WS2017/2018

F1-Praktikum
Genomforschung und Sequenzanalyse: Einführung in Methoden der Bioinformatik

Thomas Hankeln

Alignments & Datenbanksuchen

Wiederholung „Alignments“

- Dynamic Programming
  - Needleman-Wunsch: globales Alignment
  - Smith-Waterman: lokales Alignment

} optimal

- Scoring-Matrizen: PAM & BLOSUM
- Gap penalties

- Dotplots: Visualisierung von Alignments

- FASTA, BLAST, BLAT

} heuristisch

Dynamic Programming

- Wie finde ich das beste alignment?

  Alle ausprobieren?
  Zu langsam: 2 x 300 Bp = 10^{88} Möglichkeiten

- Dynamic programming: **löse kleine Sub-Probleme und konstruiere daraus Gesamtlösung**

Dynamic Programming

Zwei Sequenzen s und t der Länge L_s und L_t

$s = s1, s2, s3...sL_s$

$t = t1, t2, t3...tL_t$

Wir möchte das optimale Alignment über die volle Länge

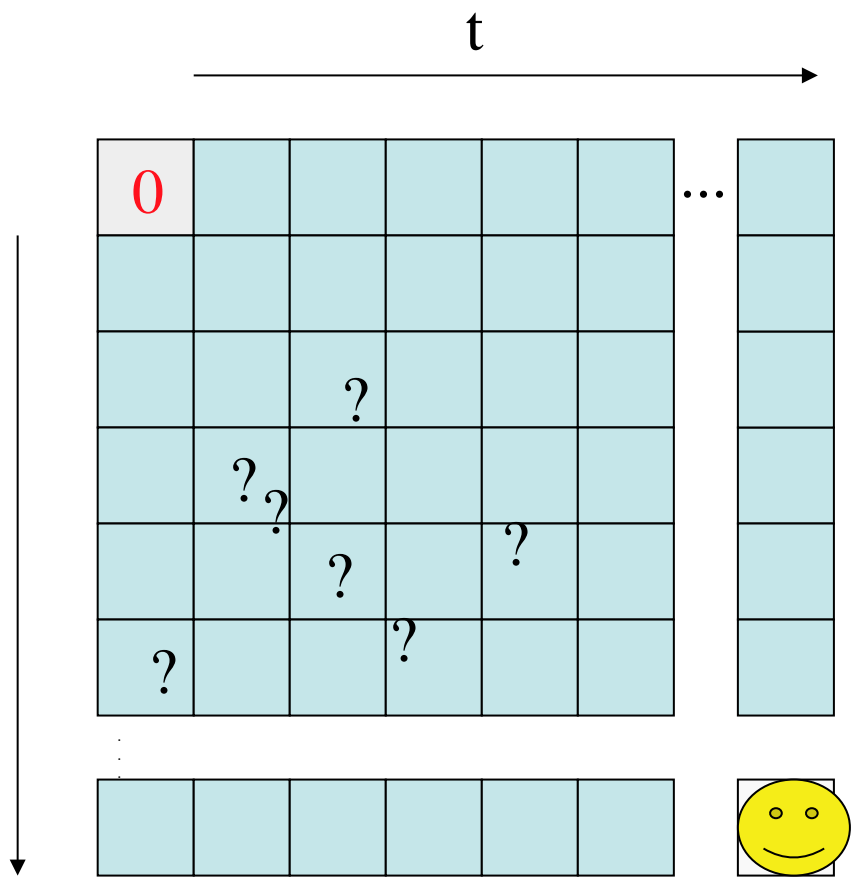
- Konstruiere Sub-Alignments
- Kombiniere diese Teillösungen rekursiv

