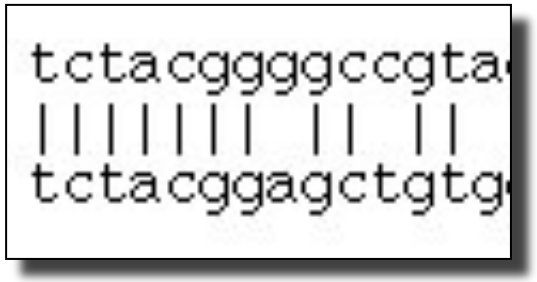
WS 2018/2019

„Genomforschung und Sequenzanalyse

- Einführung in Methoden der Bioinformatik- “

Thomas Hankeln



Alignment von DNA- und Proteinsequenzen

…das vielleicht wichtigste Werkzeug der Bioinformatik!

Wozu Alignment?

• sind zwei Gene/Proteine miteinander verwandt?
> Phylogenie & Evolution

• Finde ich ähnliche/verwandte Sequenzen (z. B. aus einem anderen Organismus) in den Datenbanken?
> „gene discovery“

• besitzt mein Protein funktionelle Abschnitte (Domänen), die man bereits von anderen Proteinen her kennt?
> Funktion, Annotation

Biologie ist eine komparative Wissenschaft!

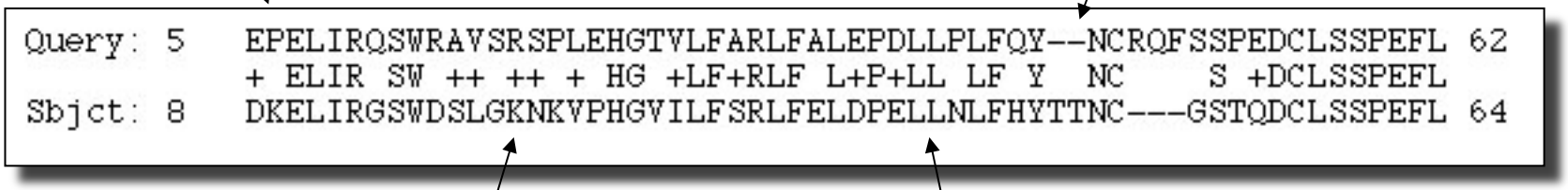
…nicht jedes Sprichwort stimmt!

...Bioinformatik & Genomforschung sind es auch!

Nt-Substitution

As-Austausch/ replacement

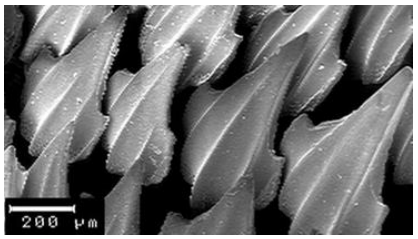
Gap bzw. InDel

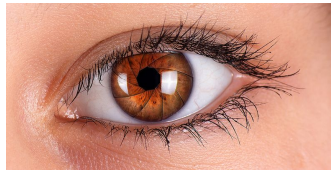


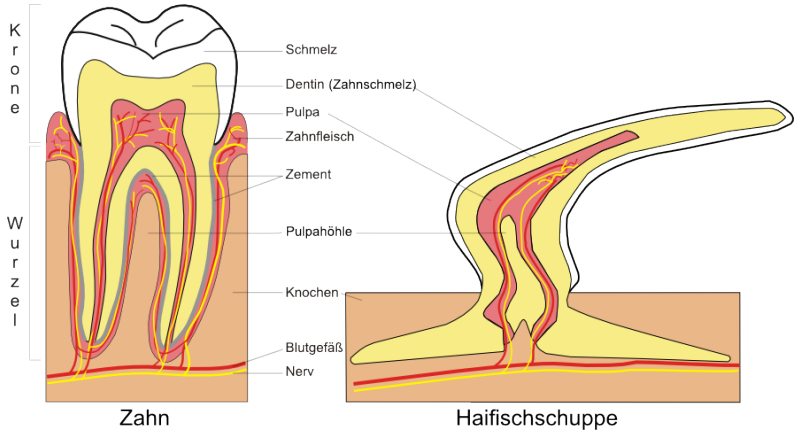
ähnliche As

identische As

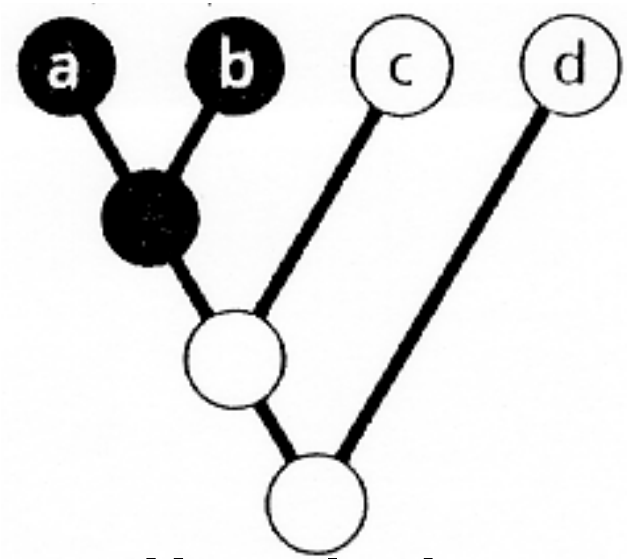
Welche Vergleiche mache Sinn?



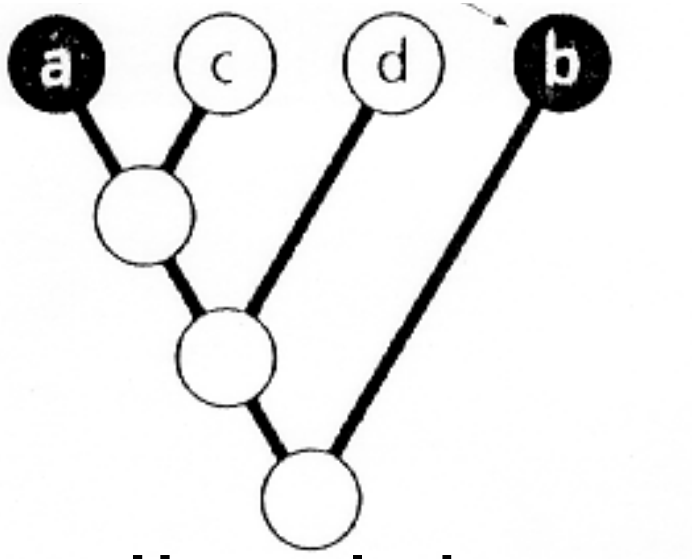




Homologie !!!



Homologie

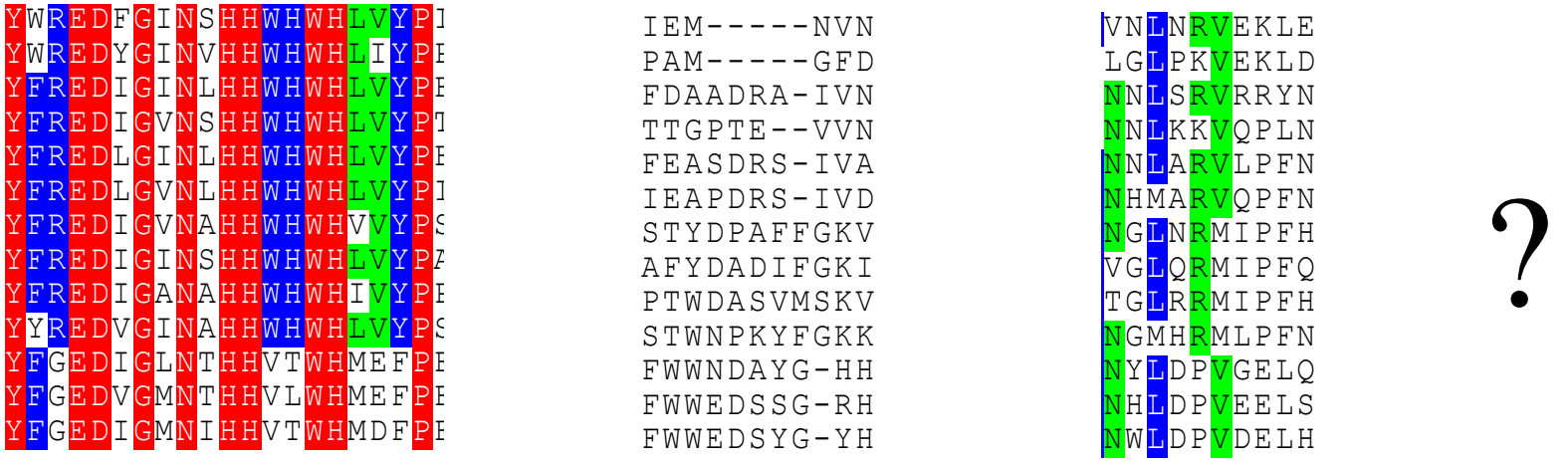


Homoplasie

• Merkmal ‚Schwarz‘ von gemeinsamem Vorläufer geerbt

• Merkmal ‚Schwarz‘ konvergent aus ‚weißen‘ Vorläufern entstanden

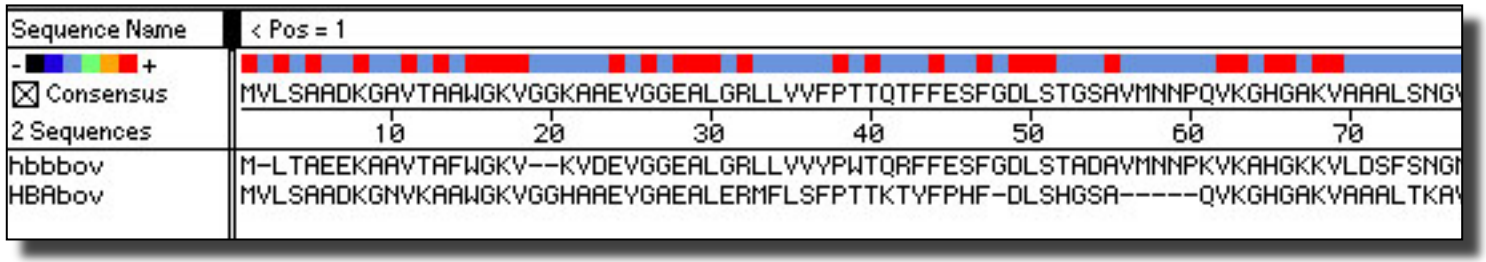
Welche Vergleiche mache Sinn?



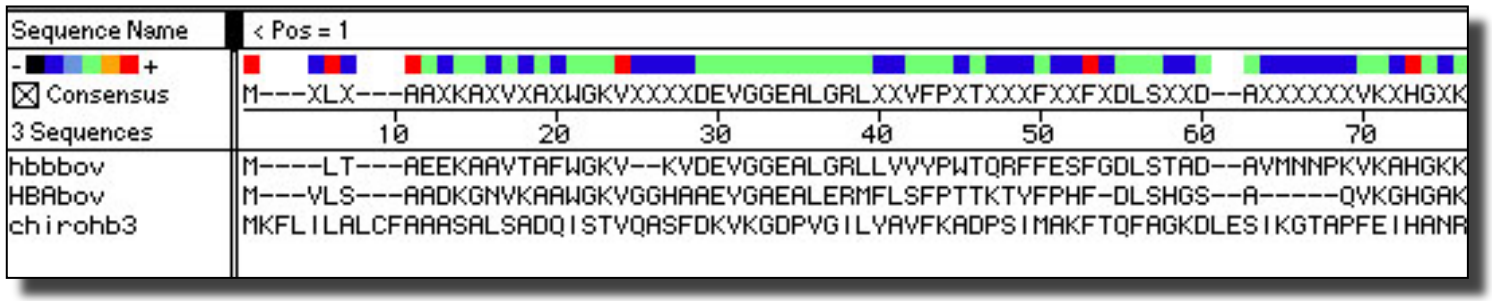
?

```
PlePPO     YWREDFGINSHHWHWHLVYPIEM-----NVNRDRKGELFYYMHQQMVARYDWERLSVNLNRVEKLE 61
PmoPPO     YWREDYGINVHHWHWHLIYPPAM-----GFDRDRKGELFYYMHQQVIARYDIERLCLGLPKVEKLD 61
BmoPPO1    YFREDIGINLHHWHWHLVYPFDAADRA-IVNKDRRGELFYYMHQQIIARYNVERMCNNLSRVRRYN 65
DmePPOA1   YFREDIGVNSHHWHWHLVYPTTGPTE--VVNKDRRGELFYYMHHQILARYNVERFCNNLKKVQPLN 64
DmePPO2    YFREDLGINLHHWHWHLVYPFEASDRS-IVAKDRRGELFYYMHQQVIARYNAERFSNNLARVLPFN 65
DmePPO3    YFREDLGVNLHHWHWHLVYPIEAPDRS-IVDKDRRGELFYYMHQQIIARYNAERLSNHMARVQPFN 65
EcaHcA     YFREDIGVNAHHWHWHVVYPSTYDPAFFGKVKDRKGELFYYMHQQMCARYDCERLSNGLNRMIPFH 66
EcaHcD     YFREDIGINSHHWHWHLVYPAFYDADIFGKIKDRKGELFYYMHQQMCARYDCERLSVGLQRMIPFQ 66
EcaHcF     YFREDIGANAHHWHWHIVYPPTWDASVMSKVKDRKGELFYYMHQQMCARYDCDRLSTGLRRMIPFH 66
LpoHc2     YYREDVGINAHHWHWHLVYPSTWNPKYFGKKKDRKGELFYYMHQQMCARYDCERLSNGMHRMLPFN 66
PvaHc      YFGEDIGLNTHHVTWHMEFPFWWNDAYG-HHLDRKGENFFWIHHQLTVRFDAERLSNYLDPVGELQ 65
PirHcC     YFGEDVGMNTHHVLWHMEFPFWWEDSSG-RHLDRKGESFFWVHHQLTVRYDAERLSNHLDPVEELS 65
PirHcA     YFGEDIGMNIHHVTWHMDFPFWWEDSYG-YHLDRKGELFFWVHHQLTARFDFERLSNWLDPVDELH 65
```

