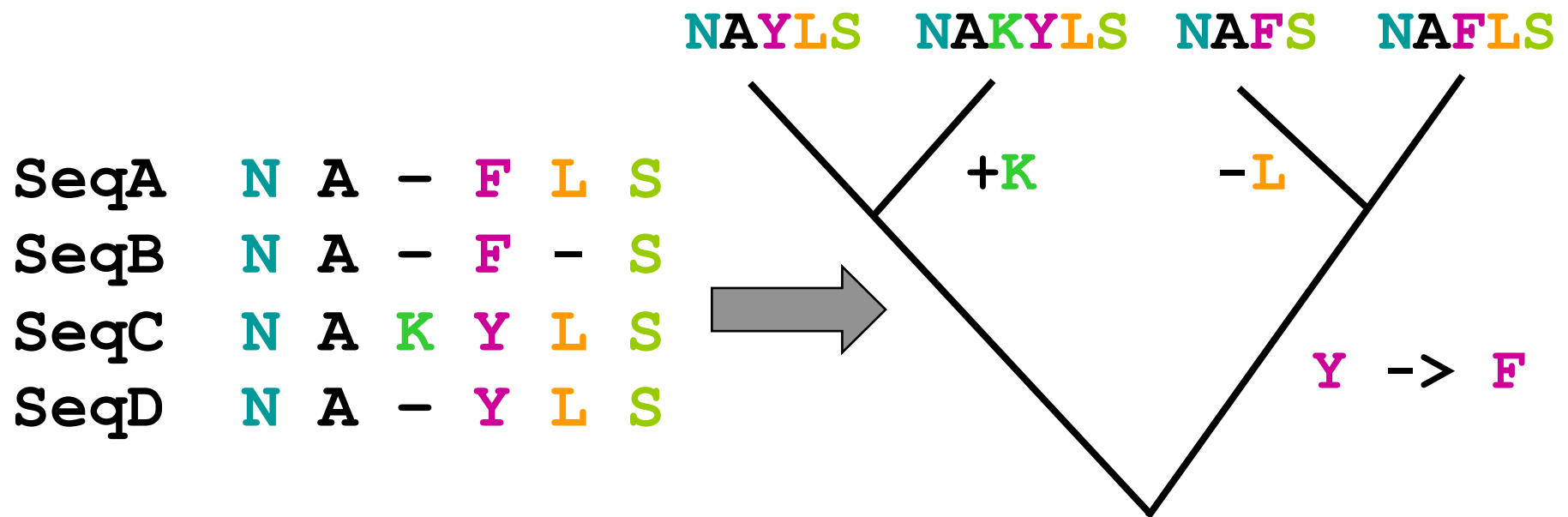
Gesucht:

SeqA	N	A	-	F	L	S
SeqB	N	A	-	F	-	S
SeqC	N	A	K	Y	L	S
SeqD	N	A	-	Y	L	S

Definitionen

- Multiple Sequence Alignment (Abkürzung = MSA)
- Deutsch: „Multiple Sequenzalignierung“
- Vergleich von > 2 DNA- oder As-Sequenzen
- Sequenzen sollen so aneinandergelegt werden, dass die Positionen **evolutiv homolog** sind!
=> Mutationen = historische Ereignisse

MSA & Evolution



Multiples Sequenz-Alignment = Evolutionshypothese!

DNA-Alignment

- DNA: **nicht-kodierende** oder **proteinkodierende** Sequenzen

1. **nicht-kodierend** (auch rRNA- oder tRNA-Gene!):

Seq1: ATCCTAGT-TCATGCATGCCTTAGCTC-TCTGAT

Seq2: ATTCTAGTTTCAT--ATGCTTCAGCTTATCTGAC

Seq3: AT-CTAGT-CCATGCATG-TTCAGCTTATGCGAC

2. **protein-kodierend** (=> Codons!):

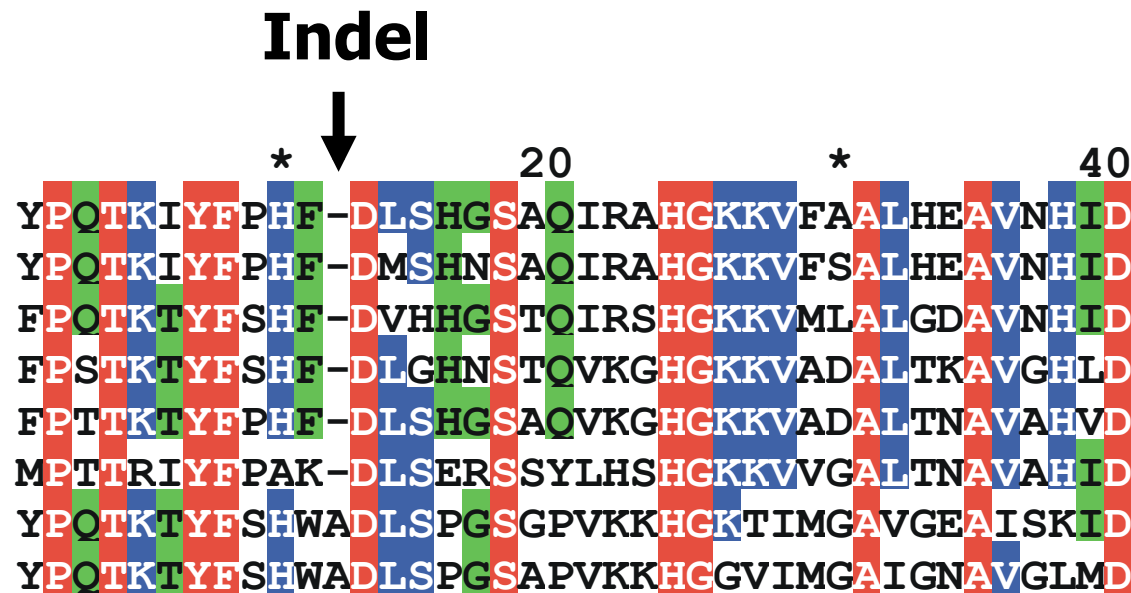
Seq1: ATC CTA GTT TCA TGC ATG CCT AAG CTC --- TGG

Seq2: ATT CTA GTT TCA --- ATG CTT CAG CTT ATC TGG

Seq3: ATC CTA GTT CCA TGC ATG TTT CAG CTT ATG CGG

=> Tripletstruktur liefert zusätzliche Information

Aminosäurealignment



Sequenzidentität = % identische AS

Sequenzähnlichkeit = % identische + % isofunktionelle AS

Insertion/Deletion = "Indel"

DNA- oder Aminosäurealignment?

Mögliche Austausche:

1. DNA

$4 \times 4 = 16$; mit Lücken (als 5. Base): $5 \times 5 = 25$ Möglichkeiten

2. Protein

$20 \times 20 = 400$; mit Lücken (als 21. AS): $21 \times 21 = 441$ Möglichk.

- mehr Informationsgehalt in Protein
- kurze Evolutionszeit: DNA, lange Distanzen: Protein

MSA: Anwendungsbeispiele

- Sequenz-“Gruppen“ definieren:
 Konservierte Domänen & Sequenzmotive
- Strukturvorhersagen bei Proteinen
- Sequenzierprojekte (Spezialfall)
- Evolutive Aussagen (**molekulare Phylogenie**)

MSA –Anwendung

- **Erkennung von Sequenz-„Gruppen“**

=> Proteinfamilien und –superfamilien

1. Proteinfamilie: hohe Identität (z.B. Hämoglobine)

2. Proteinsuperfamilie: geringe Identität (z.B. Globine)

=> häufig ähnliche Funktion!

- **Erkennen von entfernten Verwandtschaften**

=> häufig nicht mit einfachen Alignments zu erkennen

-> Protein"profile"

MSA –Anwendung

- „Entfernte Verwandte“, nur zwei Sequenzen...