Benutze dabei eine dynamic programming-Matrix M

Die DP Matrix

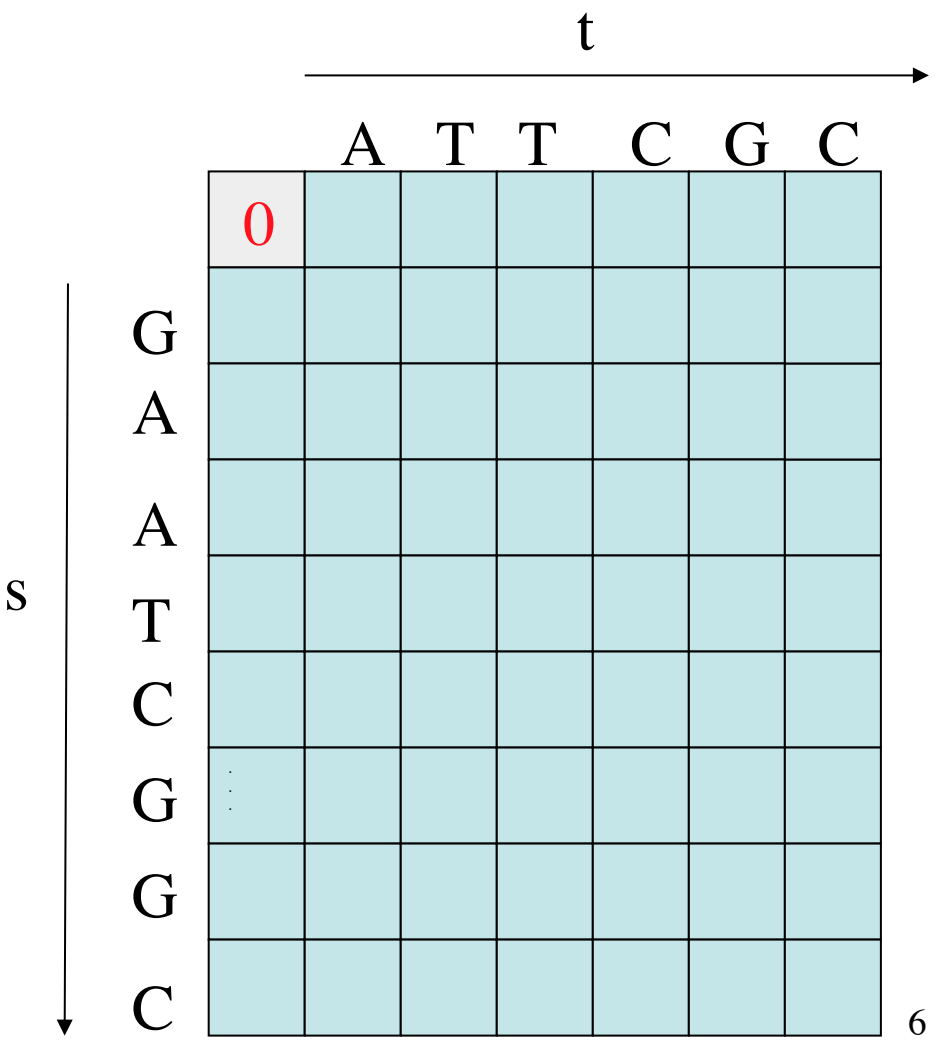
- Eintrag M (i,j): Optimaler Score für Alignment von s mit t

- Startfeld: Alignment von Nix mit Nix (= 0)



Füllen der Matrix: Schritt1

- einen Eintrag haben wir nun schon...
- von da aus gibt es drei Wege, sich durch die Matrix zu bewegen und weitere Felder auszufüllen



Füllen der Matrix

- wir habe drei Möglichkeiten in jedem Matrixfeld....

1. Aligne Nt in s mit Nt in t
2. Aligne Nt in s mit Lücke in t
3. Aligne Lücke in s mit Nt in t

(4. Aligne Lücke mit Lücke = Quatsch)