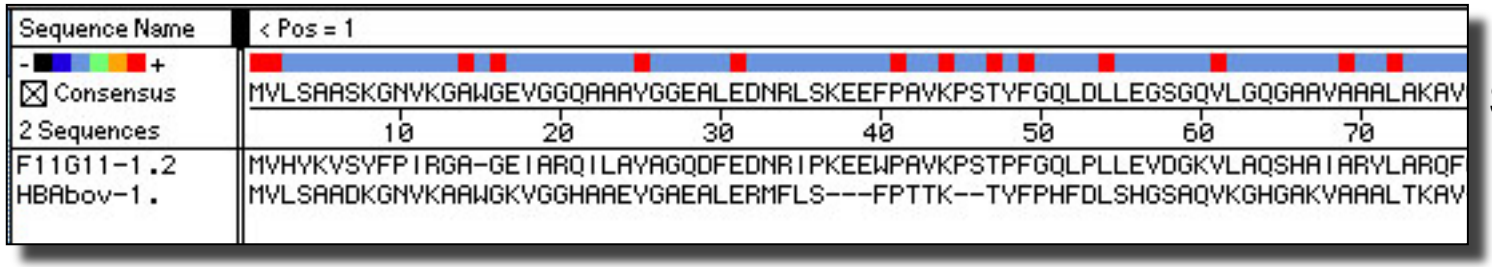
‚twilight zone‘



Hb b
Hb a



Hb b
Hb a
Mücken-Globin

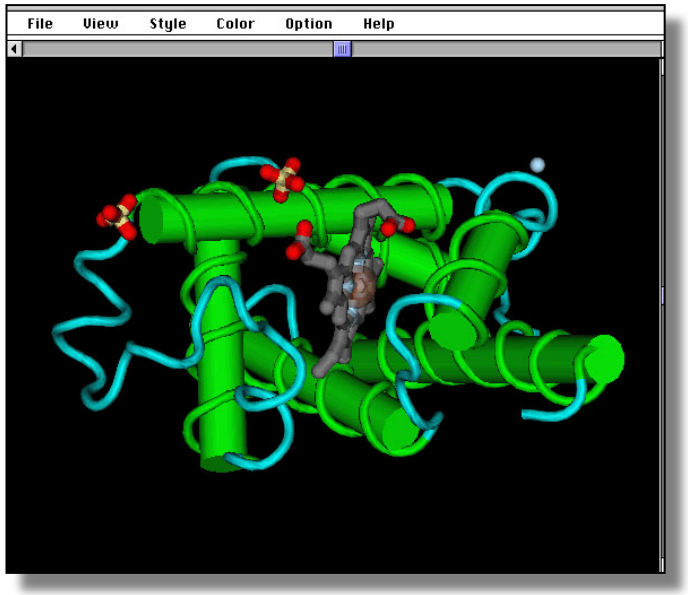


Glutathion-
S-Transferase

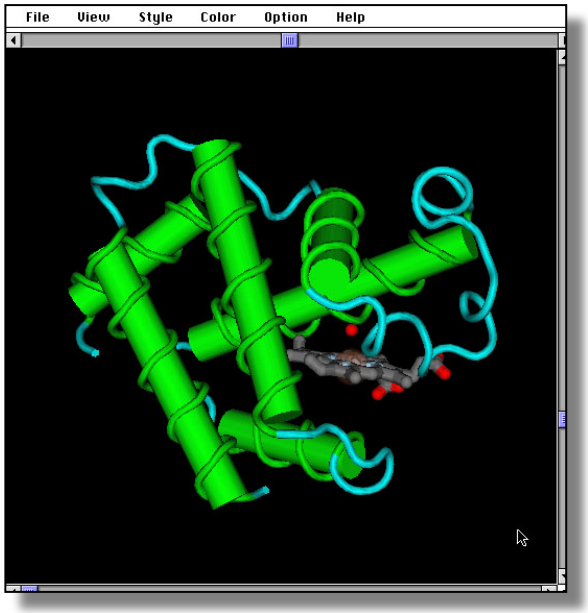
Hb a

Die interessantesten Erkenntnisse sind oft dort zu finden,
wo die Alignments schwierig und zweideutig sind

Sequenzen vs. Proteinstruktur



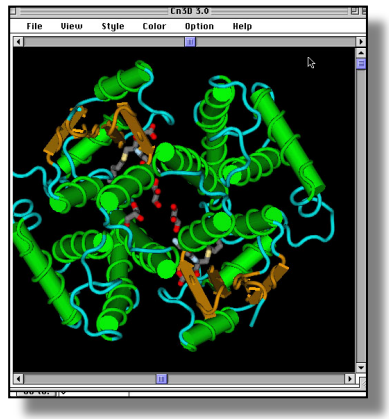
Myoglobin



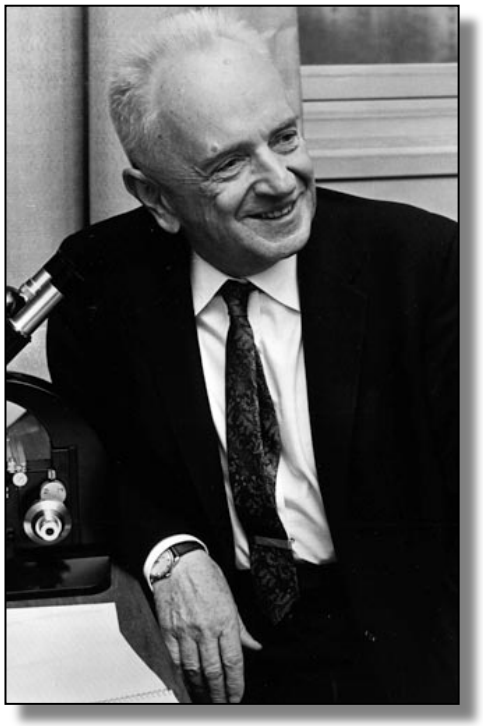
Mücken Globin

Trotz großer Veränderungen der Aminosäuresequenz kann die 3D-Struktur konserviert sein!

Aber:



Glutathion-S-Transferase

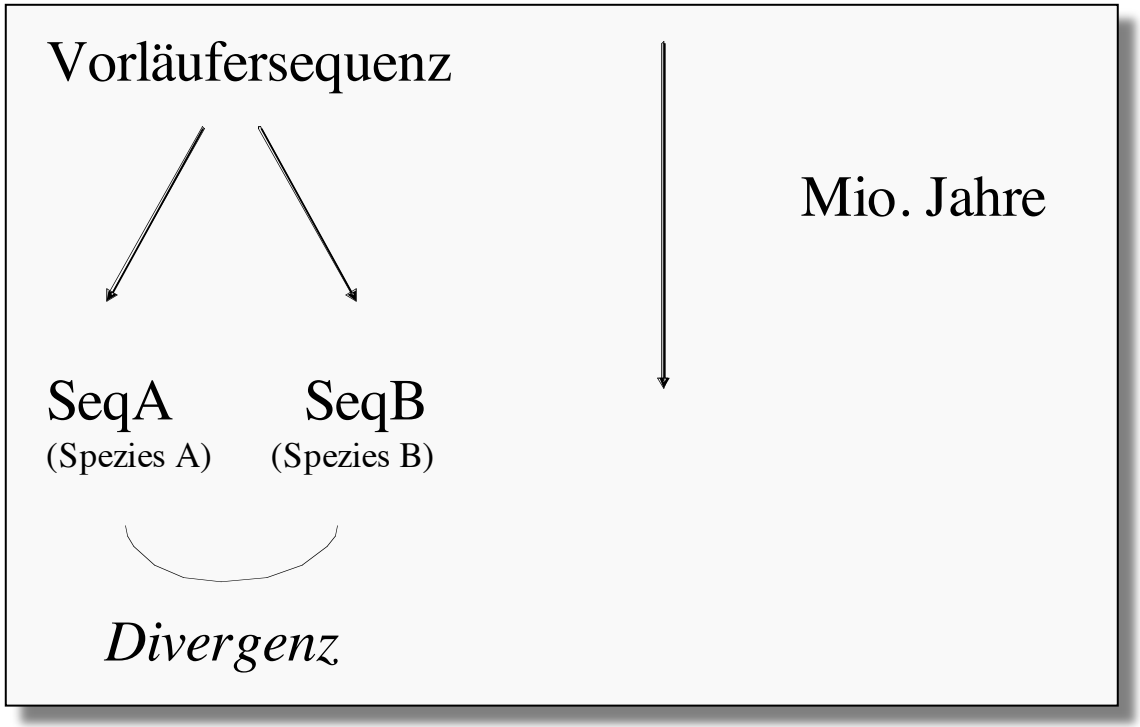


Nothing in Biology Makes Sense

Except in the Light of Evolution!

Theodosius Dobzhansky (1900-1975)

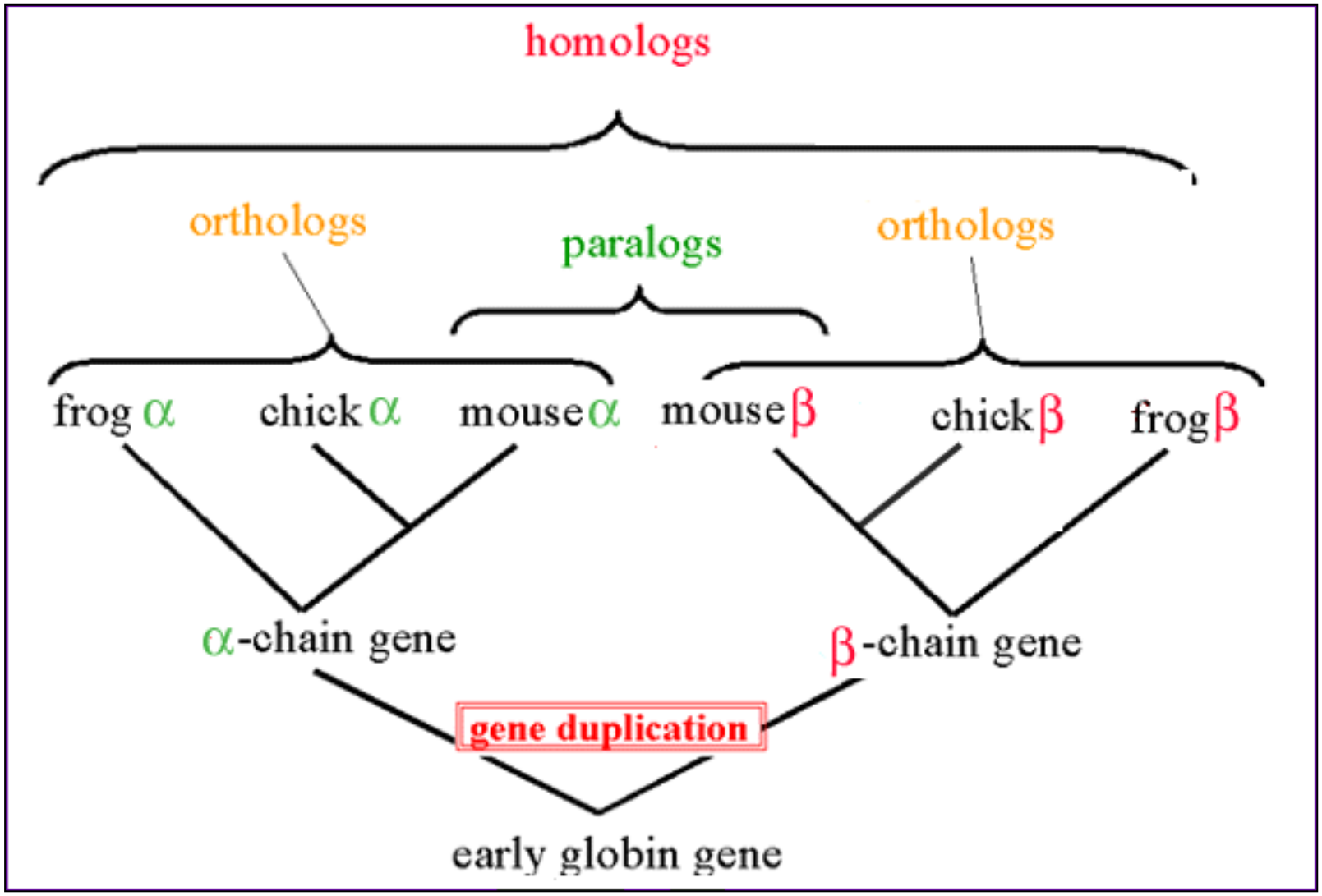
Alignments reflektieren die Evolution !!!!



Sind die zu alignierenden Sequenzen aus einer gemeinsamen Vorläufersequenz hervorgegangen?

Beim Erstellen und Bewerten von Alignments konstruieren wir evolutionäre Hypothesen!

Bei Genen: Paralog vs. Ortholog



Homologie, Identität, Ähnlichkeit

Beim Vergleich zwischen DNA-Sequenzen oder Proteinsequenzen sprechen wir zunächst immer von

- **Sequenzübereinstimmung (identity) oder**
- **Sequenzähnlichkeit (similarity)**

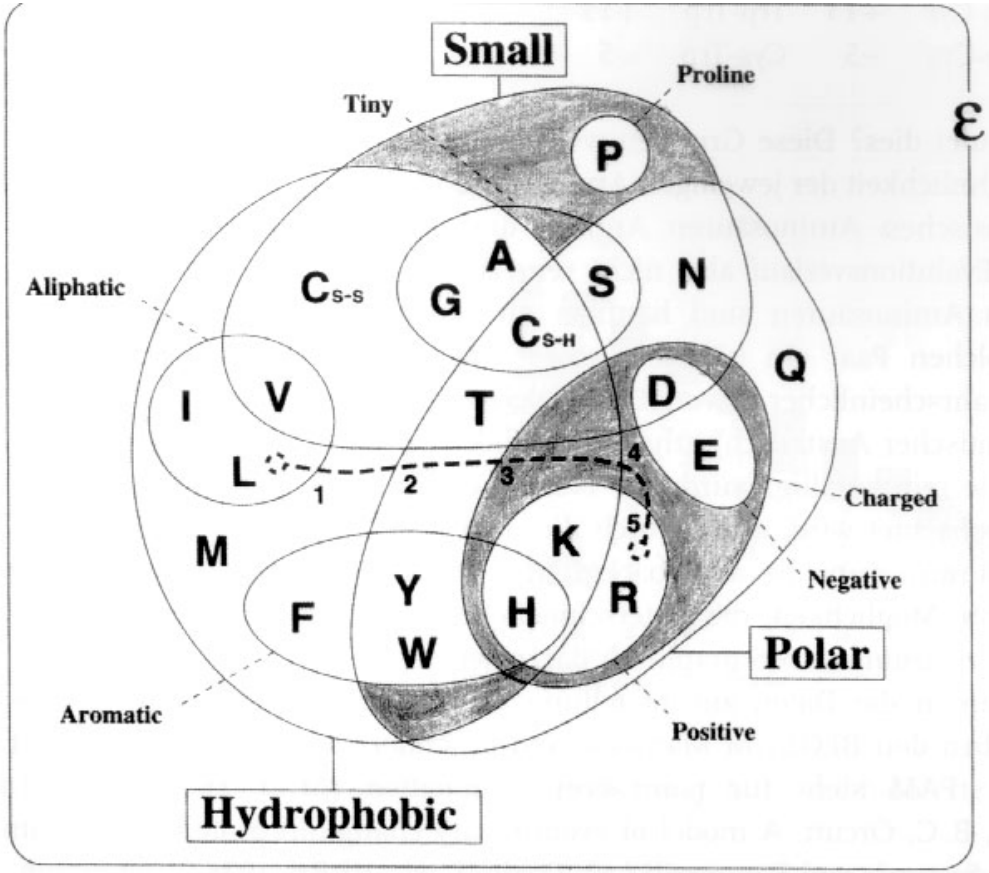
Erst aus diesem Vergleich heraus können wir überlegen, ob die gefundenen Übereinstimmungen wirklich **homolog** sind!