[illegible]

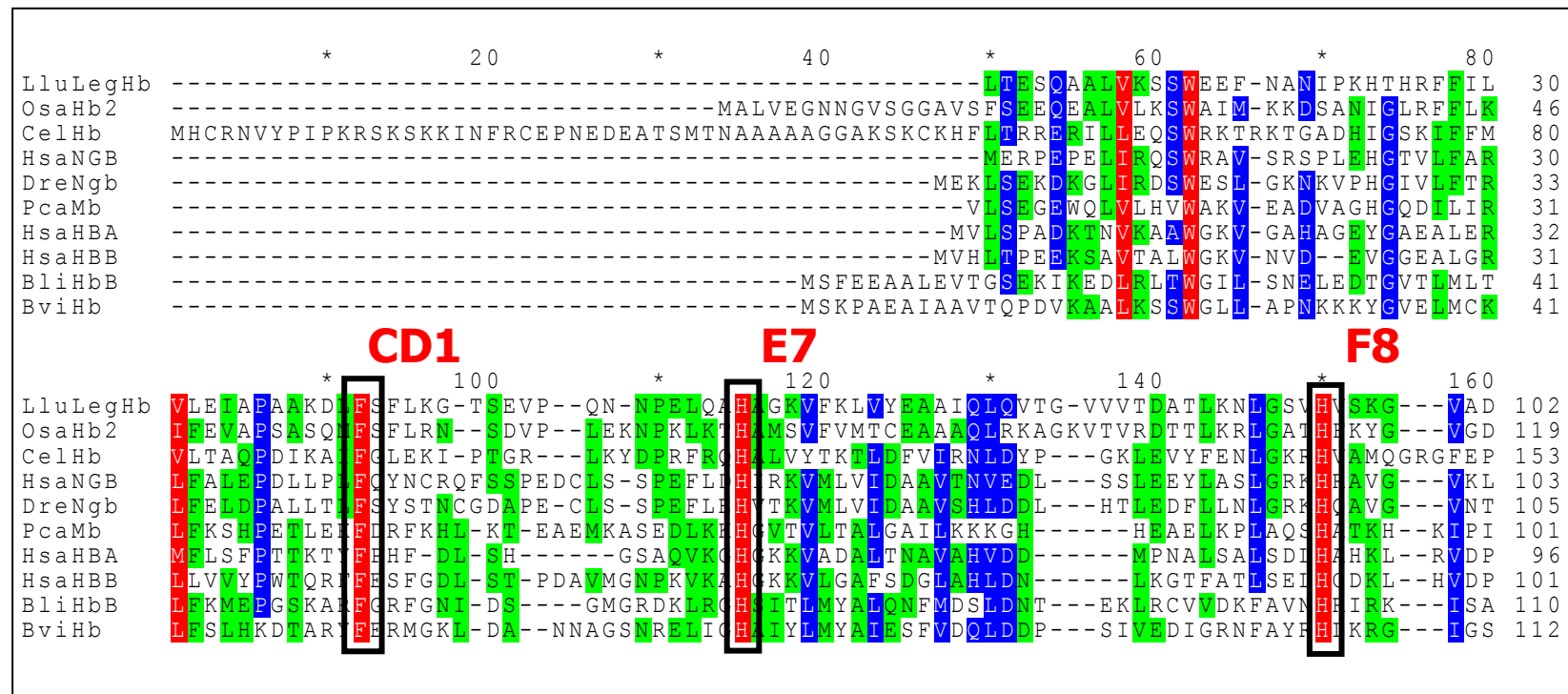
Mensch-Hämoglobin verglichen mit *C. elegans* Hämoglobin

=> nur 12% identische Aminosäuren

=> Verwandtschaft nicht abgesichert

MSA –Anwendung

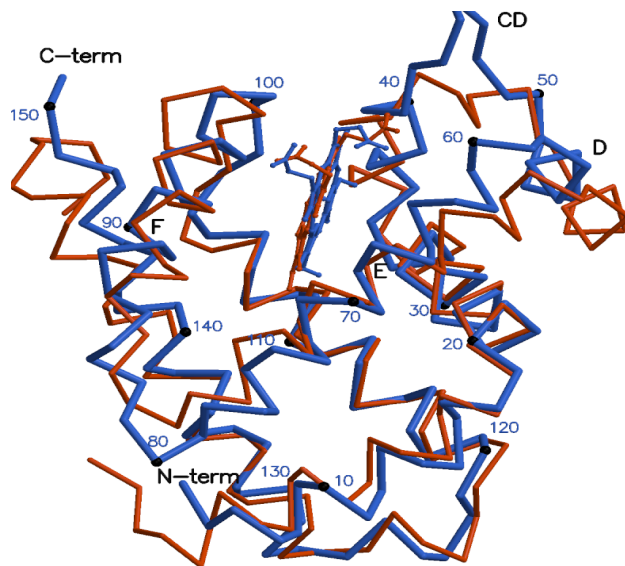
- „Entfernte Verwandte“, mehrere Sequenzen



MSA zeigt streng konservierte As-Positionen

MSA –Anwendung

- Vorhersage der Struktur von Proteinen mit divergenter Sequenz, aber sehr ähnlicher 3D-Faltung



Myoglobin

Neuroglobin

Trennung vor ~ 800 Mio Jahren!

HsaNGB	-MERPEPELTIROSWRVSRSPLEHGTVLIFARLFALEPDLIPLFOYNCROFSSPEDCLS	SPEFLDHIRKVMLVTDAAVTN	78
PcaMb	VLSEGEWOLVHLVWAKVEADVAGHGODILIRLEKSHPETIEKFDRFKHL-KTEAEMKAS	EDLKKHGVTVITAGAIKK	78

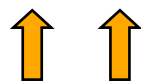
HsaNGB	VEDLSSLFEYLASDGRKHRAVG-VKLSFSFSTVGESLLYMLEKCLGPAFTPATRAAWSOLYGAVVOAMSRGWDGE----	151
PcaMb	KGHH--FAETKPTAOSHATKHKTPKYLEFTSEATTHTVTHSRHPGDEGADAOGAMNKALELFRKDTAAKYKELGYOG	153

MSA –Anwendung

- **Konservierte Domänen & Sequenzmotive**
=> Proteine mit gemeinsamer Funktion erkennen

Bsp: Cu-Bindungsstelle in Hämocyaninen und Phenoloxidase:

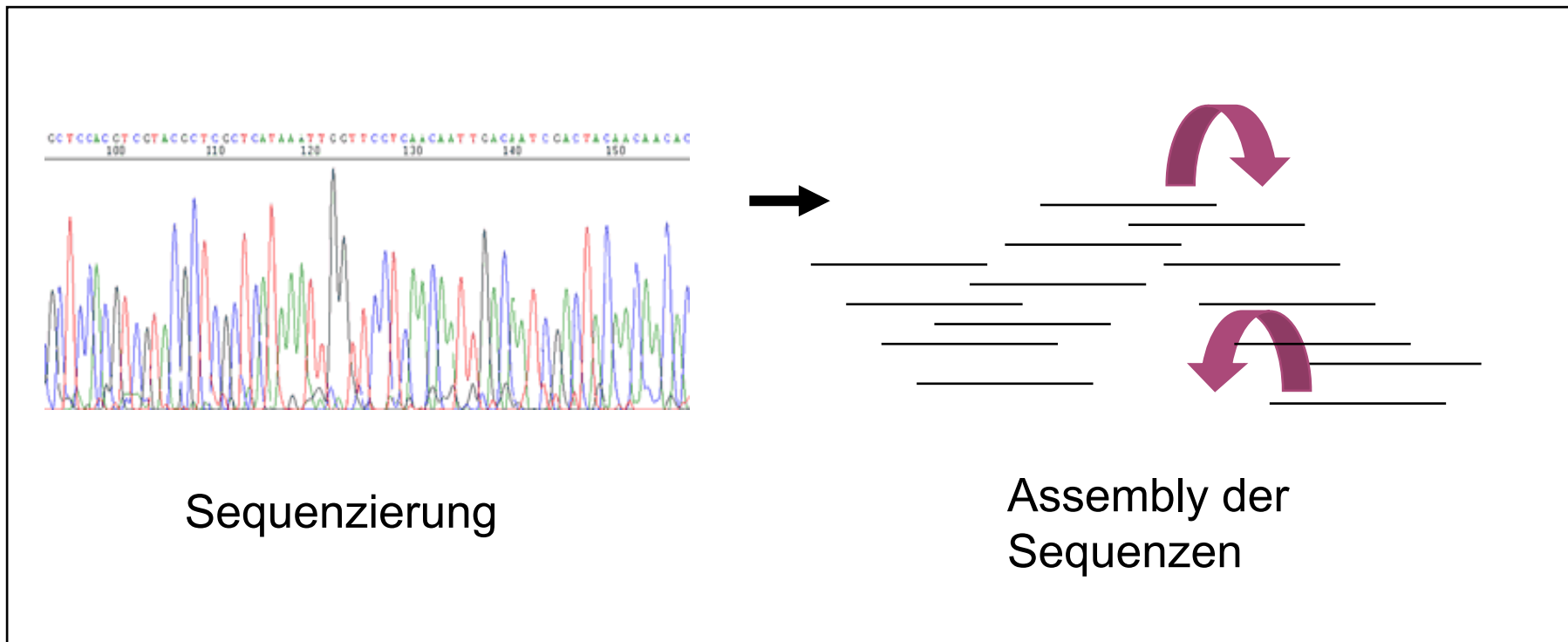
PlePPO	Y	W	R	E	D	F	G	I	N	S	H	H	W	H	W	H	L	V	Y	P	I	E	M	-	-	-	-	N	V	N	R	D	R	K	G	E	L	F	Y	Y	M	H	Q	Q	M	V	A	R	Y	D	W	E	R	L	S	V	N	L	N	R	V	E	K	L	E	61	
PmoPPO	Y	W	R	E	D	Y	G	I	N	V	H	H	W	H	W	H	L	I	Y	P	P	A	M	-	-	-	-	G	F	D	R	D	R	K	G	E	L	F	Y	Y	M	H	Q	Q	V	I	A	R	Y	D	I	E	R	L	C	L	G	L	P	K	V	E	K	L	D	61	
BmoPPO1	Y	F	R	E	D	I	G	I	N	L	H	H	W	H	W	H	L	V	Y	P	F	D	A	A	D	R	A	-	I	V	N	K	D	R	R	G	E	L	F	Y	Y	M	H	Q	Q	I	I	A	R	Y	N	V	E	R	M	C	N	N	L	S	R	V	R	R	Y	N	65
DmePPOA1	Y	F	R	E	D	I	G	V	N	S	H	H	W	H	W	H	L	V	Y	P	T	T	G	P	T	E	-	-	V	V	N	K	D	R	R	G	E	L	F	Y	Y	M	H	Q	I	L	A	R	Y	N	V	E	R	F	C	N	N	L	K	K	V	Q	P	L	N	64	
DmePPO2	Y	F	R	E	D	L	G	I	N	L	H	H	W	H	W	H	L	V	Y	P	F	E	A	S	D	R	S	-	I	V	A	K	D	R	R	G	E	L	F	Y	Y	M	H	Q	Q	V	I	A	R	Y	N	A	E	R	F	S	N	N	L	A	R	V	L	P	F	N	65
DmePPO3	Y	F	R	E	D	L	G	V	N	L	H	H	W	H	W	H	L	V	Y	P	I	E	A	P	D	R	S	-	I	V	D	K	D	R	R	G	E	L	F	Y	Y	M	H	Q	Q	I	I	A	R	Y	N	A	E	R	L	S	N	H	M	A	R	V	Q	P	F	N	65
EcaHcA	Y	F	R	E	D	I	G	V	N	A	H	H	W	H	W	H	V	Y	P	S	T	Y	D	P	A	F	F	G	K	V	K	D	R	K	G	E	L	F	Y	Y	M	H	Q	Q	M	C	A	R	Y	D	C	E	R	L	S	N	G	L	N	R	M	I	P	F	H	66	
EcaHcD	Y	F	R	E	D	I	G	I	N	S	H	H	W	H	W	H	L	V	Y	P	A	F	Y	D	A	D	I	F	G	K	I	K	D	R	K	G	E	L	F	Y	Y	M	H	Q	Q	M	C	A	R	Y	D	C	E	R	L	S	V	G	L	Q	R	M	I	P	F	Q	66
EcaHcF	Y	F	R	E	D	I	G	A	N	A	H	H	W	H	W	H	I	V	Y	P	P	T	W	D	A	S	V	M	S	K	V	K	D	R	K	G	E	L	F	Y	Y	M	H	Q	Q	M	C	A	R	Y	D	C	D	R	L	S	T	G	L	R	R	M	I	P	F	H	66
LpoHc2	Y	Y	R	E	D	V	G	I	N	A	H	H	W	H	W	H	L	V	Y	P	S	T	W	N	P	K	Y	F	G	K	K	K	D	R	K	G	E	L	F	Y	Y	M	H	Q	Q	M	C	A	R	Y	D	C	E	R	L	S	N	G	M	H	R	M	L	P	F	N	66
PvaHc	Y	F	G	E	D	I	G	L	N	T	H	H	V	T	W	H	M	E	F	P	F	W	W	N	D	A	Y	G	-	H	H	L	D	R	K	G	E	N	F	F	W	I	H	H	Q	L	T	V	R	F	D	A	E	R	L	S	N	Y	L	D	P	V	G	E	L	Q	65
PirHcC	Y	F	G	E	D	V	G	M	N	T	H	H	V	L	W	H	M	E	F	P	F	W	W	E	D	S	S	G	-	R	H	L	D	R	K	G	E	S	F	F	W	V	H	H	Q	L	T	V	R	Y	D	A	E	R	L	S	N	H	L	D	P	V	E	E	L	S	65
PirHcA	Y	F	G	E	D	I	G	M	N	I	H	H	V	T	W	H	M	D	F	P	F	W	W	E	D	S	Y	G	-	Y	H	L	D	R	K	G	E	L	F	F	W	V	H	H	Q	L	T	A	R	F	D	E	R	L	S	N	W	L	D	P	V	D	E	L	H	65	



Cu - Bindungsstellen

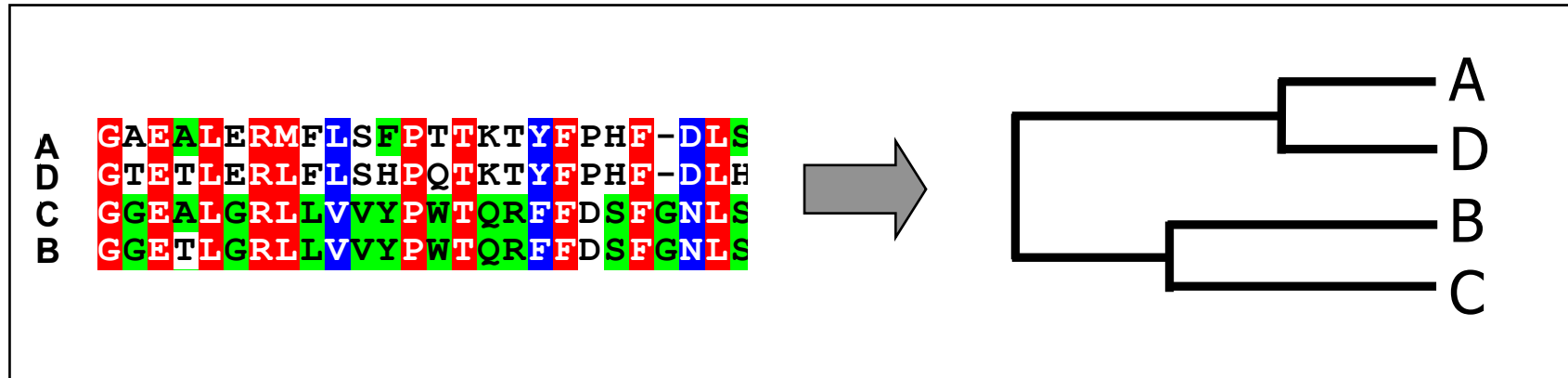
MSA –Anwendung

- Sequenzierprojekte

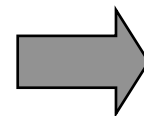


MSA –Anwendung

- **Evolute Aussagen**
=> **Phylogenetische Rekonstruktion (Stammbäume)**



Garbage IN



Garbage OUT

Wie kommt man zum MSA?