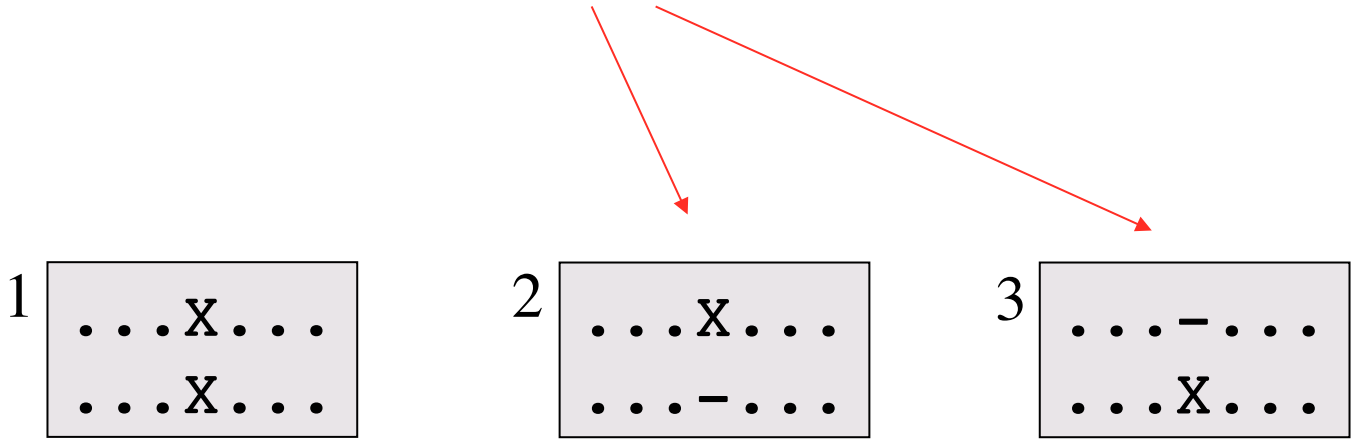
1
```
...x...
...x...
```

2
```
...x...
...-...
```

3
```
...-...
...x...
```

Füllen der Matrix: gap penalty

> gap penalty einführen (z. B. -1)

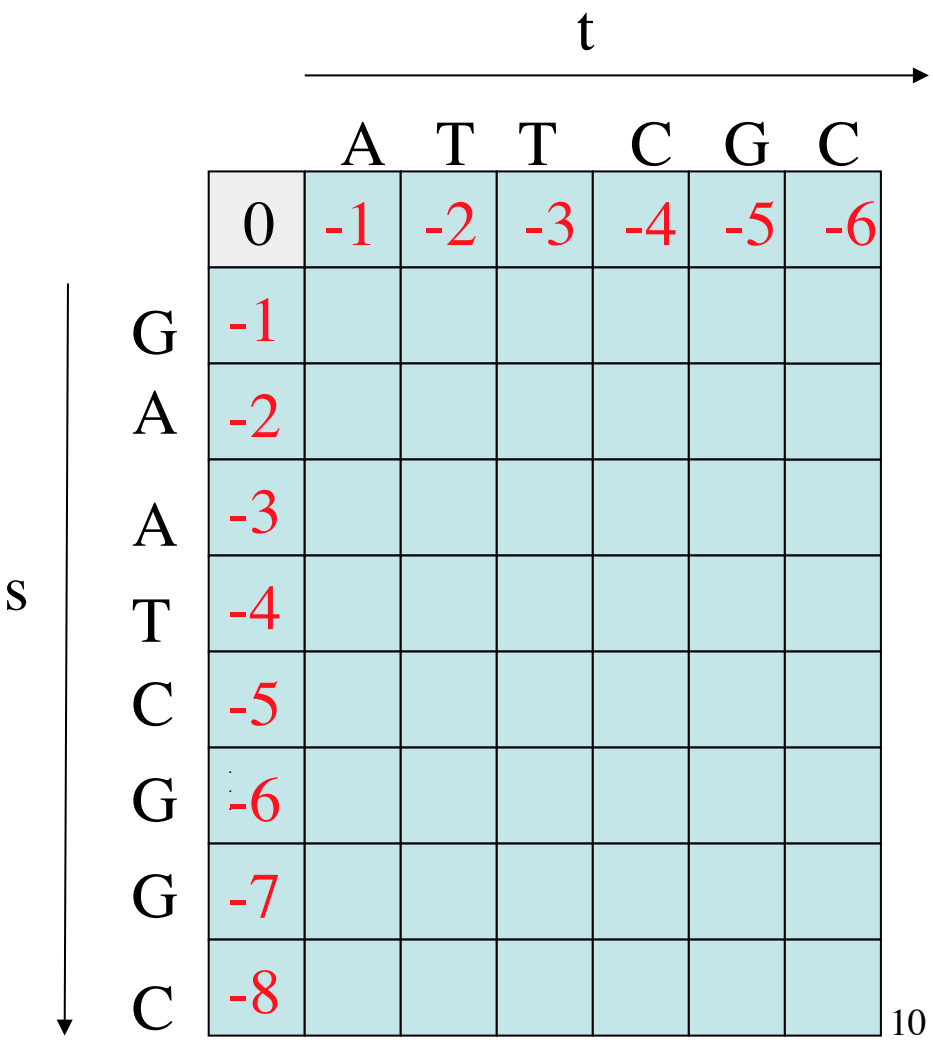


Füllen der Matrix: matches

- Ziel: Maximaler Score
- Identitäts-Scores: Matches = +1, Mismatches = 0

Füllen der Matrix: Schritt 2

- gap penalty = -1
- obere Reihe:
  aligne t mit gaps in s
- linke Spalte:
  aligne s mit gaps in t