Protein-Sequenzen: Identität & Ähnlichkeit

```
 Score = 91.3 bits (223), Expect = 4e-18
 Identities = 59/156 (37%), Positives = 88/156 (55%), Gaps = 14/156 (8%)

Query: 4   MYKKILYPTDFSETAEIALKHVKAFKTLKAEEVILLHVIDEREIKKRDIFSLLLGVAGLN 63
           +YKKI+ PTD S+ +  A KH          EV  ++V+D          S  +G+
Sbjct: 25  LYKKIVIPTDGSDVSLEAAKHAINIAKEFDAEVYAIYVVD---------VSPFVGLPA-- 73

Query: 64  KSVEEFENELKNKLTEEAKNKMENIKKELEDVGFKVKDIIVVGIPHEEIVKIAEDEGVDI 123
           +   E  +EL   L EE +  ++ +KK  E+ G K+   ++ G+P  EIV+ AE +  D+
Sbjct: 74  EGSWELISEL---LKEEGQEALKKVKKMAEEWGVKIHTEMLEGVPANEIVEFAEKKKADL 130

Query: 124 IIMGSHGKTNLKEILLGSVTENVIKKSNKPVLVVKR 159
           I+MG+ GKT L+ ILLGSV E VIK ++ PVLVVK+
Sbjct: 131 IVMGTTGKTGLERILLGSVAERVIKNAHCPVLVVKK 166
```

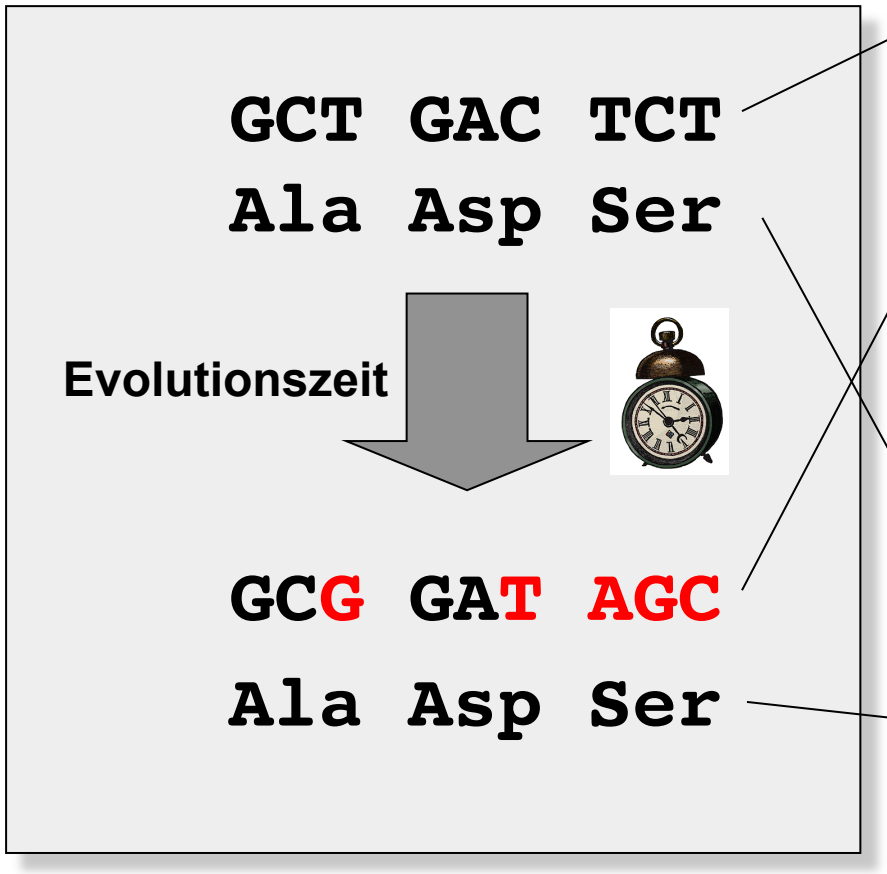
Bei Proteinsequenz-Alignments unterscheidet man **Sequenzidentität** und **Sequenzähnlichkeit** (= Identität plus iso-funktionelle As)

Funktionelle Ähnlichkeit von Amino-säuren



1.29 Dieses Diagramm ist eine recht anschauliche Weise, den unterschiedlichen Grad von Ähnlichkeit zwischen Aminosäuren zu verdeutlichen. Die 20 natürlichen Aminosäuren werden durch ein Set von 10 physiko-chemischen Eigenschaften beschrieben und den überlappenden Sektoren zugeordnet. So ist z. B. Lysin (K) positiv, geladen, polar und hydrophob. Zur Bewertung des Konservierungsgrades einer Position in einem Alignment wird deren Konservierungsnummer C_n berechnet: $C_n = 10 - P$, mit P als der Anzahl der Sektorengrenzen, die überschritten werden müssen, um alle Aminosäuren dieser Alignmentposition zu erreichen und 10 der Gesamtzahl von Eigenschaften. Ist z. B. eine Alignmentposition (Säule) nur von L und R besetzt, so müssen für eine solche Substitution fünf Sektorengrenzen überschritten werden: $C_n = 10 - P = 5$ (aus Livingstone, Barton, in: Methods in Enzymology, Vol. 266: *Computer Methods for Macromolecular Sequence Analysis*, R. F. Doolittle, ed., pp. 497–512. Abdruck mit Genehmigung von Academic Press).

Vergleiche ich auf DNA- oder auf Protein-Ebene?



DNA mutiert schnell:
„stille“ Mutationen sind „selektiv neutral“ und häufen sich an

Aminosäuren bleiben lange Zeit gleich („konserviert“):
Selektion auf Funktion!

Vergleiche ich auf DNA- oder auf Protein-Ebene?

GCT GAC TCT

Ala Asp Ser

Evolutionszeit

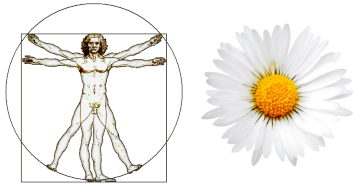
GCG GAT AGC

Ala Asp Ser

Konsequenz:

- Suche auf DNA-Ebene funktioniert gut zwischen nahe verwandten Taxa oder Genen

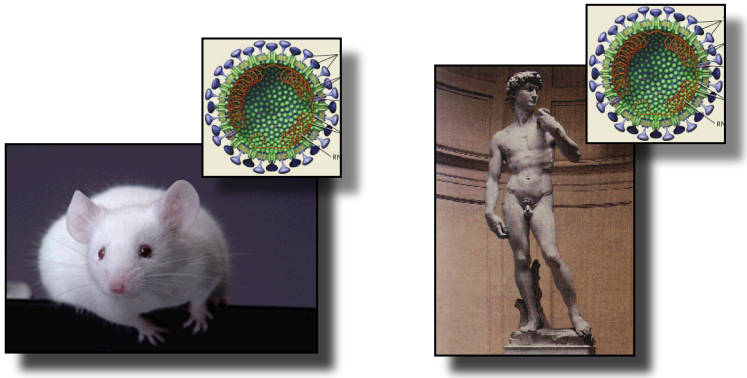
- Suche auf Aminosäureebene kann auch noch Ähnlichkeiten von entfernt verwandten Sequenzen detektieren



Alignment:

Wann DNA? Wann Protein?

Eng verwandte SARS-Varianten in der Population



Corona-Virus-Gruppen aus verschiedenen Spezies

Warum ist es nicht einfach, das „beste“ Alignment zu konstruieren?

- 2 Sequenzen à 300 Bp
  = 10^{88} mögliche Alignments!!!

- Computer-Algorithmen erforderlich, die
  ohne Ausprobieren aller Möglichkeiten auskommen.

- **„Regelwerk“ notwendig**, um bestmögliches
Alignment zu erkennen

Warum ist es nicht einfach, das „beste“ Alignment zu konstruieren?

```
seqA   TCAGACGATTG (11)
seqB   TCGGAGCTG (9)

I.     TCAG-ACG-ATTG
       TC-GGA-GC-T-G

II.    TCAGACGATTG
       TCGGAGCTG--

III.   TCAG-ACGATTG
       TC-GGA--GCTG
```

Annahmen über den Ablauf der Sequenz-Evolution:

I. Keine mismatches

II. Keine internen Lücken

III. „Von beidem Etwas“

Aber was ist richtig?

(manchmal)

…etwas einfacher geht‘s mit dem 20 As-Alphabet von Proteinen

Finde das optimale Alignment:

```
THIS IS A RATHER LONGER SENTENCE THAN THE NEXT
THIS IS A SHORT SENTENCE



THIS IS A RATHER LONGER - SENTENCE THAN THE NEXT
|||| || | --*|-- -|---| - |||||||| ---- --- ----
THIS IS A --SH-- -O---R T SENTENCE ---- --- ----
or
THIS IS A RATHER LONGER SENTENCE THAN THE NEXT
|||| || | ------ ------ |||||||| ---- --- ----
THIS IS A SHORT- ------ SENTENCE ---- --- ----
```

Warum ist ein „richtiges“ Alignment so problematisch?

- **Zwei beliebige Sequenzen lassen sich prinzipiell immer alignen!**
- **Es gibt viele mögliche Alignments**
- **Sequenz-Alignments müssen also in ihrer ‚Güte‘ bewertet werden, um das ‚optimale Alignment‘ zu finden**
- **Häufig wird es mehrere gleich gute Lösungen geben**

```
ACGTACGTACGTACGTACGTACGTACGT
   |    |    |    |    |    |    |
 GATCGATCGATCGATCGATCGATCGATC
```

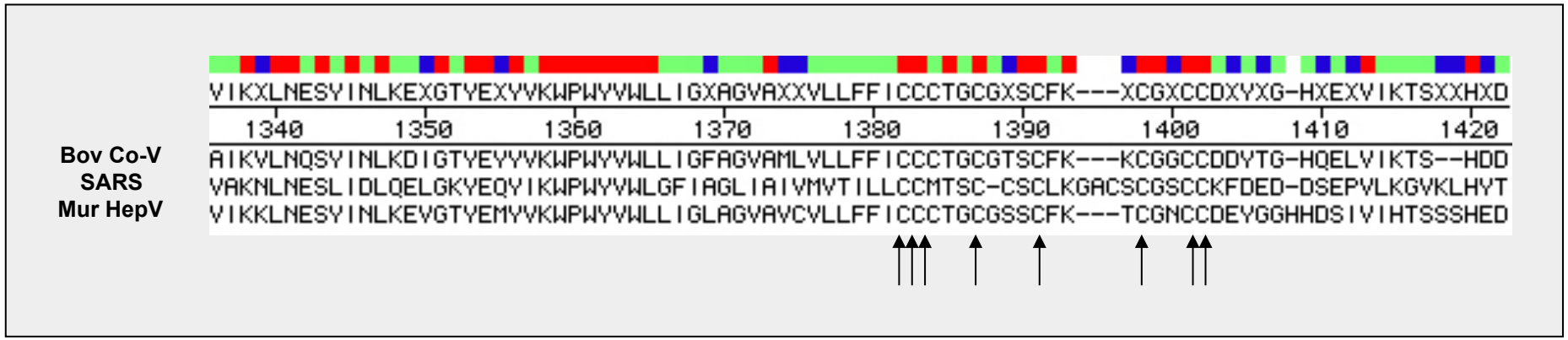
```
ACGTACGTACGTACGTACGTACGTACGT
     |    |    |    |    |    |
  GATCGATCGATCGATCGATCGATCGATC
```

Wie erstellt man ein möglichst „richtiges“ Alignment ?

Wir brauchen **„evolutionäre Modelle“,** um die beobachteten Sequenzveränderungen richtig zu bewerten:

- wie häufig mutiert ein A nach G bzw. nach C od. T (Transitionen : Transversionen)?
- wie häufig entstehen In/Dels relativ zu Substitutionen?
- wie häufig wird während der Proteinevolution eine Aminosäure durch irgendeine andere Aminosäure ersetzt?

Was bedeutet es, ein „Evolutionsmodell“ zu haben?



Ein ‚Evolutionsmodell‘ basiert auf empirischen Daten! Zum Beispiel:
Ich weiß, die Aminosäure Cystein ist für die Proteinstruktur äußerst wichtig!

- Cysteine sind also **konserviert** während der Evolution von Proteinen!
- Cysteine können daher beim Alignment zweier Proteinsequenzen als **Ankerpunkte** dienen
- ein Alignment mit übereinanderstehenden Cysteinen würde danach mit Pluspunkten **‚belohnt‘**

Ein Alignment ist .....

Es gibt kein „richtiges“ Alignment, sondern nur.....

Um die Qualität eines Alignments zu bewerten, brauche ich ....

Um zu verhindern, dass ein unsinniges Alignment gemacht wird, muss ich verhindern, dass .....

...zunächst zur Behandlung von Lücken!

Ein einfacher Score-Wert zur Bewertung eines Alignments…

$$S = Y - \sum W_k$$

S = Similarity-Score
Y = Anzahl an Matches

W_k = gap penalty für gaps der Länge k

Das Setzen einer Lücke wird durch einen negativen Score (gap penalty) bestraft!

Gap-Penalty

Mit Setzen der gap penalty trifft man Annahmen über die relative Häufigkeit von indel-Mutationen während der Evolution!