- Alignment per Hand/Auge:
gut bei relativ ähnlichen Sequenzen
- Mit dem Computer:
Exakte Algorithmen
Heuristische Verfahren

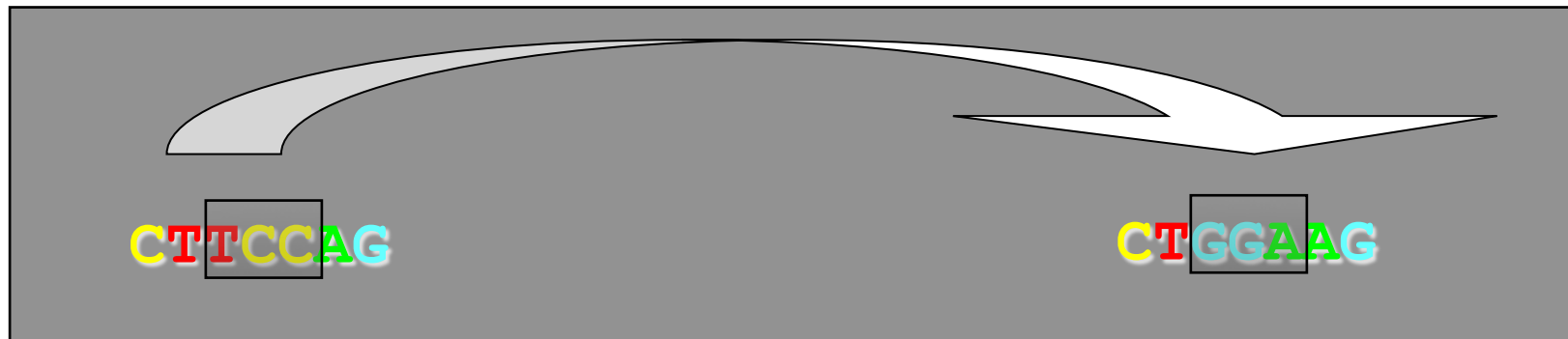
Alle Verfahren brauchen Bewertungsschemata



Kennen wir! => z.B. PAM & BLOSUM-Matrizen, gap penalty

Manchmal hilft visuelle Bewertung...

Ancistr.	TTGACATAAGCTTCCAGTTT
Brach.	TTGACATAAGCTTCGAGTTT
Meteorium	TTGACATAAACTGGAAGTTT
Papillaria	TTGACATAAACTTCTAGTTT



Bewertung eines MSA

- Optimales Alignment ist ein Alignment mit minimalen Kosten (oder maximaler „Belohnung“ für Ähnlichkeit).
- Algorithmen sollen ein Alignment erstellen, welches die Kosten minimiert bzw. die Ähnlichkeit optimiert.

MSA und Gaps^{*}

- Wie wahrscheinlich ist es, dass zwischen Sequenzpaaren und Sequenzgruppen Lücken (gaps) auftreten?
- Gaps stellen Insertionen und Deletionen (Indels) dar, die in der Evolution aufgetreten sind
- Ihre Wahrscheinlichkeit kann definiert werden.

^{*} Das Problem kennen wir bereits von den dualen Alignments 20

Was kostet ein Gap?

Hämoglobin- und Myoglobinsequenzen von Mensch

CLUSTALX: Gap creation weight 20, gap length weight 0.1*

HsaHBA	GAEALERMFLSFPTTKTYFPHF	-----	DLSHGSAQVKGHGKKVADAL	TNAVAVDDMPN	54
HsaHBZ	GTETLERLFLSHPQTKTYFPHF	-----	DLHPGSAQLRAHGSKVVAAVGDAVKSIDDIGG		54
HsaHBE	GGEALGRLLVVYPWTQRFFDSFGNLSSPSAILGNPKVKAHGKKVLTSFGDAIKNMDNLKP				60
HsaHBG	GGETLGRLLVVYPWTQRFFDSFGNLSSASAIMGNPKVKAHGKKVLTSLGDAIKHLDDLKG				60
HsaHBB	GGEALGRLLVVYPWTQRFFESFGDLSTPDAMGNPKVKAHGKKVLGAFSDGLAHLDNLKG				60
HsaHBD	GGEALGRLLVVYPWTQRFFESFGDLSSPDAMGNPKVKAHGKKVLGAFSDGLAHLDNLKG				60
HsaMB	GQEVLIIRLFKGFHPEFLEKFDKFKHLKSEDEMKASEDLKKHGATVLTALGGILKKKGHHEA				60

* synonym: „gap opening penalty“, „gap extension penalty“

Was kostet ein Gap?

Hämoglobin- und Myoglobinsequenzen von Mensch

CLUSTALX: Gap creation weight 10, gap length weight 0.1

HsaHBA	G	A	E	A	L	E	R	M	F	L	S	F	P	T	T	K	T	Y	P	H	F	-	D	L	S	H	-	-	-	-	-	G	S	A	Q	V	K	G	H	G	K	K	V	A	D	A	L	T	N	A	V	A	H	V	D	D	M	P	N	54	
HsaHBZ	G	T	E	T	L	E	R	L	F	L	S	H	P	Q	T	K	T	Y	P	H	F	-	D	L	H	P	-	-	-	-	-	G	S	A	Q	L	R	A	H	G	S	K	V	V	A	A	V	G	D	A	V	K	S	I	D	D	I	G	G	54	
HsaHBE	G	G	E	A	L	G	R	L	L	V	V	P	W	T	Q	R	F	F	D	S	F	G	N	L	S	S	P	S	A	I	L	G	N	P	K	V	K	A	H	G	K	K	V	L	T	S	F	G	D	A	I	K	N	M	D	N	L	K	P	60	
HsaHBG	G	G	E	T	L	G	R	L	L	V	V	P	W	T	Q	R	F	F	D	S	F	G	N	L	S	S	P	S	A	I	L	G	N	P	K	V	K	A	H	G	K	K	V	L	T	S	L	G	D	A	I	K	H	L	D	D	L	K	G	60	
HsaHBB	G	G	E	A	L	G	R	L	L	V	V	P	W	T	Q	R	F	F	E	S	F	G	D	L	S	T	P	D	A	V	M	G	N	P	K	V	K	A	H	G	K	K	V	L	G	A	F	S	D	G	L	A	H	L	D	N	L	K	G	60	
HsaHBD	G	G	E	A	L	G	R	L	L	V	V	P	W	T	Q	R	F	F	E	S	F	G	D	L	S	S	P	D	A	V	M	G	N	P	K	V	K	A	H	G	K	K	V	L	G	A	F	S	D	G	L	A	H	L	D	N	L	K	G	60	
HsaMB	G	Q	E	V	L	I	R	L	F	K	G	H	P	E	T	L	E	K	F	D	K	F	K	H	L	K	S	E	D	E	M	K	A	S	E	D	L	K	K	H	G	A	T	V	L	T	A	L	G	G	I	L	K	K	K	G	H	H	E	A	60

Welches Alignment ist aber nun besser?

Was kostet ein Gap?