Füllen der Matrix

- für jedes Matrixfeld die drei Möglichkeiten berechnen:

    1. Aligne beide Nts und addiere score (+1 oder 0)
    2. Inseriere gap in s und addiere -1
    3. Inseriere gap in t und addiere -1

- Das Ganze rekursiv auf das jeweilige Nachbarfeld beziehen:

$$M(i,j) = \max \begin{cases} M(i-1, j-1) + \text{score}(a,b) \\ M(i-1, j) + (-1) \\ M(i, j-1) + (-1) \end{cases}$$

Füllen der Matrix

Nur nochmal eine andere Formalisierung des Gesagten....

		x_1	x_2	$\ldots$	x_{i-1}		x_i	$\ldots$	x_m
	$F(0,0)$	$F(1,0)$	$F(0,2)$				$\vdots$		
y_1	$F(0,1)$						$\vdots$		
y_2	$F(0,2)$						$\vdots$		
$\vdots$							$\vdots$		
y_{j-1}					$F(i-1,j-1)$		$F(i,j-1)$		
						$\nwarrow$	$\uparrow$		
y_j	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$F(i-1,j)$	$\leftarrow$	$F(i,j)$		
$\vdots$									
y_n									

Füllen der Matrix

$$M(i,j) = \max \begin{cases} M(i-1, j-1) + score(a,b) \\ M(i-1, j) + (-1) \\ M(i, j-1) + (-1) \end{cases}$$

↖ 0 + 0 = 0
← -1 + (-1) = -2
↑ -1 + (-1) = -2

max

↖ -1 + 1 = 0
← -2 + (-1) = -3
↑ 0 + (-1) = -1

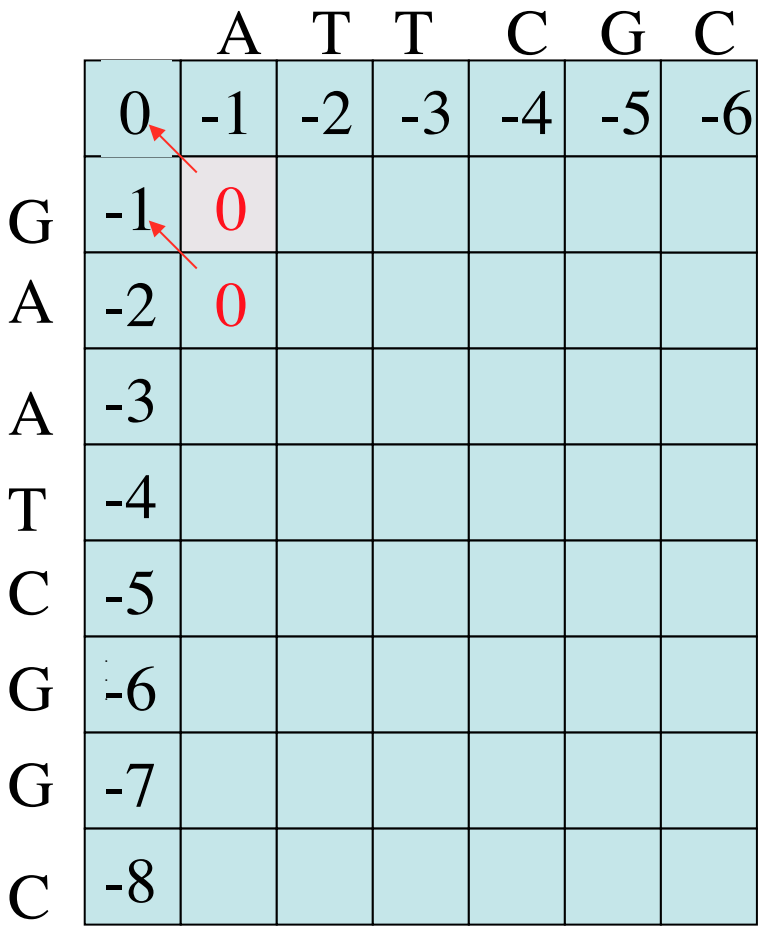
max

		A	T	T	C	G	C
	0	-1	-2	-3	-4	-5	-6
G	-1	0					
A	-2	0					
A	-3						
T	-4						
C	-5						
G	-6						
G	-7						
C	-8						

Aufgabe: Füllen der Matrix

- Bestimme den optimalen Pfad des Alignments und Schreibe das Alignment

Achtung: es kann mehrere gleichwertig optimale Alignments geben!