- gap opening penalty

  …Kosten für das Setzen einer Lücke

- gap extension penalty

  …Kosten für die Verlängerung einer Lücke

Gap-Penalty

- **‚lineare' gap penalty:**

$$W = (d\ g)$$ mit g = gap-Länge
und d = gap opening penalty

> nimmt an, dass gaps umso unwahrscheinlicher sind, je länger sie erscheinen (macht lange gaps unnötig „teuer")

- **‚affine' gap penalty:**

$$W = (d + g\ e)$$ mit e = gap extension penalty (e < d)

> lange Indels werden weniger bestraft als bei linearer Gap penalty. Man nimmt dann an, dass z. B. gaps von der Länge 1 oder der Länge 10 nicht drastisch unterschiedlich häufig während der Evol. entstehen

Auswirkungen der gap penalty

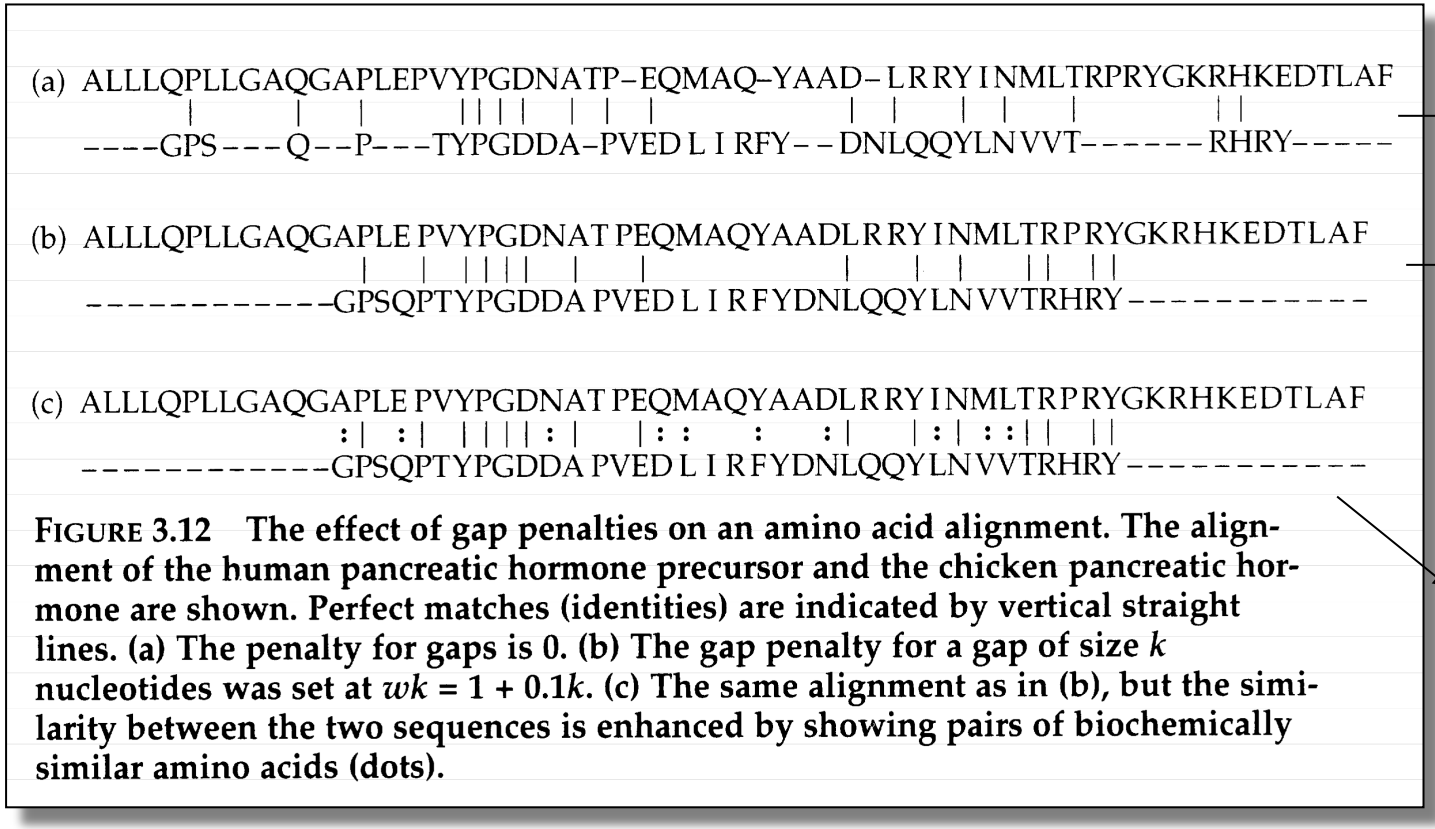


FIGURE 3.12 The effect of gap penalties on an amino acid alignment. The alignment of the human pancreatic hormone precursor and the chicken pancreatic hormone are shown. Perfect matches (identities) are indicated by vertical straight lines. (a) The penalty for gaps is 0. (b) The gap penalty for a gap of size k nucleotides was set at $wk = 1 + 0.1k$. (c) The same alignment as in (b), but the similarity between the two sequences is enhanced by showing pairs of biochemically similar amino acids (dots).

→ Penalty = 0

→ Penalty $w_k=1+0.1k$

Anzeigen der biochemisch verwandten As macht deutlich, daß das Alignment (b) Sinn macht

…und jetzt zu den Austauschen!

- in sog. „Substitutionsmatrizen“ wird die relative Häufigkeit erfasst, mit der Nukleotide oder Aminosäuren während der Evolution ausgetauscht werden.

Eine einfache Identitätsmatrix bei Nukleotidsequenzen...

	A	C	G	T
A	1			
C	0	1		
G	0	0	1	
T	0	0	0	1

- alle Richtungen von Nt-Austauschen sind gleich wahrscheinlich

- bei jedem „match“ beider Sequenzen gibt es 1 Punkt für den Übereinstimmungs-Score

DNA-Alignment-Bewertung

```
seqA   TCAGACGATTG (11)
seqB   TCGGAGCTG (9)

I.     TCAG-ACG-ATTG
       TC-GGA-GC-T-G

II.    TCAGACGATTG
       TCGGAGCTG--

III.   TCAG-ACGATTG
       TC-GGA--GCTG
```

D = 7 – 6(3+1x0.1) = - 11.6

D = 4 – (3+2x0.1) = **+ 0.8**

D = 6 – 2(3+1x0.1) – (3+2x0.1)
= - 3.4

Match = +1

Gap-Parameter:

d = **3** (gap opening)

e = 0.1 (gap extension)

Bei hoher gap opening penalty!

DNA-Alignment-Bewertung

```
seqA   TCAGACGATTG (11)
seqB   TCGGAGCTG (9)

I.     TCAG-ACG-ATTG
       TC-GGA-GC-T-G

II.    TCAGACGATTG
       TCGGAGCTG--

III.   TCAG-ACGATTG
       TC-GGA--GCTG
```

?

Match = +1

Gap-Parameter:

d = 1 (gap opening)

e = 0.1 (gap extension)

Bei niedriger gap opening penalty!

Substitutions-Matrizen für Proteine

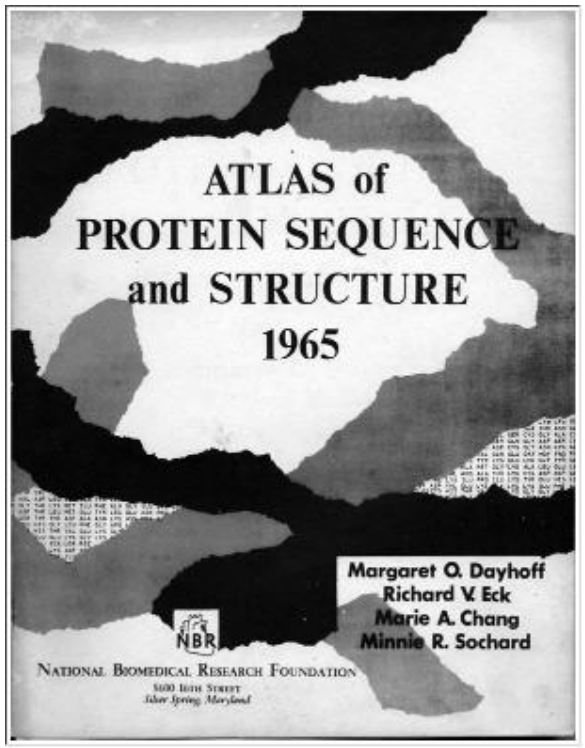
• chemisch-funktionelle Ähnlichkeit der As bestimmt Wahrscheinlichkeit eines Austauschs während der Evolution. Daher...

• ...sind die „Kosten“ bzw. die „Belohnung“ für bestimmte Austausche unterschiedlich hoch!

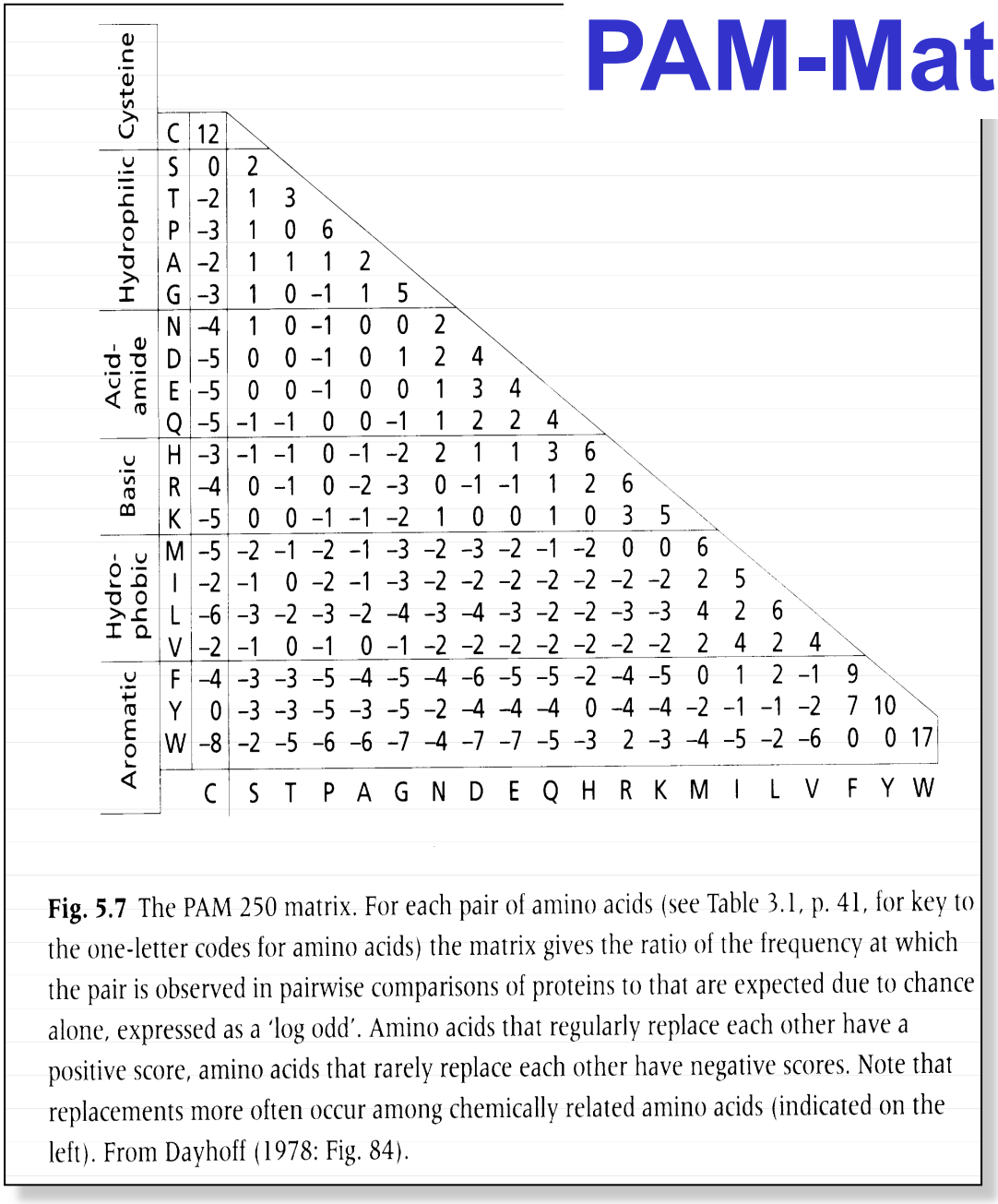
• Definition der Kosten erfolgt über Matrizen:

> **PAM-Matrizen** (Dayhoff 1978)
> **BLOSUM-Matrizen** (Henikoff & Henikoff 1992)
u. einige mehr

Margaret O. Dayhoff
1925-1983



PAM-Matrizen



		C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
Cysteine	C	12																			
Hydrophilic	S	0	2																		
	T	−2	1	3																	
	P	−3	1	0	6																
	A	−2	1	1	1	2															
	G	−3	1	0	−1	1	5														
Acid-amide	N	−4	1	0	−1	0	0	2													
	D	−5	0	0	−1	0	1	2	4												
	E	−5	0	0	−1	0	0	1	3	4											
	Q	−5	−1	−1	0	0	−1	1	2	2	4										
Basic	H	−3	−1	−1	0	−1	−2	2	1	1	3	6									
	R	−4	0	−1	0	−2	−3	0	−1	−1	1	2	6								
	K	−5	0	0	−1	−1	−2	1	0	0	1	0	3	5							
Hydro-phobic	M	−5	−2	−1	−2	−1	−3	−2	−3	−2	−1	−2	0	0	6						
	I	−2	−1	0	−2	−1	−3	−2	−2	−2	−2	−2	−2	−2	2	5					
	L	−6	−3	−2	−3	−2	−4	−3	−4	−3	−2	−2	−3	−3	4	2	6				
	V	−2	−1	0	−1	0	−1	−2	−2	−2	−2	−2	−2	−2	2	4	2	4			
Aromatic	F	−4	−3	−3	−5	−4	−5	−4	−6	−5	−5	−2	−4	−5	0	1	2	−1	9		
	Y	0	−3	−3	−5	−3	−5	−2	−4	−4	−4	0	−4	−4	−2	−1	−1	−2	7	10	
	W	−8	−2	−5	−6	−6	−7	−4	−7	−7	−5	−3	2	−3	−4	−5	−2	−6	0	0	17

Fig. 5.7 The PAM 250 matrix. For each pair of amino acids (see Table 3.1, p. 41, for key to the one-letter codes for amino acids) the matrix gives the ratio of the frequency at which the pair is observed in pairwise comparisons of proteins to that are expected due to chance alone, expressed as a 'log odd'. Amino acids that regularly replace each other have a positive score, amino acids that rarely replace each other have negative scores. Note that replacements more often occur among chemically related amino acids (indicated on the left). From Dayhoff (1978: Fig. 84).

- „Percent Accepted Mutation“
  („accepted during evolution“)
- 1 PAM entspricht 1% As-Aus-tausch
- positiver Wert = Aminosäuren, die sich häufig in Alignments gegenüberstehen und somit ‚funktionell konserviert' sind

z. B. W-W 17
C -C 12
P – P 6

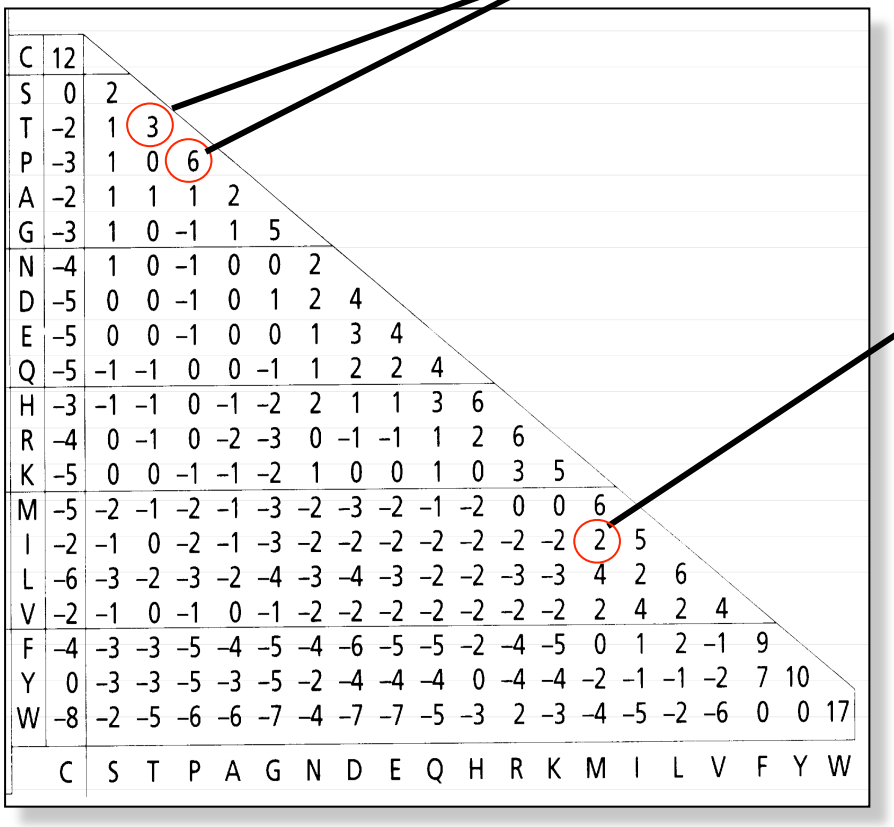
aber W-V - 6

Bewertung eines As-Alignments

Sequenz 1

Sequenz 2

```
PTHPLASKTQILPEDLASEDLTI
||||||     |      ||  ||
PTHPLAGERAIGLARLAEEDFGM
```



	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	12																			
S	0	2																		
T	-2	1	3																	
P	-3	1	0	6																
A	-2	1	1	1	2															
G	-3	1	0	-1	1	5														
N	-4	1	0	-1	0	0	2													
D	-5	0	0	-1	0	1	2	4												
E	-5	0	0	-1	0	0	1	3	4											
Q	-5	-1	-1	0	0	-1	1	2	2	4										
H	-3	-1	-1	0	-1	-2	2	1	1	3	6									
R	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6								
K	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5							
M	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6						
I	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5					
L	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6				
V	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4			
F	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9		
Y	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10	
W	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17

```
P:P = +6
T:T = +3
...
I:M = +2


Score   =
6+3+...+2 = XX
```