Erst Hinzunahme einer **Hai-Hb-alpha**-Sequenz bringt die Lösung:

SacHbA	G	A	E	S	L	A	R	M	F	A	T	T	P	S	T	K	T	Y	F	S	K	F	T	D	E	S	A	-	-	-	-	N	G	K	R	V	K	A	H	G	G	K	V	L	N	A	V	A	D	A	T	D	H	L	D	N	V	A	G	55	
HsaHBA	G	A	E	A	L	E	R	M	F	L	S	F	P	T	T	K	T	Y	F	P	F	-	D	L	S	H	-	-	-	-	G	S	A	Q	V	K	G	H	G	K	K	V	A	D	A	L	T	N	A	V	A	H	V	D	D	M	P	N	54		
HsaHBZ	G	T	E	T	L	E	R	L	F	L	S	H	P	Q	T	K	T	Y	F	P	H	-	D	L	H	P	-	-	-	-	G	S	A	Q	L	R	A	H	G	S	K	V	V	A	A	V	G	D	A	V	K	S	I	D	D	I	G	54			
HsaHBB	G	G	E	A	L	G	R	L	L	V	V	Y	P	W	T	Q	R	F	F	E	S	F	D	L	S	T	P	D	A	V	M	G	N	P	K	V	K	A	H	G	K	K	V	L	G	A	F	S	D	G	L	A	H	L	D	N	L	K	G	60	
HsaHBD	G	G	E	A	L	G	R	L	L	V	V	Y	P	W	T	Q	R	F	F	E	S	F	G	D	L	S	T	P	D	A	V	M	G	N	P	K	V	K	A	H	G	K	K	V	L	G	A	F	S	D	G	L	A	H	L	D	N	L	K	G	60
HsaHBE	G	G	E	A	L	G	R	L	L	V	V	Y	P	W	T	Q	R	F	F	D	S	F	G	N	L	S	S	P	S	A	I	L	G	N	P	K	V	K	A	H	G	K	K	V	L	T	S	F	G	D	A	I	K	N	M	D	N	L	K	P	60
HsaHBG	G	G	E	T	L	G	R	L	L	V	V	Y	P	W	T	Q	R	F	F	D	S	F	G	N	L	S	S	A	S	A	I	M	G	N	P	K	V	K	A	H	G	K	K	V	L	T	S	L	G	D	A	I	K	H	L	D	D	L	K	G	60
HsaMB	G	Q	E	V	L	I	R	L	F	K	G	H	P	E	T	L	E	K	F	D	K	F	K	H	L	K	S	E	D	E	M	K	A	S	E	D	L	K	K	H	G	A	T	V	L	T	A	L	G	I	L	K	K	K	G	H	E	A	60		

=> Alignment # 2 (10 gap creation weight) ist plausibel!

Bewertung eines MSA I: „sum of pairs-score“

Beispiel (DNA) für die Berechnung der Kosten (c) mit:
match c = 0; mismatch c = 2; gap c = 1

S1	=	-	G	C	T	G	A	T	A	T	A	A	C	T
S2	=	G	G	G	T	G	A	T	-	T	A	G	C	T
S3	=	A	G	C	G	G	A	-	A	C	A	C	C	T

sum of pairs:		4	0	4	4	0	0	2	2	4	0	6	0	0 = 26

Vorgehensweise:

- Summieren der Kosten der Paare ("sum of pairs") der 1. Spalte:
 $c(-,G) + c(G,A) + c(-,A) = 4$, usw.
- Summieren der Spaltenwerte = 26 (Gesamtkosten!)

Bewertung eines MSA I: „sum of pairs-score“

```
S1 = - G C T G A T A T A A C T
S2 = G G G T G A T - T A G C T
S3 = A G C G G A - A C A C C T
-----
sum of pairs: 4 0 4 4 0 0 2 2 4 0 6 0 0 = 26
```

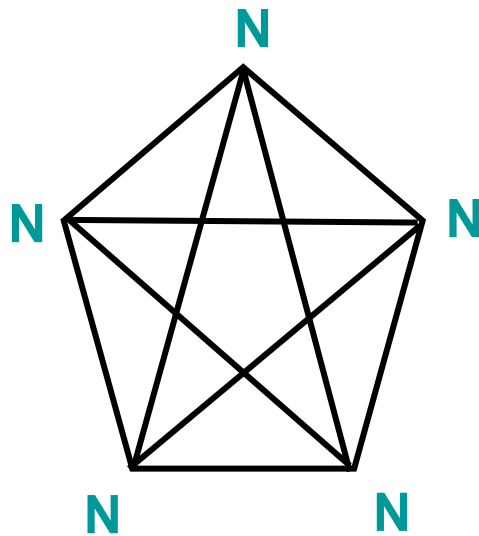
```
S1 = G - C T G A T A T A A C T
S2 = G G G T G A - T T A G C T
S3 = A G C G G A - A C A C C T
-----
sum of pairs: 4 2 4 4 0 0 2 4 4 0 6 0 0 = 30
```

Bewertung eines MSA I:

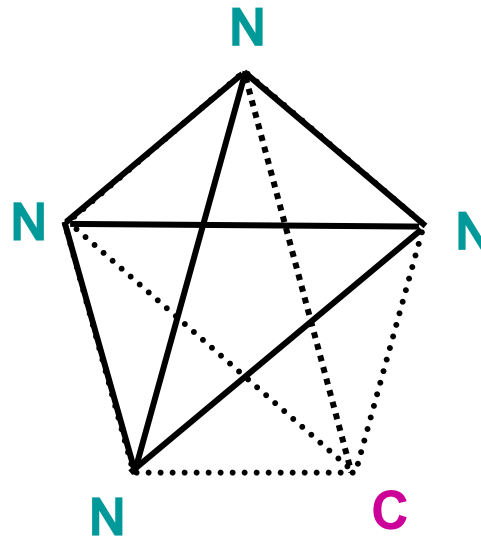
„sum of pairs“

Beispiel: Protein
Bewertung mit "sum of pairs",
und BLOSUM-Matrix

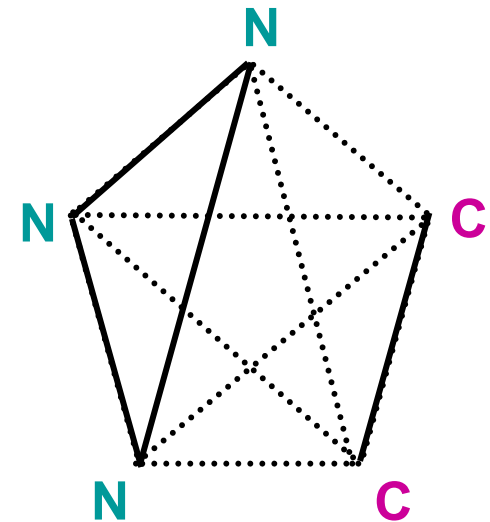
Seq 1N.....N.....N...
Seq 2N.....N.....N...
Seq 3N.....N.....N...
Seq 4N.....N.....C...
Seq 5N.....C.....C...