NW- Globales Alignment

- Alle möglichen optimalen Alignments werden durch „trace-back“ als Pfad in der Matrix gefunden

- Sequenzen sind von Anfang bis Ende alignt.

Die Lösung…

$$A_1 = \begin{bmatrix} \texttt{GAATCGGC} \\ \texttt{-ATTCG-C} \end{bmatrix}$$

$$A_2 = \begin{bmatrix} \texttt{GAATCGGC} \\ \texttt{-ATTC-GC} \end{bmatrix}$$

$s\backslash t$		A	T	T	C	G	C
	0	-1	-2	-3	-4	-5	-6
G	-1	0 (↖)	-1 (↖ ←)	-2 (↖ ←)	-3 (↖ ←)	-3 (↖)	-4 (←)
A	-2	0 (↖)	0 (↖)	-1 (↖ ←)	-2 (↖ ←)	-3 (↖ ←)	-3 (↖)
A	-3	-1 (↖ ←)	0 (↖)	0 (↖)	-1 (↖ ←)	-2 (↖ ←)	-3 (↖ ←)
T	-4	-2 (↑)	0 (↖)	1 (↖)	0 (↖ ←)	-1 (↖ ←)	-2 (↖ ←)
C	-5	-3 (↑)	-1 (↑)	0 (↑ ↖)	2 (↖)	1 (←)	0 (↖ ←)
G	-6	-4 (↑)	-2 (↑)	-1 (↑ ↖)	1 (↑)	3 (↖)	2 (←)
G	-7	-5 (↑)	-3 (↑)	-2 (↑ ↖)	0 (↑)	2 (↑ ↖)	3 (↖)
C	-8	-6 (↑)	-4 (↑)	-3 (↑ ↖)	-1 (↑ ↖)	1 (↑)	3 (↖)

↖ match

← Gap in s

↑ Gap in t

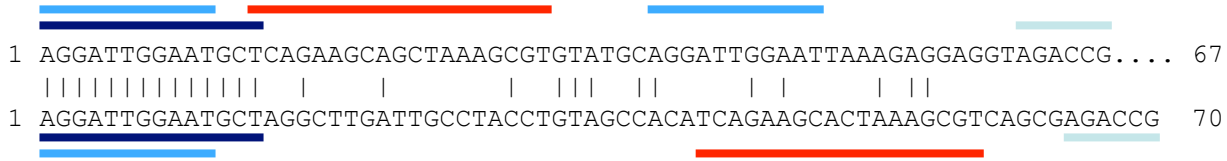
NW- Globales Alignment

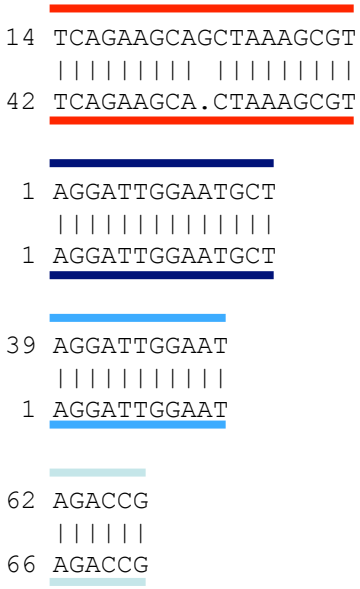
Input: two sequences x and y
Output: optimal alignment and score α
Initialization:
Set $F(0,0) := 0$
Set $F(i,0) := -id$ and $T(i,0) := (i-1,0)$ for all $i = 1,2,\ldots,m$
Set $F(0,j) := -jd$ and $T(0,j) := (0,j-1)$ for all $j = 1,2,\ldots,n$
Recurrence:
for $i = 1,2,\ldots,m$ **do**:
  for $j = 1,2,\ldots,n$ **do**:

$$\text{Set } F(i,j) := \max \begin{cases} F(i-1,j-1) + s(x_i, y_i) \\ F(i-1,j) - d \\ F(i,j-1) - d \end{cases}$$