Verändert nach Folie von Stefan Wiemann, DKFZ

Erstellung einer PAM -Matrix

1. Vergleiche problemlos zu alignierende, nahe verwandte Proteinsequenzen (>1572 Austausche in 71 Sequenzgruppen mit > 85% Sim.)
2. Kalkuliere die Austauschwahrscheinlichkeit jeder As relativ zu ihrer Häufigkeit in den Sequenzen und zur allgemeinen Mutabilität der Sequenzen
3. Berechne Score-Wert für As als „log odds“-Wert, d.h. als Logarithmus der Wahrscheinlichkeiten:

$$S(a,b) = \log \frac{p_{ab}}{q_a q_b}$$

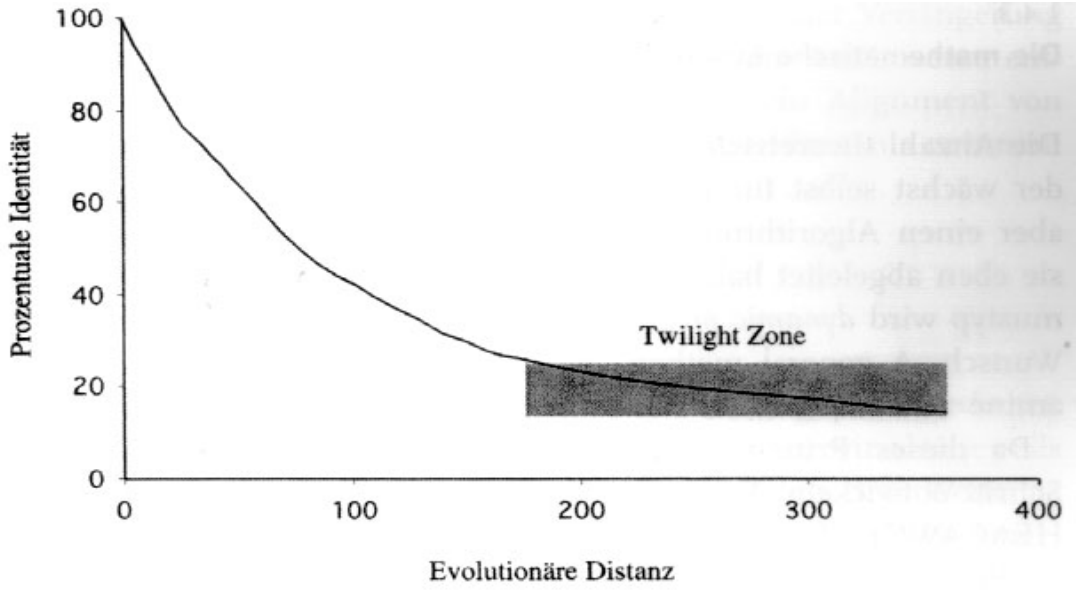
joint probab.

Background frequency

4. „Log odds“-Scores können bei der Bewertung des Alignments anhand der Matrix einfach addiert werden

Erstellung einer PAM -Matrix

- Achtung: PAM 60/80/120/250 -Matrizen wurden durch Multiplikation der PAM 1 mit sich selbst extrapoliert!
- PAM 250 bedeutet, dass man durch multiple Austausche an derselben Position durchaus 250% As-Austausche erwarten kann. Dennoch haben diese entfernt verwandten Sequenzen noch bis 20% As-Ähnlichkeit:



PAM - Matrizen in der Kritik

1. zu kleiner Proteindatensatz bei Dayhoff:

   die Original-PAMs basieren auf wenigen Proteinfamilien!
   > besser: JTT-Matrizen nehmen 1991er-Datenbank

2. Extrapolationsverfahren:

   die gebräuchlichen PAM100 oder PAM250 Matrizen sind bedingt durch die Extrapolation nur Vorhersagen

3. Weitere unzulässige Annahmen:

   „Alle As-Positionen evolvieren unabhängig“; „Alle Positionen mutieren mit gleicher Wahrscheinlichkeit“

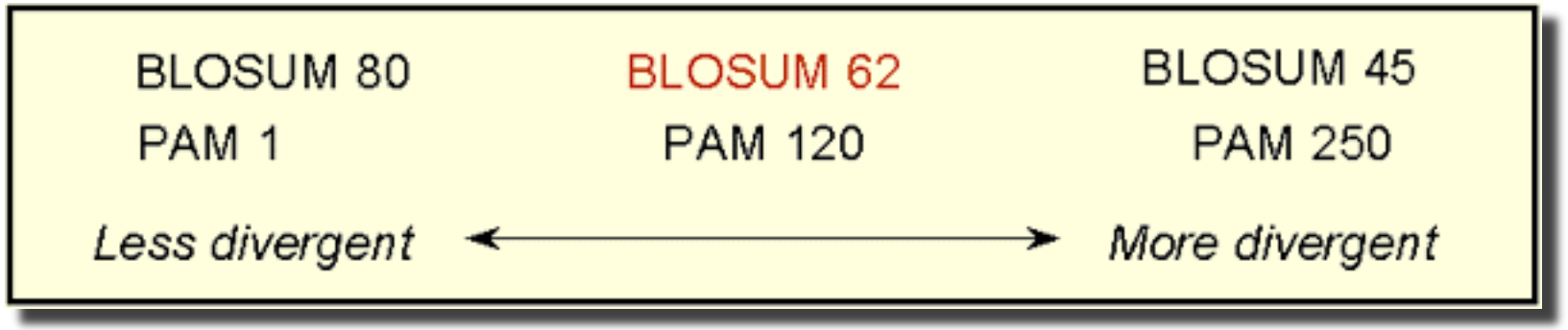
Henikoff & Henikoff 1992

BLOSUM - Matrizen

- Blocks Substitution Matrix

- Scores errechnet nach beobachteten Austauschfrequenzen in „Blöcken“ aus lokalen Alignments auch z.T. entfernt sequenzverwandter, jedoch klar biochemisch verwandter Proteine (> 500 versch. Familien!)
- z. B. BLOSUM 62 : Scores abgeleitet von Sequenzen mit höchstens 62% As-Identität
- BLOSUM-Matrizen-Werte sind nicht extrapoliert u. nicht abhängig von einem evolutionären Modell.
- besonders geeignet für Alignment entfernt verwandter Proteinsequenzen

BLOSUM vs. PAM



PAM	60	für 60% ähnliche Proteine
	80	50%
	120	40%

Achtung: die meisten Matrizen in Vergleichsprogrammen haben assoziierte (und oft auch optimierte) Gap penalty-Scores! Vorsicht bei drastischen Änderungen der gap penalty-Werte relativ zu den Substitutions-Scores.

Wir haben also Kriterien (Substitutionsmatrizen, gap penalties), um Alignments zu bewerten.

Aber wie werden Alignments überhaupt erstellt?

Bisher nur „Regeln“ für Alignment besprochen. Nun endlich...

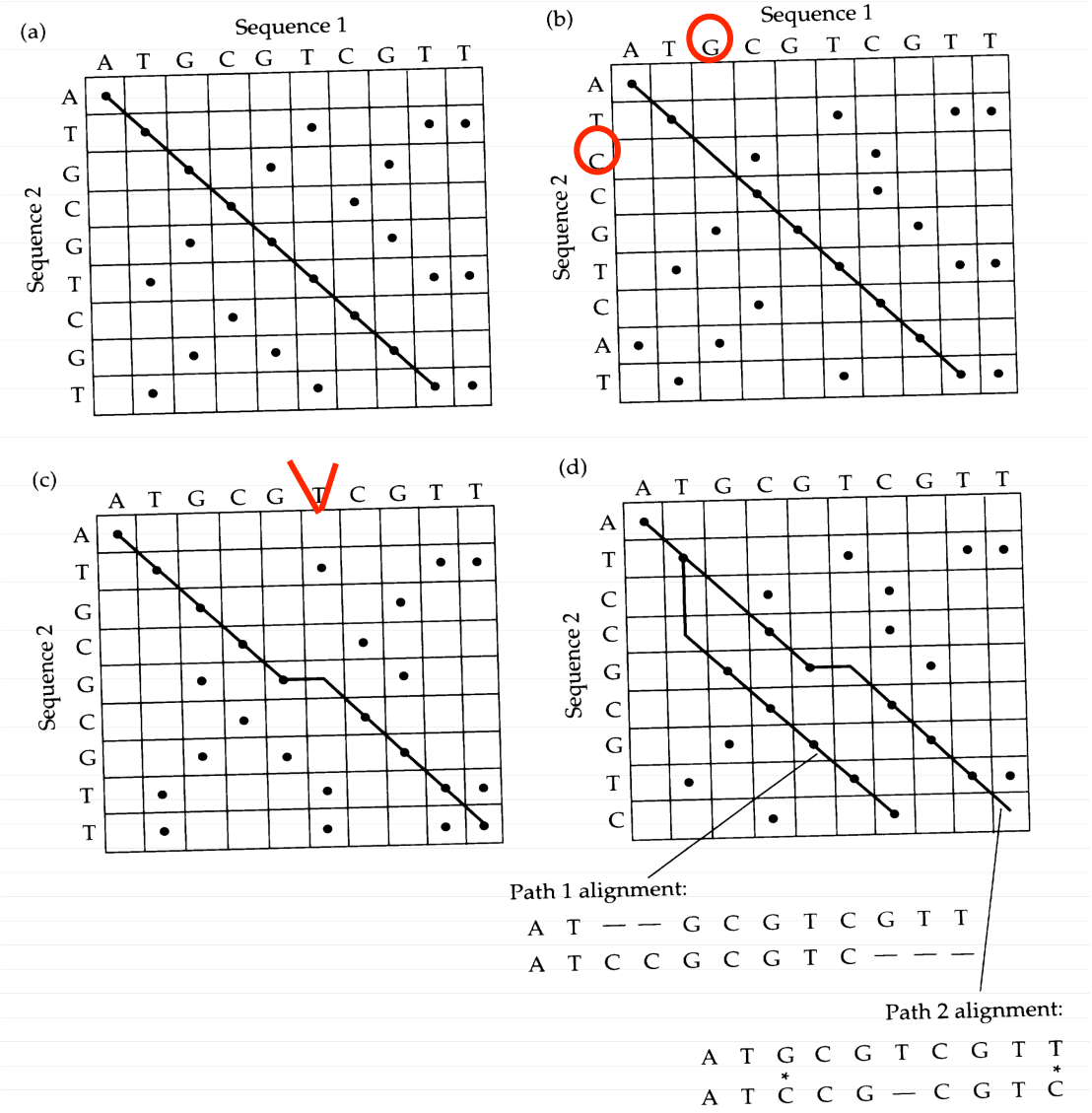
Alignment-Methoden

- Dot-Matrix-Vergleich (Grafik)
- algorithmischer Vergleich

  „dynamic programming“
  „word/k-tuple“-Methoden

Vergleich von DNA per Dot-Plot

Gibbs, McIntyre 1970

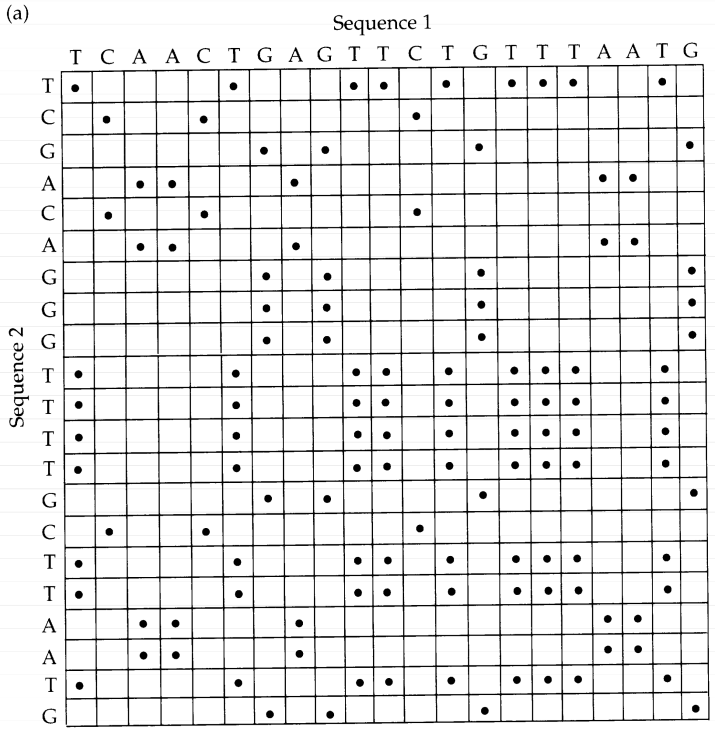


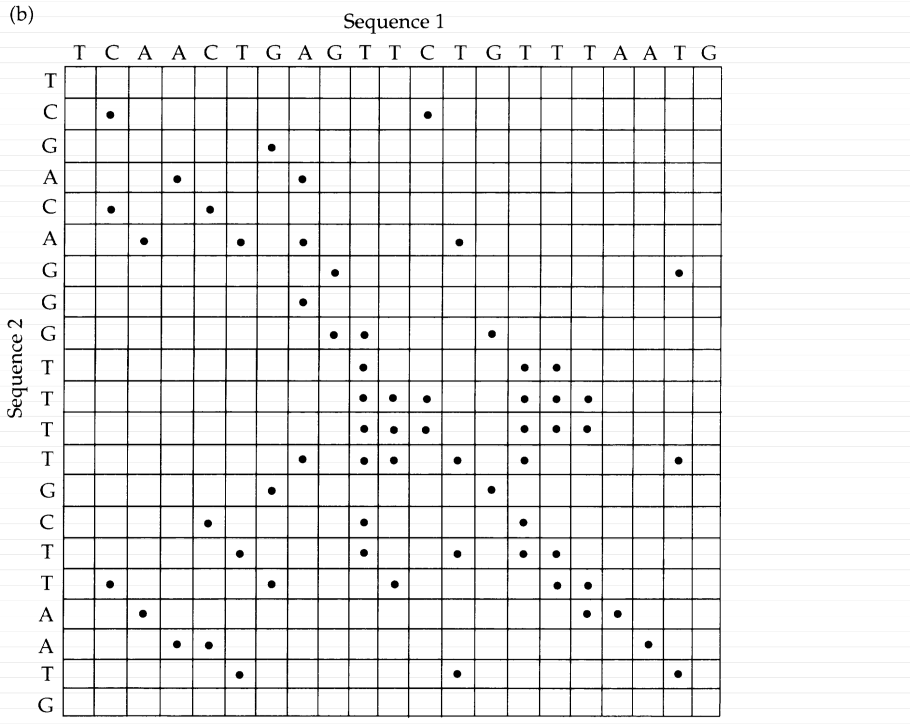
Dot plot Programm:
http://bioweb.pasteur.fr/seqanal/interfaces/dottup.html