Spalte 1

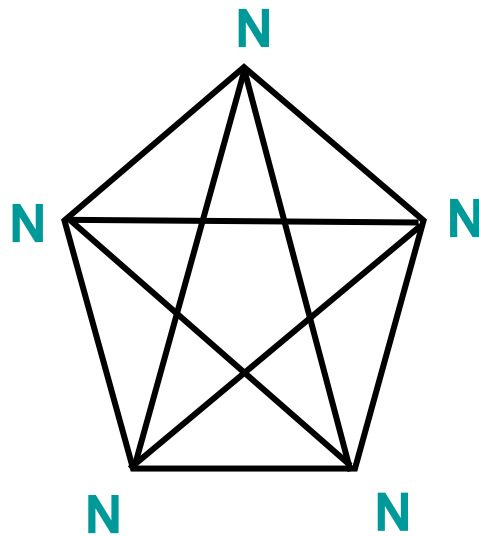


Spalte 2

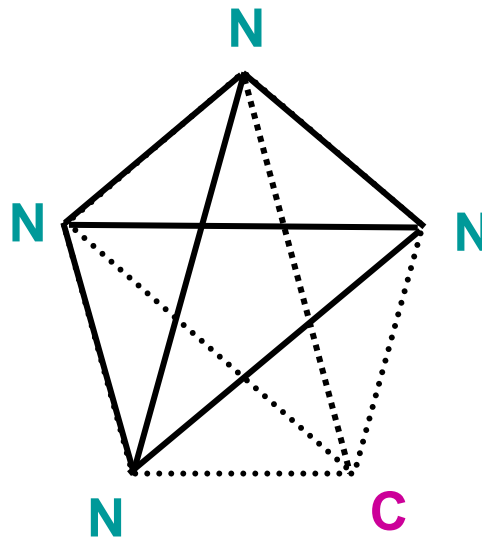


Spalte 3

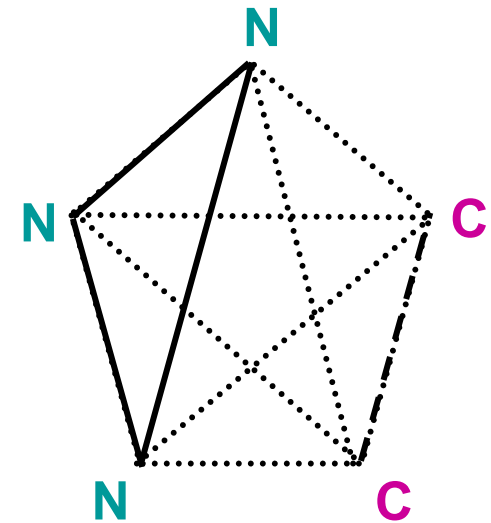
Bewertung eines MSA I: „sum of pairs“



Spalte 1



Spalte 2



Spalte 3

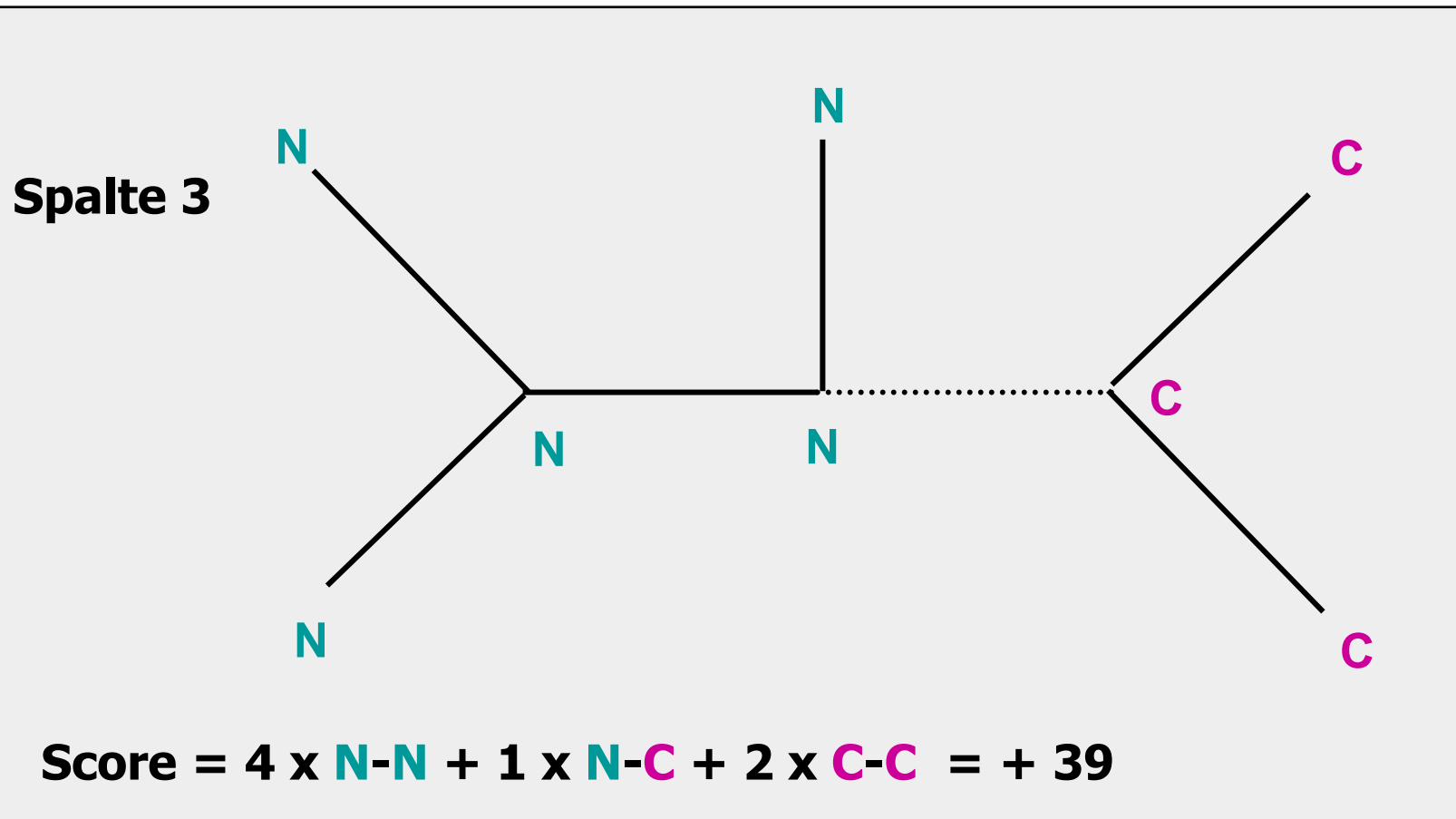
BLOSUM 62 scores

N – N:	+6
N – C:	-3
C – C:	+9

Spalte 1:	10 x N-N	= + 60
Spalte 2:	6 x N-N + 4 x N-C	= + 24
Spalte 3:	3 x N-N + 6 x N-C + 1 x C-C	= + 9

Bewertung eines MSA II

Bewertung mithilfe eines Stammbaums (z.B. CLUSTAL):



Lösungsansätze

- Exakte Algorithmen
 - ➔ • Progressive Verfahren
 - Iterative Verfahren
 - Divide-and-Conquer-Verfahren
 - Profile Hidden Markov Modelle
- } Heuristische Verfahren

Exakte Algorithmen

Wir erinnern uns: **Needleman & Wunsch** für paarweises globales Alignment

		R	L	G	Y	E
	0	1	2	3	4	5
R	1	0	1	2	3	4
G	2	1	1	1	2	3
Y	3	2	2	2	1	2
A	4	3	3	3	2	2

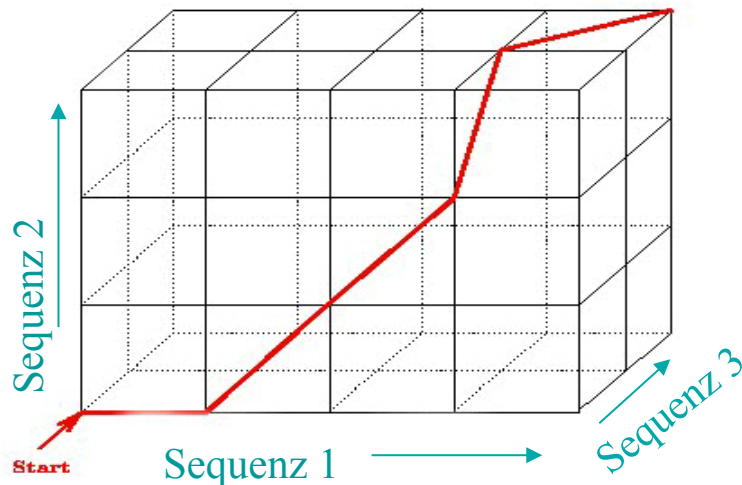
Kosten c als:

$$c(a,b) = c(a,-) = 1$$

$$c(a,a) = 0$$

Lösung: **R L G Y E**
R - G Y A

N & W mit k Sequenzen



Needleman-Wunsch:
in 3 Dimensionen!

Verwendung von k -
dimensionalen Gittergraphen

- produziert das bestmögliche globale MSA
- sehr rechenintensiv, wenn k groß (Aufwand: n^M)
- praktikabel bis 10 x 1000 As (Programm MSA auf Supercomputern)

Daher...