    Set backtrace $T(i,j)$ to the maximizing pair (i',j') (encoded as $\in \{\leftarrow, \nwarrow, \uparrow\}$)
The best score is $\alpha := F(m,n)$
Set $(i,j) := (m,n)$
Traceback:
repeat
  if $T(i,j) = (i-1,j-1)$ **print** $\binom{x_{i-1}}{y_{j-1}}$
  else if $T(i,j) = (i-1,j)$ **print** $\binom{x_{i-1}}{-}$ **else print** $\binom{-}{y_{j-1}}$
  Set $(i,j) := T(i,j)$
until $(i,j) = (0,0)$.

Global vs. Lokal





Lokales Alignment

Wenn die Sequenzen divergenter sind als zuvor...

- nur kurze konservierte Bereiche
- Abschnitte un-alignierbarer Sequenzen
- globales A. unsinnig

Verwende *Smith-Waterman* lokales Alignment:

$$M(i,j) = \max \begin{cases} M(i-1, j-1) + \text{score}(a,b) \\ M(i-1, j) + (-1) \\ M(i, j-1) + (-1) \\ 0 \quad \text{<hier ist der Unterschied!} \end{cases}$$

...wie Reset-Knopf: starte erneut wenn Score unter Null fällt

Lokales Alignment

- lokales Alignment kann überall, auch innerhalb der Matrix starten

- obere Zeile und linke Spalte, sowie alle negativen Positionen werden mit 0 gefüllt

- trace-back vom Feld mit höchstem Score zuerst starten

- bei mehreren Startstellen: mehrere lokal konservierte Abschnitte

Lokales Alignment

optimale lokale Alignments:

```
TCG
TCG

A-TCGGC
ATTC-GC

TCGGC
TC-GC

AT-CGGC
ATTCG-C

TCGGC
TCG-C
```

		A	T	T	C	G	C
	0	0	0	0	0	0	0
G	0	0	0	0	0	1	0
A	0	1	0	0	0	0	0
A	0	1	0	0	0	0	0
T	0	0	2	1	0	0	0
C	0	0	1	1	2	1	1
G	0	0	0	0	1	3	2
G	0	0	0	0	0	2	2
C	0	0	0	0	1	1	3

Zur Erinnerung...

- **„Optimales alignment“ heißt:**
  höchster score, gegeben die Matrix und die gap penalty
- **dies ist nicht unbedingt das biologisch sinnvollste Alignment**
- **Paarweise Alignment-tools produzieren immer etwas, manchmal auch Sinnloses...**
- **DNA: Paarweise alignment-tools können nur Nt-matches finden, wenn der gleiche Strang zweier Sequenzen verglichen wird**

Gap Penalty

- linear:

$$W = -(G_{open} \times \text{Länge } n)$$

- affin:

$$W = -(G_{open} + (G_{ext} \times (n-1))$$

```
     GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKL
lin  GSAQVKGHGKK--------VA--D----A-SALSDLHAHKL

     GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKL
aff  GSAQVKGHGKKVADA---------------SALSDLHAHKL
```

Gap Penalty

- affin: $W = G_{open} + (G_{ext} \times (n - 1))$

  mit $G_{ext} < G_{open}$ werden weniger aber, grössere Lücken favorisiert

- Werte zu gross: keine Lücken;
  Werte zu klein: zu viele Lücken

- gap penalties sind auf scoring-Matrix abgestimmt

- für overlap alignment: end-gaps nicht bestraft

Scoring-Matrizen

- 4 x 4 für DNA
- 20 x 20 für Proteine

- Belohnungsscores für konservierte Positionen

- PAM (percent accepted mutation)
- BLOSUM (blocks substitution matrix)

PAM-Familie

- 1 PAM = 1 % Austausch (ca. 10 Mio Jahre)

  Was bedeutet dann PAM 250?

- umfasst multiple Austausche A > B > A

- PAM (klein) für ähnliche Sequenzen
  PAM (groß) für divergente Sequenzen

PAM 1 etc.