Dot-Plot mit ‚sliding window‘

Problem:
Diagonale ist versteckt

Lösung:
„sliding window“als Filter
3 nt window size, 2/3 müssen passen



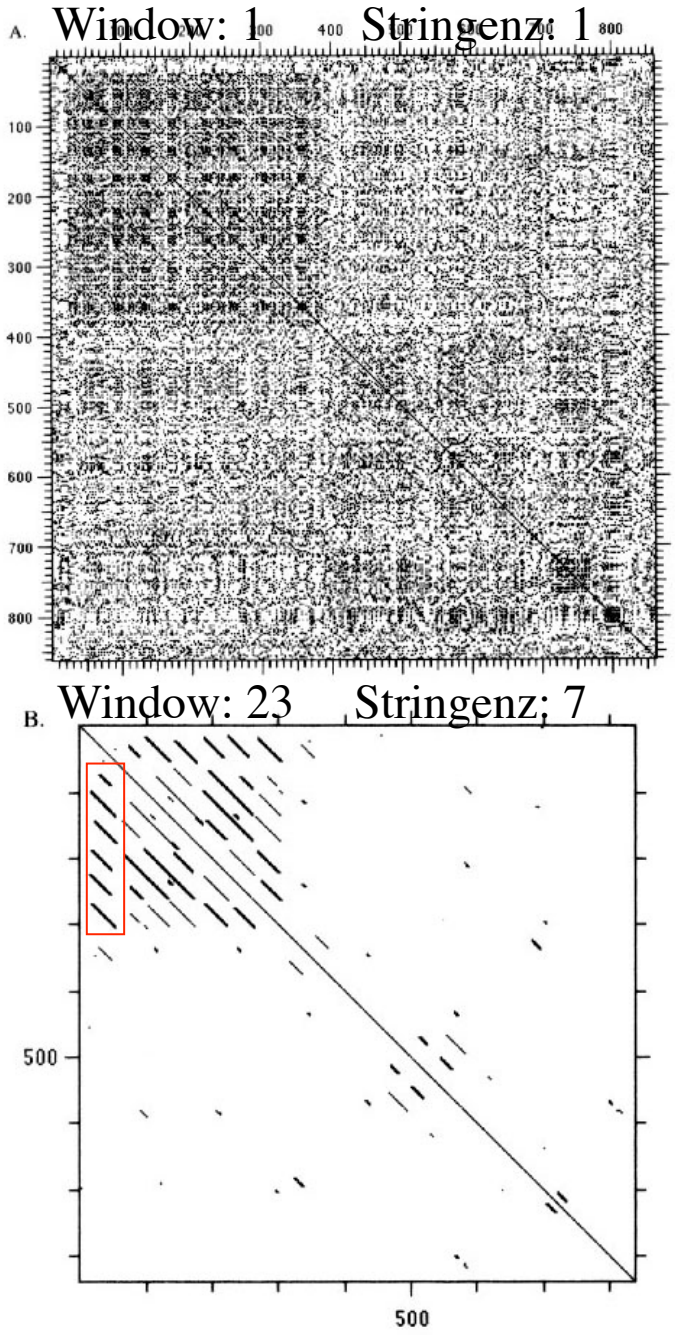


Dot-Plot mit ‚sliding window‘

Je kleiner das Fenster, desto größer die Auswirkung zufälliger Matches.

Große Fenster sind nicht gut geeignet für Entdeckung kurzer Matches.

Stringenz bestimmt sichtbares Ergebnis in hohem Maße.



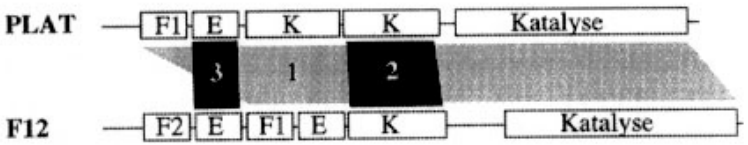
Dot-Plot zeigt Repetitionen

- LDL-Rezeptorgen mit sich selbst verglichen

- kurze Diagonalen abseits der Hauptdiagonalen zeigen repetitive Sequenzregionen an!

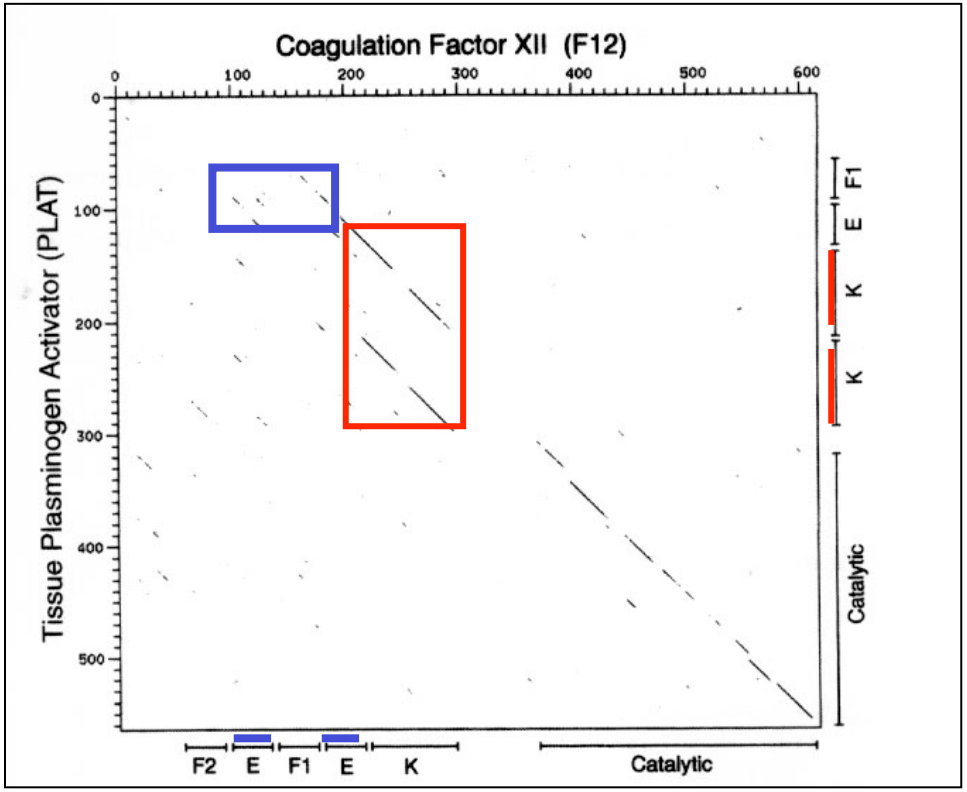
- bei Erhöhung der Stringenz auf 15/23 verschwinden die zusätzlichenDiagonalen:
> die Repetitionen sind nur bis zu einem gewissen Grad ähnlich

Dot-Plot zeigt Protein-Domänen

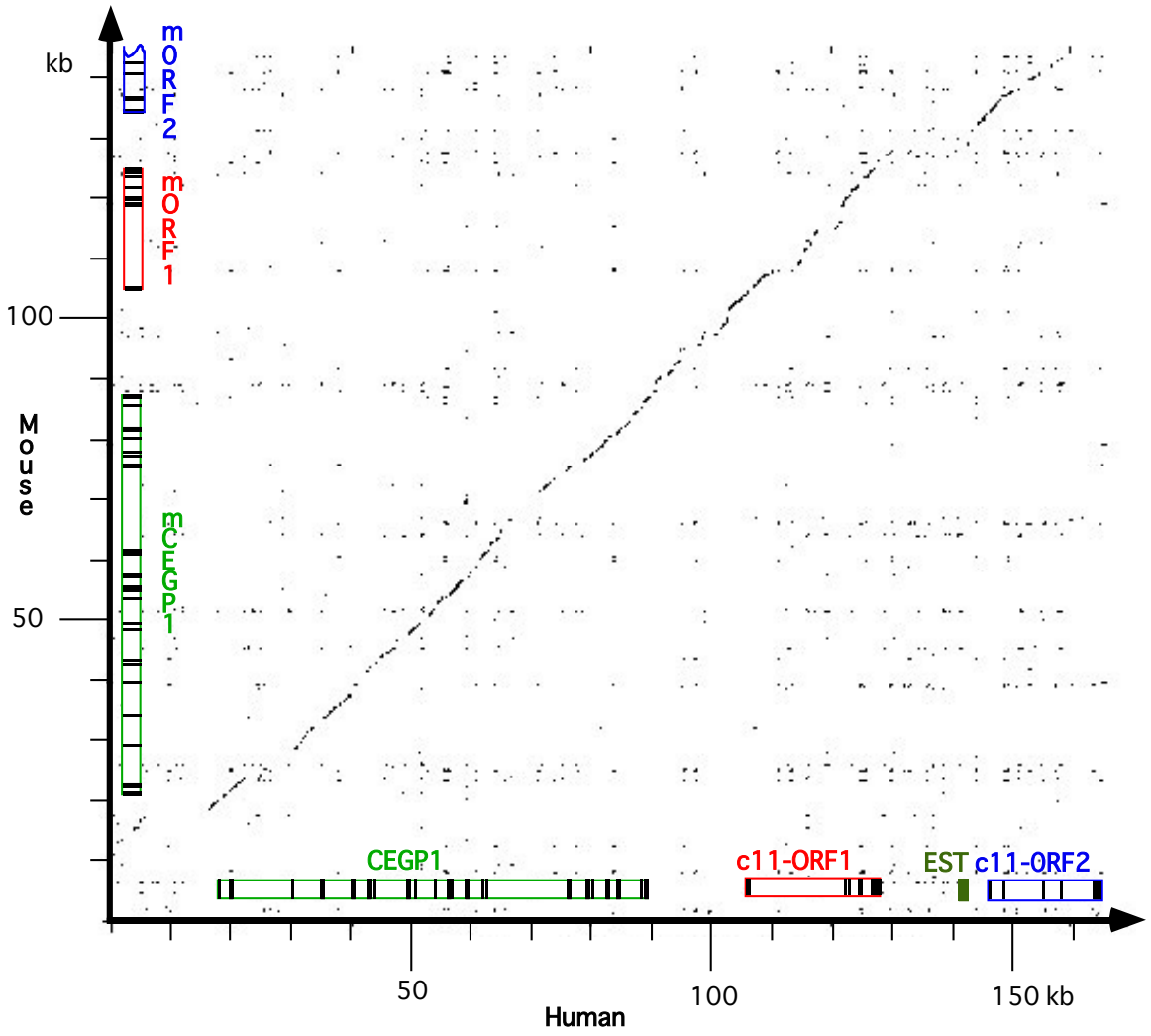


1.37 Die menschlichen Proteine PLAT (T-Plasminogen Activator) und F12 (Coagulationsfaktor XII) benutzen die gleichen 'Bausteine' in unterschiedlicher Anzahl und Reihenfolge. Neben dem optimalen lokalen Alignment (Bereich F1-Domäne bis N-Terminus) müssen die *zusätzlichen* lokalen Alignments 2 (verdoppelte K-Domäne in PLAT) und 3 (verdoppelte E-Domäne in F12) erkannt werden, um die Domänenstruktur beider Proteine zu verstehen (s. a. Abb. 1.36).

Dot-Plots ermöglichen Entdeckung repetitiver Domänen in Proteinen



Dotplot zeigt syntäne Genombereiche



Mensch-Maus-Genomvergleich über ca. 150 kb

Warum ist ein „richtiges“ Alignment so problematisch?

- 2 x 300 Bp = 10^{88} mögliche Alignments!!!
- Computer-Algorithmen erforderlich, die ohne ausführliche Suche auskommen.

Alignment-Algorithmen:

Globales vs. Lokales Alignment

- **globales** Alignment: z. B. Needleman-Wunsch
  Nachteil: sehr aufwendig und langsam, aber komplettes alignment bis zu den Enden

```
ATTGTCCATGCAGCCTGAA
TATCGGGATGC--CTTATT
```

- **lokales** Alignment: z. B. Smith-Waterman
  schnell, aber nur die lokal passenden Teile werden alignt, Lücken v. a. an den Enden sind nicht dargestellt

```
-------ATGC--------
-------ATGC--------
```

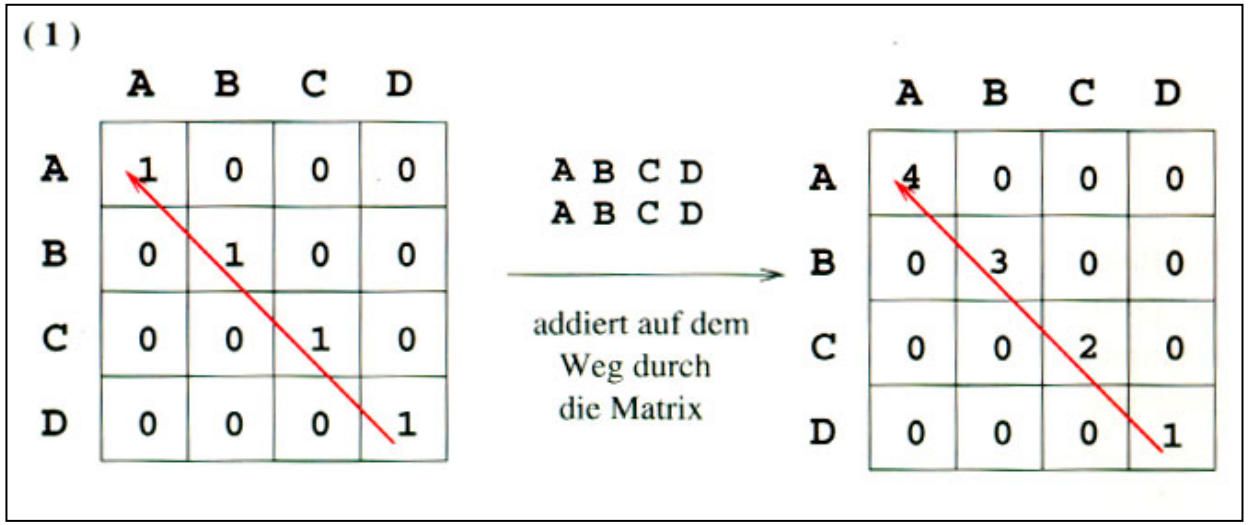
Needleman-Wunsch (N-W) 1970

- GLOBALES ALIGNMENT!
- Bei Erstellung des Alignments werden zunächst kleine Problem-Schritte gelöst. Dann wird aus den Teillösungen das Gesamtalignment rekonstruiert
- Algorithmus: „dynamic programming“