Heuristische Verfahren

- Progressive Verfahren
- Iterative Verfahren
- Divide-and-Conquer-Verfahren
- Hidden Markov Modelle

u. v. m.

...gibt 's „global“ und „lokal“

Progressives Alignment

1. paarweiser Vergleich aller Sequenzen miteinander => Berechnung der Distanzen zweier Sequenzen
2. gruppiert Sequenzen nach Ähnlichkeit (Cluster-Bildung)
3. Erstellung paarweiser Alignments
4. sukzessives Alignment nach Ähnlichkeit, dabei die ähnlichsten Sequenzpaare zuerst

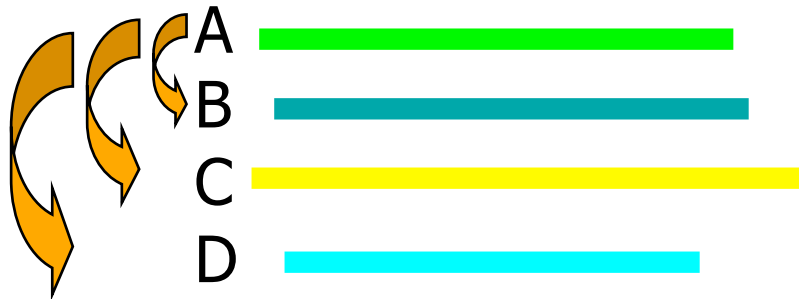
Feng & Doolittle (1987): **PileUp** (GCG package)

Higgins and Sharp (1988), Thompson et al. (1994): **CLUSTAL**

Notredame et al. (2000): **T-COFFEE**

Progressives Alignment

1) Sequenzvergleich



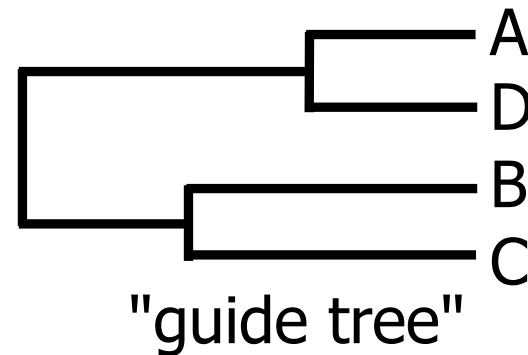
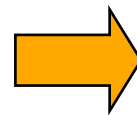
Alle Sequenzen werden miteinander verglichen
(Option A: schnelles "quick and dirty" Alignment
Option B: exaktes, langsames Needleman-Wunsch)
=> Berechnen der Distanzen

Progressives Alignment

2) Ähnliche Sequenzen werden gruppiert

=> Cluster-Analyse = Erstellung eines hierarchischen Stammbaums ("guide tree").

	A	B	C	D
A	-	0.75	0.89	0.27
B		-	0.45	0.82
C			-	0.77
D				-

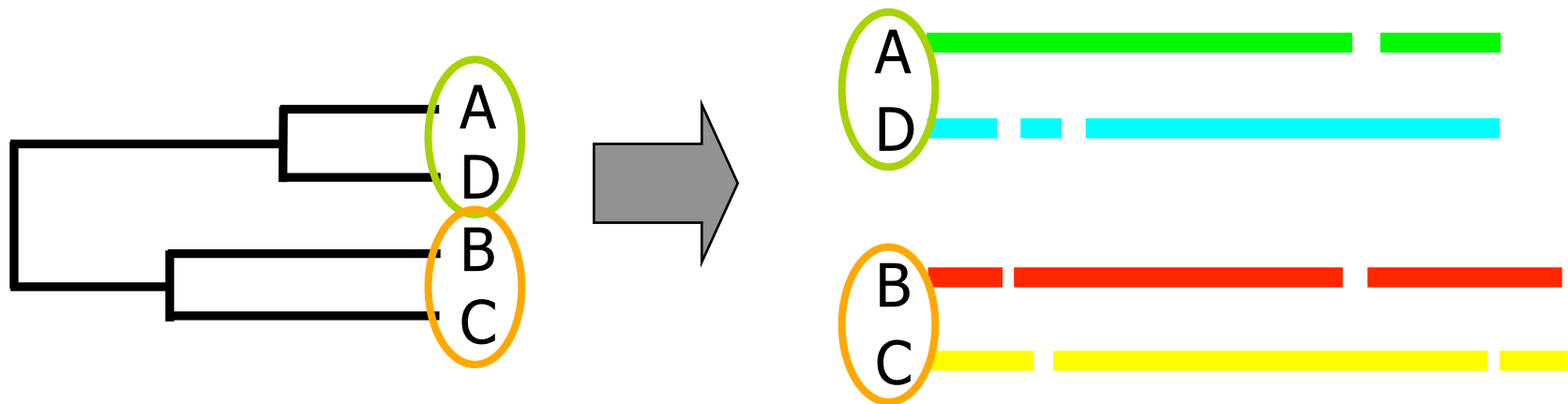


PileUp = UPGMA

Clustal = Neighbor Joining

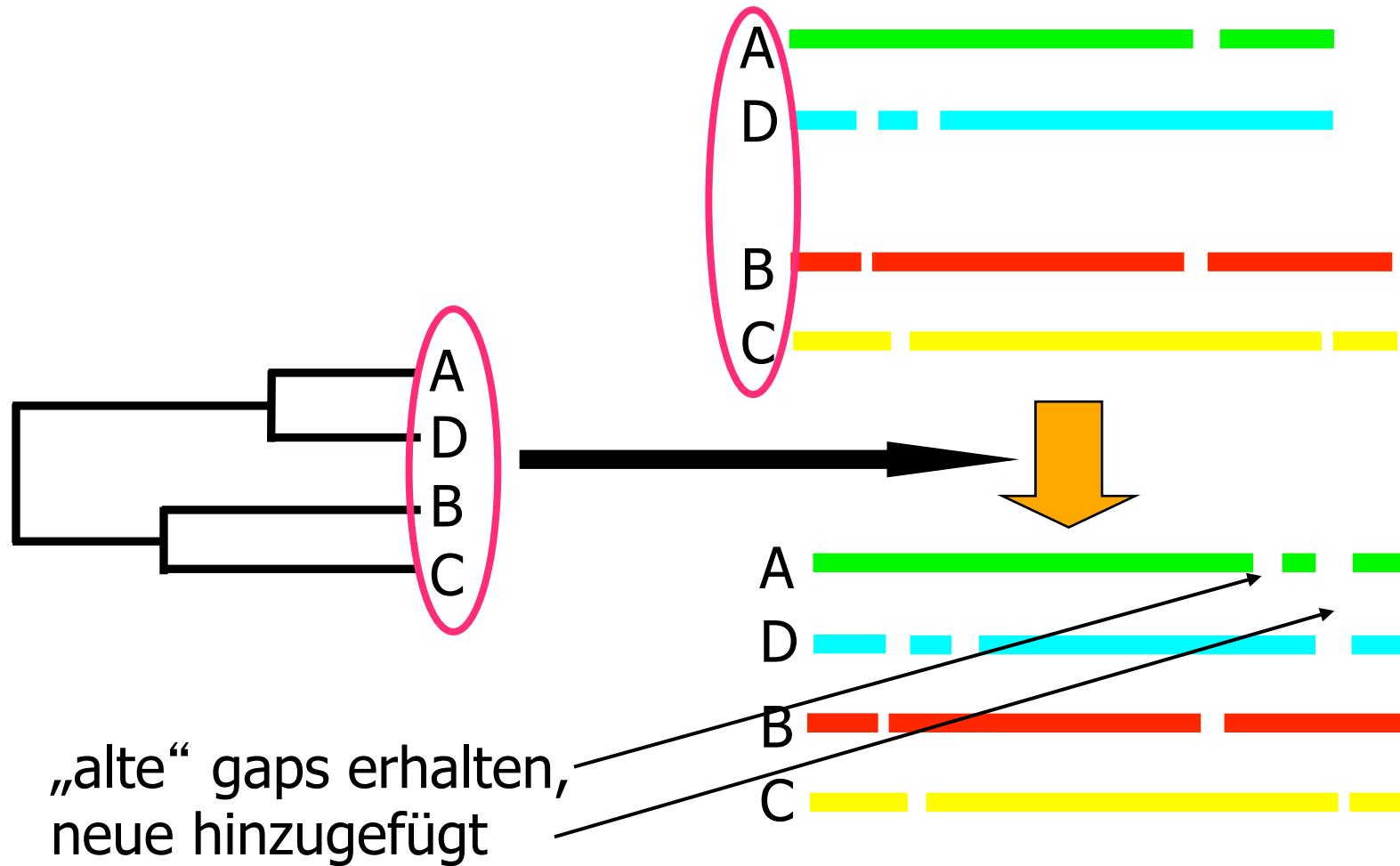
Progressives Alignment

3) Alignment von nahe verwandten Sequenzen; die ähnlichsten zuerst.