- Speicherbedarf: ≈ m x n

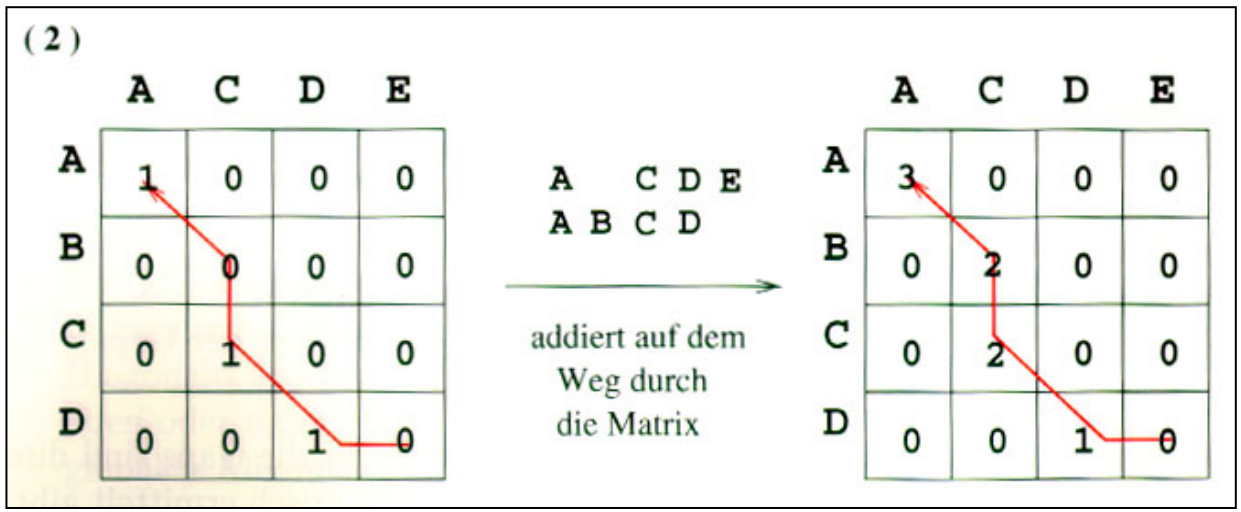
Bsp: Seq 1 (1kb) x Seq 2 (10kb) > 10 Mb Speicher

Vorgehensweise von NW



Einfachster Fall:

Zwei identische Sequenzen gegeneinander...



Etwas schwieriger...

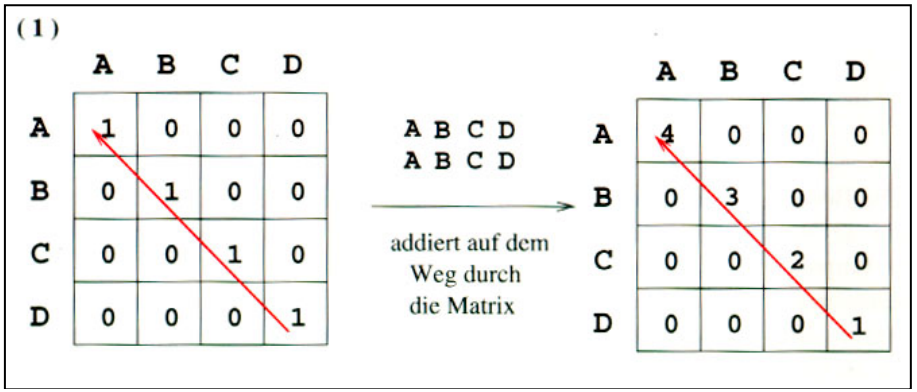
Vorgehensweise von NW

- wie beim Dotplot wird zunächst eine **zweidimensionale Matrix** mit den beiden zu vergleichenden Sequenzen erstellt

- in alle **Zellen der Matrix** wird der **Alignment-Score** für die jeweils verglichenen Sequenzpositionen hineingeschrieben. Die Berechnung des Score-Werts erfolgt natürlich anhand einer Substitutionsmatrize.

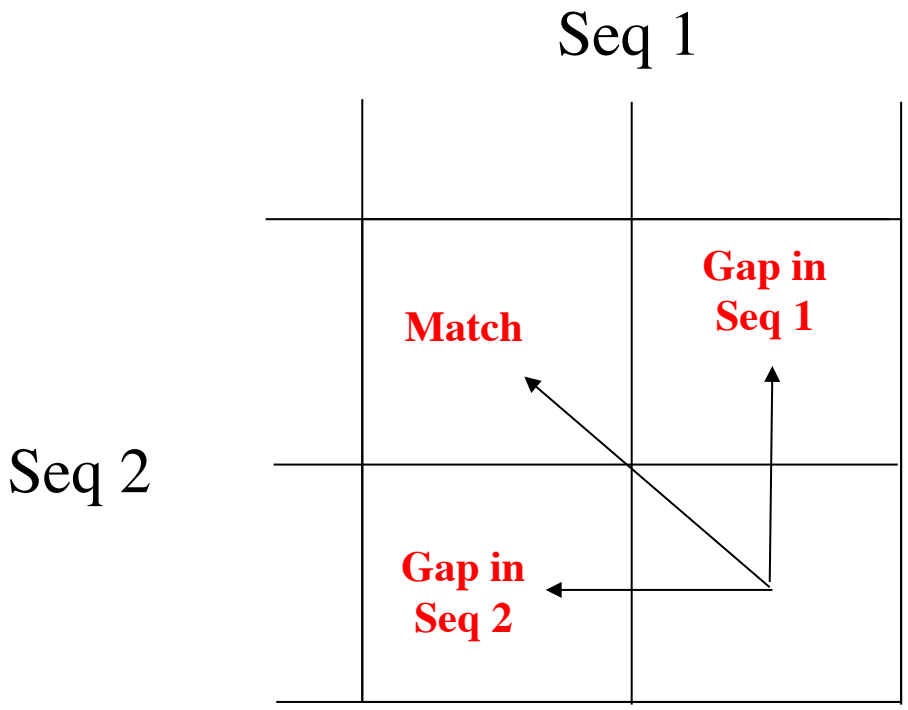
- das Alignment ergibt sich als Pfad durch die Matrix („**Traceback**“ vom End-Feld hin zum Start-Feld).

- Der Pfad mit der höchsten Endsumme repräsentiert das Alignment...



Needleman-Wunsch

- auch die Möglichkeit eines *gaps* muss bei Weg durch die Matrix berücksichtigt werden: **> Verlassen der Diagonalen**



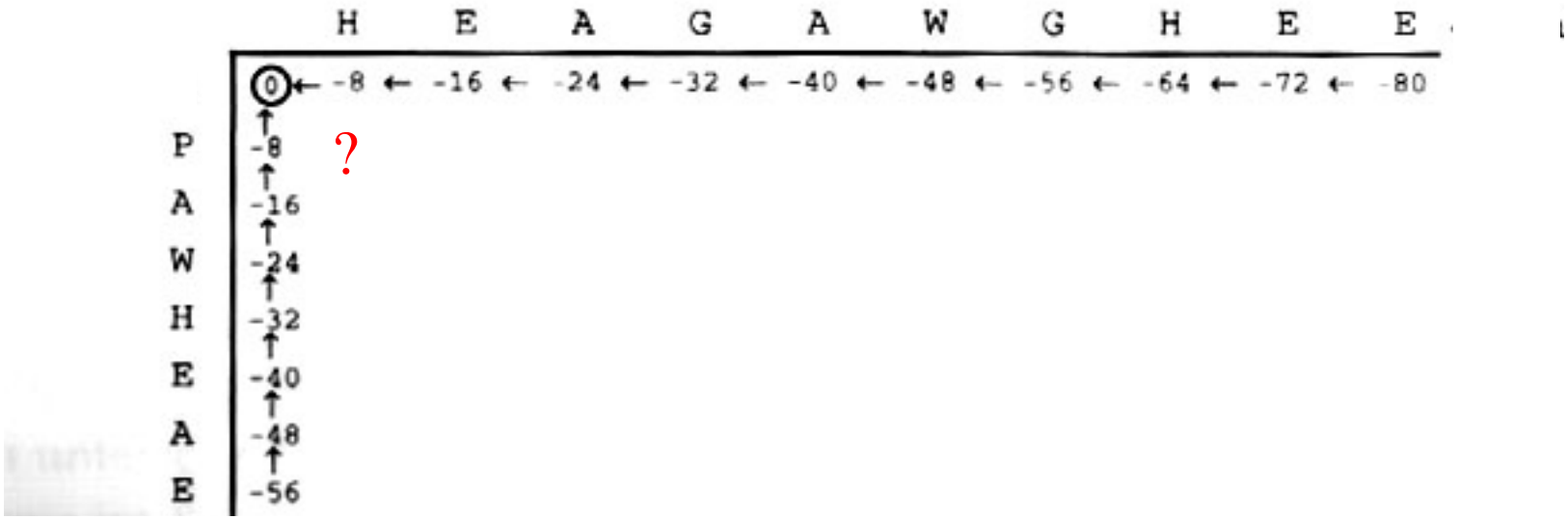
Needleman-Wunsch

Substitutions-Matrix:

	H	E	A	G	W
P	-2	-1	-1	-2	-4
A	-2	-1	5	0	-3
W	-3	-3	-3	-3	15
H	10	0	-2	-2	-3
E	0	6	-1	-3	-3

Gap penalty -8

Sequenz X	HEAGAWGHEE
Sequenz Y	PAWHEAE

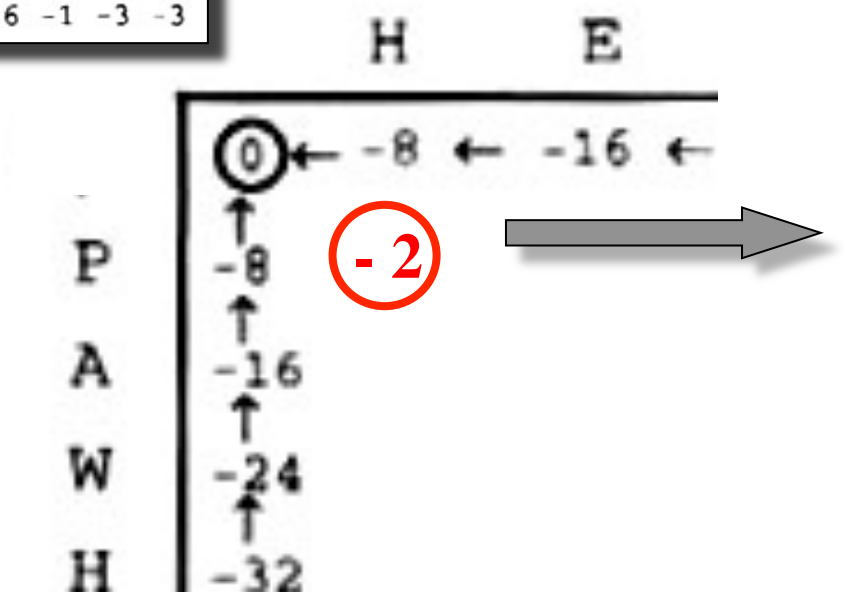


Aus:
Rauhut, Bioinformatik

Needleman-Wunsch

d = - 8

	H	E	A	G	W
P	-2	-1	-1	-2	-4
A	-2	-1	5	0	-3
W	-3	-3	-3	-3	15
H	10	0	-2	-2	-3
E	0	6	-1	-3	-3



3 Möglichkeiten!
Welche hat höchsten Score?

H
P

S = 0 + (-2) = - 2 !

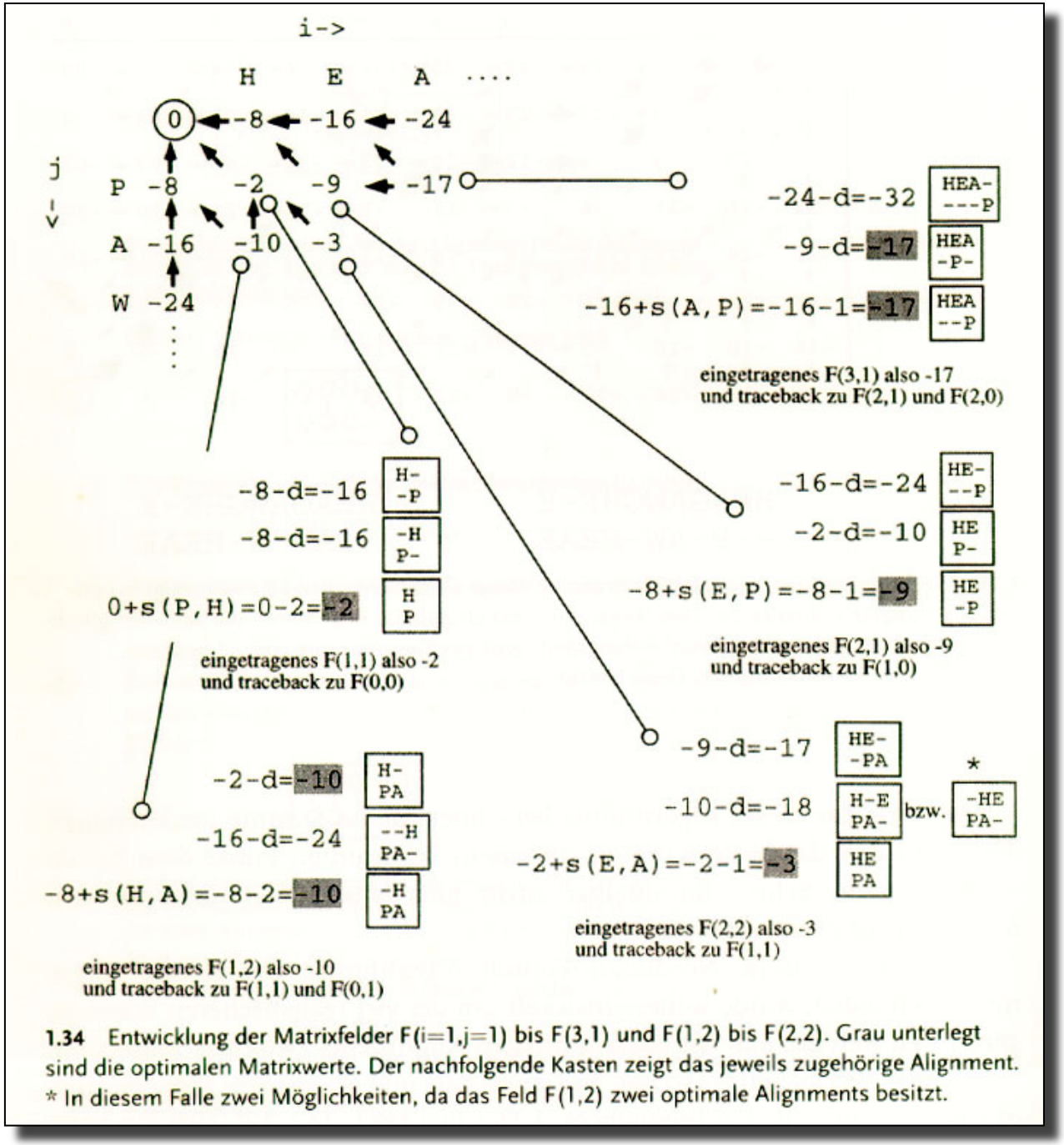
H -
- P

S = - 8 + d = - 16

- H
P -

S = - 8 + d = - 16

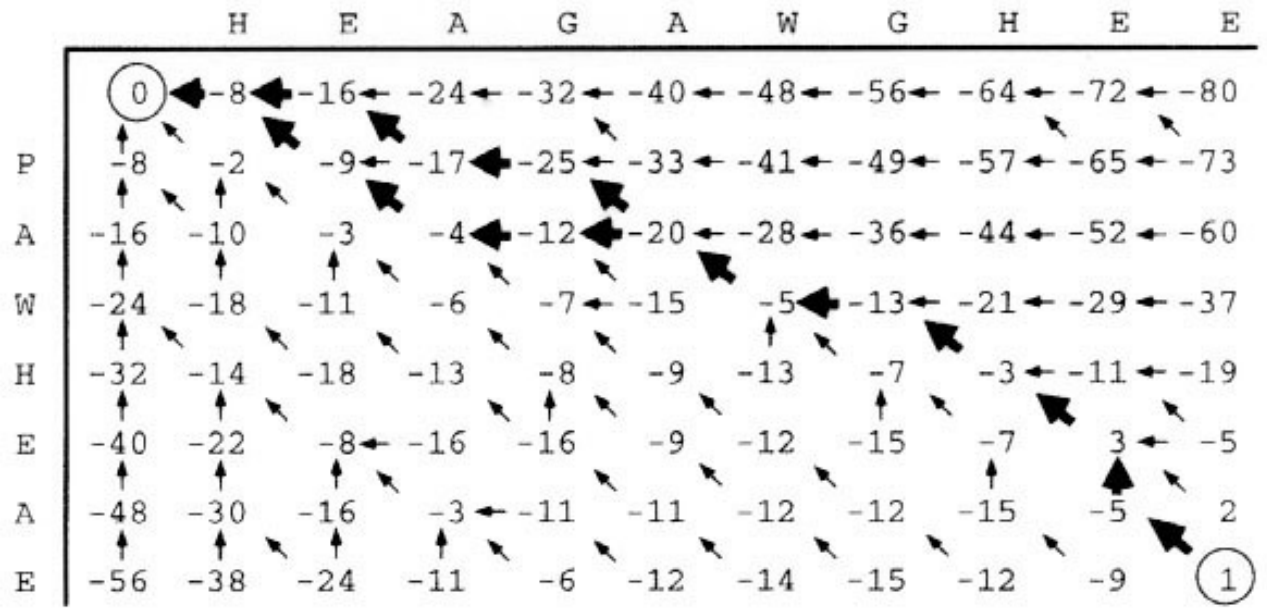
Und so geht‘s weiter....



1.34 Entwicklung der Matrixfelder F(i=1,j=1) bis F(3,1) und F(1,2) bis F(2,2). Grau unterlegt sind die optimalen Matrixwerte. Der nachfolgende Kasten zeigt das jeweils zugehörige Alignment.
* In diesem Falle zwei Möglichkeiten, da das Feld F(1,2) zwei optimale Alignments besitzt.

Aus:
Rauhut, Bioinformatik

Needleman-Wunsch



		H	E	A	G	A	W	G	H	E	E
	0	-8	-16	-24	-32	-40	-48	-56	-64	-72	-80
P	-8	-2	-9	-17	-25	-33	-41	-49	-57	-65	-73
A	-16	-10	-3	-4	-12	-20	-28	-36	-44	-52	-60
W	-24	-18	-11	-6	-7	-15	-5	-13	-21	-29	-37
H	-32	-14	-18	-13	-8	-9	-13	-7	-3	-11	-19
E	-40	-22	-8	-16	-16	-9	-12	-15	-7	3	-5
A	-48	-30	-16	-3	-11	-11	-12	-12	-15	-5	2
E	-56	-38	-24	-11	-6	-12	-14	-15	-12	-9	1

HEAGAWGHE-E
--P-AW-HEAE

HEAGAWGHE-E
-PA--W-HEAE

1.33 Die fettgedruckten traceback-Pfeile der vervollständigten Matrix zeigen zwei Wege, um vom Endfeld mit dem score +1 zum Anfangsfeld mit dem score 0 zurückzufinden. Diese beiden Wege entsprechen den beiden gezeigten optimalen *globalen* Alignments, die beide ein gleichwertiges Gesamt-score von +1 besitzen.

Trace-Back der Pfeile zeigt den Weg des optimalen Alignments

Aus:
Rauhut, Bioinformatik

Aber…

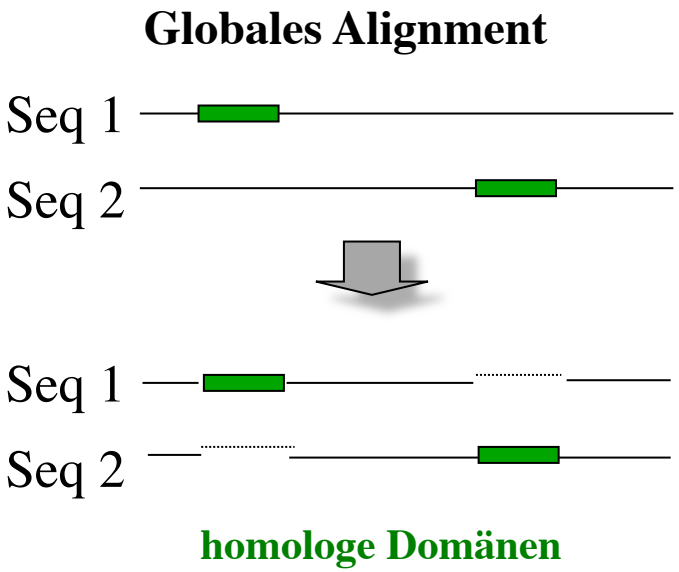
• das globale N-W Alignment funktioniert nur gut bei relativ nahe verwandten Sequenzen!!!

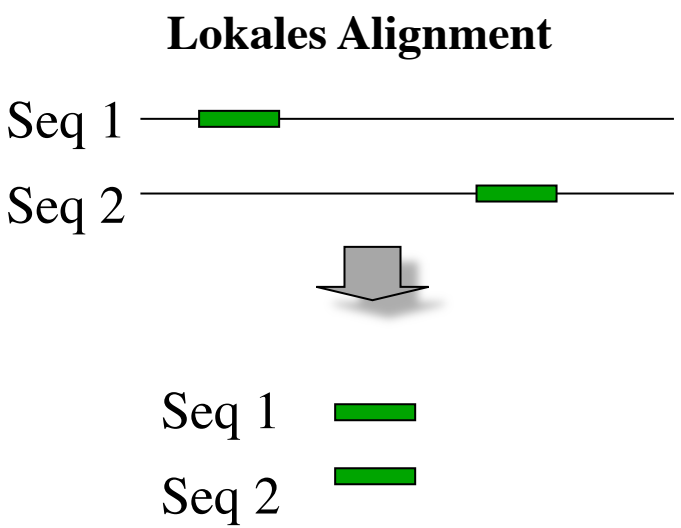
• das globale Alignment findet nicht...

...homologe Domänen, wenn diese an unterschiedlichen Stellen in den Sequenzen liegen

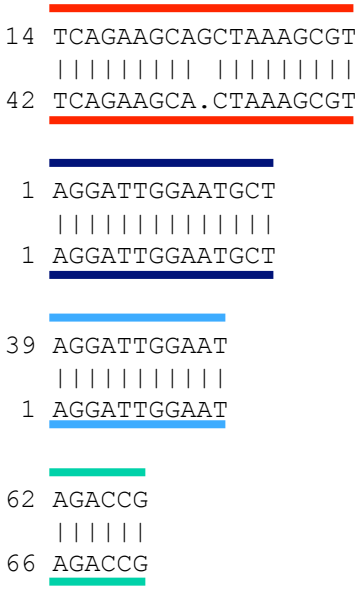
...kleinere konservierte Bereiche in ansonsten divergenten Sequenzen

Global vs. Lokal





Global vs. Lokal



Smith-Waterman (S -W) 1981

- finde den längsten gemeinsamen Bereich zweier Sequenzen mit der größten Ähnlichkeit

- liefert nur EIN lokales Alignment (das mit höchstem Score)!!!

- Berechnung ähnlich wie bei N-W, aber...

  ....wenn alle 3 Alignmentmöglichkeiten an einer Position negative Scores haben, wird Score-Wert auf **0** gesetzt

  > Pfade starten und enden innerhalb der Matrix

Smith-Waterman

		H	E	A	G	A	W	G	H	E	E
	0	0	0	0	0	0	0	0	0	0	0
P	0	0	0	0	0	0	0	0	0	0	0
A	0	0	0	5	0	5	0	0	0	0	0
W	0	0	0	0	2	0	20	12	4	0	0
H	0	10	2	0	0	0	12	18	22	14	6
E	0	2	16	8	0	0	4	10	18	28	20
A	0	0	8	21	13	5	0	4	10	20	27
E	0	0	6	13	18	12	4	0	4	16	26

Somit ergibt sich das folgende optimale **lokale** Alignment

AWGHE
AW–HE

mit einem score von 28

1.35 Die Berechnung der Matrix erfolgt genauso wie in Abb. 1.33 und 1.34. Sind aber die drei Möglichkeiten kleiner als Null, so wird der Matrixwert 0 eingetragen. Das *lokale* Alignment beginnt nun an der Matrixposition mit dem höchsten score (hier +28) und endet wenn ein Nullfeld erreicht wird. Die traceback-Pfeile dieses Alignments sind fettgedruckt.

Aus:
Rauhut, Bioinformatik

Zusammenfassung

- N-W maximiert matches und minimiert gap-Anzahl
  > optimales Alignment mit höchstem Score

- N-W nur tauglich für relative ähnliche Sequenzen ohne Änderungen in ihrer Domänen-Architektur

- N-W- Programme:

GAP	in GCG-Programmpaket
Needle	http://mobyle.pasteur.fr/cgi-bin/portal.py?#forms::needle

Zusammenfassung

- S-W maximiert matches durch Einführung von gaps
  > ein optimales lokales Alignment

- S-W sehr sensitiv für Domänensuche

- S-W liefert immer Ergebnis! Anwender muss beurteilen!

- S-W- Programme:

BESTFIT	in GCG-Programmpaket
WATER	http://mobyle.pasteur.fr/cgi-bin/portal.py

GCG-Programme

A. GAP (Needleman-Wunsch algorithm)

Percent Similarity: 44.651 Percent Identity: 36.279

```
  1 MSTKKKPLTQEQLEDARRLKAIYEKKKNELGLSQESVADKMGMGQSGVGA 50
    |·|          ||    |      |  ::|· | · | ···  ·|··  ·:·
  1 MNT........QLMGER....IRARRKK.LKIRQAALGKMVGVSNVAISQ 37

 51 LFNGINALNAYNAALLAKILKVSVEEFSPSIAREIYEMYEAVSMQPSLRS 100
       |· |    |·| |  |  :   ·   :   |  ·   ·    |·
 38 WERSETEPNGENLLALSKALQCSPDYLLKGDLSQTNVAYHS...RHEPRG 84

101 EYEYPVFSHVQAGMFSPELRTFTKGDAERWVSTTKKASDSAFWLEVEGNS 150
      ||·  |  |  ||        ·  ·  |   |  |   ||    |·  ·|||·|·|·|
 85 ..SYPLISWVSAGQWMEAVEPYHKRAIENWHDTTVDCSEDSFWLDVQGDS 132

151 MTAPTGSKPSFPDGMLILVDPEQAVEPGDFCIARLGGD.EFTFKKLIRDS 199
    ||||·|    |  |:||·|||||||      |    :|:|  |:  |  |||||:  |·
133 MTAPAG..LSIPEGMIILVDPEVEPRNGKLVVAKLEGENEATFKKLVMDA 180

200 GQVFLQPLNPQYPMIPCNESCSVVGKVIASQWPEETFG 237
    |·  ||  ||||||||||   |  ·|  ::|  |:  ·  ··
181 GRKFLKPLNPQYPMIEINGNCKIIGVVVDAKLAN..LP 216
```

B. BESTFIT (Smith-Waterman algorithm)

Percent Similarity: 58.871 Percent Identity: 48.387

```
104 YPVFSHVQAGMFSPELRTFTKGDAERWVSTTKKASDSAFWLEVEGNSMTA 153
    ||·  |  |  ||      ·   :  |    |  |   ||    |:  ·|||:|:|:|||||
 86 YPLISWVSAGQWMEAVEPYHKRAIENWHDTTVDCSEDSFWLDVQGDSMTA 135

154 PTGSKPSFPDGMLILVDPEQAVEPGDFCIARLGGD.EFTFKKLIRDSGQV 202
    |·|    |  |:||·||||||      |    :|:|  |:  |  ||||||:  |·|·
136 PAG..LSIPEGMIILVDPEVEPRNGKLVVAKLEGENEATFKKLVMDAGRK 183

203 FLQPLNPQYPMIPCNESCSVVGKVIAS 229
    ||  ||||||||||   |  ·|  ::|  |:  ·
184 FLKPLNPQYPMIEINGNCKIIGVVVDA 210
```

GAP (global)

BESTFIT (lokal)

- Proteine enthalten oft wiederholte Domänen!

  Die Programme LALIGN (fasta.bioch.virginia.edu/fasta_www/**lalign**.htm) und SIM (http://www.expasy.ch/tools/sim.html) können mehrere lokale Alignments in absteigender Qualität anzeigen!

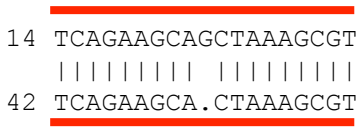
- Die Programme FASTA und BLAST2 (eigentlich für Datenbank-Suchen gemacht; kein dyn. Program.-Algorithmus) können auch lokale Alignments machen.

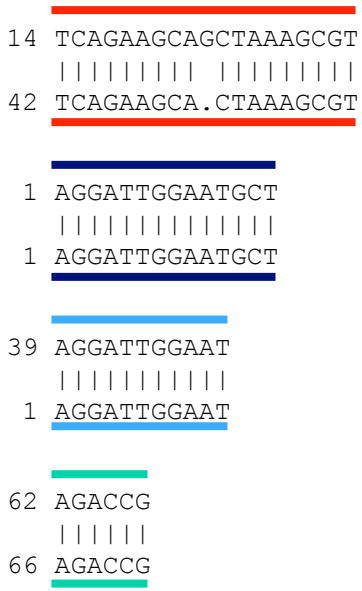
  http://www.ncbi.nlm.nih.gov/blast/bl2seq/bl2.html

  http://fasta.bioch.virginia.edu/

GLOBAL

LOKAL





Algorithm: Bestfit (Smith & Waterman)

identifiziert Region mit bester lokaler Ähnlichkeit

Algorithmus: SIM (Huang & Miller)

identifiert **alle** Regionen mit lokaler Ähnlichkeit