Progressives Alignment

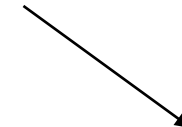
4) Sukzessives globales Alignment



Progressive Alignment

- Probleme -

- Welche Matrizen? Welche gap penalty?
- Erfolg hängt von den ersten paarweisen Alignments ab: problematisch, wenn selbst die nächsten Verwandten bereits unähnlich sind
- Fehler propagieren sich in das MSA
- Insertionen werden beim paarweisen Vergleich und beim MSA penalisiert: „Über-Alignment“ mit kompakten`ungapped blocks` und anderen Regionen mit gehäuften gaps



Praktische Hinweise für globale MSAs

- Keine sehr unterschiedlich langen Sequenzen global alignen: **Sequenzen trimmen** oder lokales MSA machen (z. B. PSI-BLAST)
- Protein-MSA: in Visualisierungsprogrammen (z.B. GeneDoc) unsichere Abschnitte auf **chemisch ähnliche As** hin absuchen. Falls nicht vorhanden, Abschnitte aus Alignment entfernen.
- Per Expertenwissen und 3D-Strukturinformation **Schlüsselpositionen im MSA identifizieren** und diese konstant halten

Praktische Hinweise für globale MSAs

Beispiel:

		*	20	*	40	*	60	*	80	
LluLegHb	-----	-----	-----	-----	-----	-----	-----	-----	-----	30
OsaHb2	-----	-----	-----	-----	-----	-----	-----	-----	-----	46
CelHb	MHCRNVYPIPKRSKSKKINFRCEPNEDEATSMTNAAAAAGGAKSKCKHF	-----	-----	-----	-----	-----	-----	-----	-----	80
HsaNGB	-----	-----	-----	-----	-----	-----	-----	-----	-----	30
DreNgb	-----	-----	-----	-----	-----	-----	-----	-----	-----	33
PcaMb	-----	-----	-----	-----	-----	-----	-----	-----	-----	31
HsaHBA	-----	-----	-----	-----	-----	-----	-----	-----	-----	32
HsaHBB	-----	-----	-----	-----	-----	-----	-----	-----	-----	31
BliHbB	-----	-----	-----	-----	-----	-----	-----	-----	-----	41
BviHb	-----	-----	-----	-----	-----	-----	-----	-----	-----	41

CD1

E7

F8

		*	100	*	120	*	140	*	160					
LluLegHb	VLEIAPAAKD	H	SFLKG-TSEVP--QN-NPELQA	H	GKVFKLVYEAAIQLOVVTG-VVVT	DATLKNLGSV	H	VS	SKG	---	VAD	102		
OsaHb2	IEVAPSASQ	H	SFLRN--SDVP--LEKNPKLKT	H	MSVFMVTC	EAAQLRKAGKVTVRD	TTTLKRLGAT	H	KEYG	---	VGD	119		
CelHb	VLTAQPDIAK	F	LEKI-PTGR--LKYDPRFRQ	H	LVYTKTLD	FVIRNLDYP--GKLE	VYFENL	GK	H	VAMQGRGFEP	153			
HsaNGB	LEALEPDLLP	F	CYNCRQFSSPEDCLS-SPEFLH	H	IRKVMLVIDAAV	NVEDL---SSLE	EYLASL	GR	H	HA	V	---	VKL	103
DreNgb	LFELDPALLT	H	SYSTNCGDAPE-CLS-SPEFLH	H	TKVMLVIDAAV	SHLDDL---HTLE	DFLLNL	GR	H	HA	V	---	VNT	105
PcaMb	LFKSHPETLE	H	IRFKHL-KT-EAEMKASEDLK	H	CVTLTALGAILKKKGH	---	---	---	---	---	---	---	KIPI	101
HsaHBA	MFLSFPTTKTY	H	EHF-DL-SH-----GSAQVK	H	CKKVADALTNAVAHVDD	---	---	---	---	---	---	---	RVDP	96
HsaHBB	LVVYPWTQRI	H	ESFGDL-ST-PDAVMGNPKVK	H	CKKVLGAFSDGLAHLDN	---	---	---	---	---	---	---	HVDP	101
BliHbB	LFKMEPGSKA	H	GRFGNI-DS---GMGRDKLR	H	ITLMYALQNFMDSLDNT	---	---	---	---	---	---	---	ISA	110
BviHb	LFSLHKDTAR	H	ERMGKL-DA--NNAGSNRELIC	H	ITLYLMYAIESFVDQLDDP	---	---	---	---	---	---	---	IGS	112



Praktische Hinweise für globale MSAs

Table 1

Summary of MSA programs that we consider to be the best currently available

Program	Advantages	Cautions
CLUSTALW	Uses less memory than other programs	Less accurate or scalable than modern programs
DIALIGN	Attempts to distinguish between alignable and non-alignable regions	Less accurate than CLUSTALW on global benchmarks
MAFFT, MUSCLE	Faster and more accurate than CLUSTALW; good trade-off of accuracy and computational cost. Options to run even faster, with lower average accuracy, for high-throughput applications.	For very large data sets (say, more than 1000 sequences) select time- and memory-saving options
PROBCONS	Highest accuracy score on several benchmarks	Computation time and memory usage is a limiting factor for large alignment problems (>100 sequences)
ProDA	Does not assume global alignability; allows repeated, shuffled and absent domains.	High computational cost and less accurate than CLUSTALW on global benchmarks
T-COFFEE	High accuracy and the ability to incorporate heterogeneous types of information	Computation time and memory usage is a limiting factor for large alignment problems (>100 sequences)

Clustal Omega, ClustalW and ClustalX Multiple Sequence Alignment

http://www.clustal.org/

UNIMAIL JGU Anmelden Google Phylogeny p... and system ilias 78 molgen UCSC LEO News ElektrZeitschr

 **Clustal: Multiple Sequence Alignment** 

Multiple alignment of nucleic acid and protein sequences



Clustal Omega

- Latest version of Clustal - fast and scalable (can align hundreds of thousands of sequences in hours), greater accuracy due to new HMM alignment engine
- Command line/web server only (GUI public beta available soon)



ClustalW/ClustalX

- "Classic Clustal"
- GUI (ClustalX), command line (ClustalW), web server versions available

Heuristische Verfahren

- Progressive Verfahren
- Iterative Verfahren
- Divide-and-Conquer-Verfahren
- Profile, Hidden Markov Modelle



...besonders gebräuchlich zur Erstellung von **lokalen MSAs**, mit denen z. B. Protein-Domänen definiert werden
(werden später in der VL besprochen)

Websites

- http://en.wikipedia.org/wiki/List_of_sequence_alignment_software
- Clustal: <http://www.ebi.ac.uk/clustalw/>
- MAFFT: <http://www.ebi.ac.uk/Tools/msa/>
- Editor für PC: <http://www.psc.edu/biomed/genedoc/>
- Editor für Mac/PC/Linux: <http://pbil.univ-lyon1.fr/software/seaview.html>