1. Finde sehr ähnliche Sequenzen (1% Divergenz)
2. Mache globales alignment (per Hand, 71 Gruppen)
3. Zähle die Austausche (1572)
4. Berechne die Scorewerte für die einzelnen Austausche

5. PAM2 = PAM1 x PAM1, PAM3 = PAM1 x PAM2 etc
   Extrapolation für größere Divergenzen durch Multiplikation

PAM

PAM	Number of observed substitutions per 100 amino acids
1	1
11	10
23	20
38	30
56	40
80	50
120	60
159	70
250	80

BLOSUM

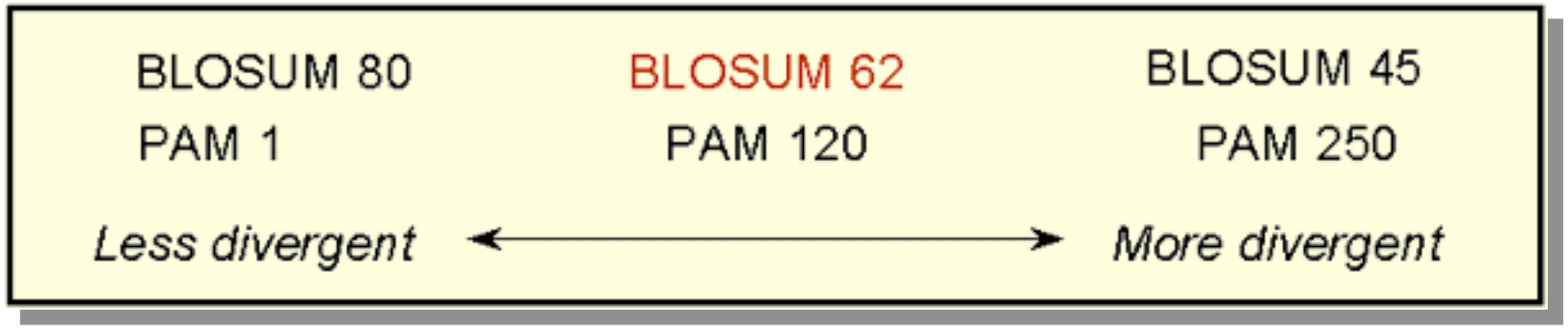
1. Suche alignbare Blöcke ohne gaps
   (2000 Blöcke, 500 Proteinfamilien)

2. Zähle Austausche direkt, kalkuliere log odds-Werte
   Log (a,b) / (a x b)

3. Mache dies ohne Extrapolation für Blöcke unterschiedlichen Grades an Ähnlichkeit

➢ BLOSUM-Matrizen für unterschiedliche phylogenetische Abstände

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.

Matrizen und gap penalties bestimmen Signifikanz-Scores

A. Search with MJ0050

	BLOSUM50 -10/-2				BLOSUM62 -7/-1				BLOSUM62 -11/-1			
The best scores are:	s-w	E()	%_id	alen	s-w	E()	%_id	alen	s-w	E()	%_id	alen
NP_416010 glutamate decarb.	250	e-11	24.9	401	216	e-7	25.3	415	137	e-8	22.9	332
NP_417379 glycine decarb.	169	e-05	22.1	420	163	0.001	23.3	430	88	0.004	22.1	331
NP_417025 aminotransferase	122	0.02	23.6	254	119	0.12	24.5	257	76	0.04	23.7	118
NP_414772 aminoacyl-his.	110	0.15	23.4	188	108	0.74	23.2	311	57	6.9	23.4	188
NP_415139 alkyl hydroperoxide	99	1.1	26.9	156	104	1.5	24.5	233	62	2.0	28.9	97

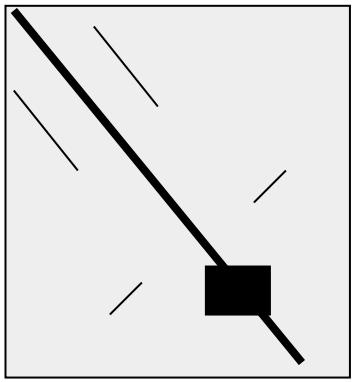
B. Search with MJ1633

	BLOSUM50 -10/-2				BLOSUM62 -7/-1				BLOSUM62 -11/-1			
The best scores are:	s-w	E()	%_id	alen	s-w	E()	%_id	alen	s-w	E()	%_id	alen
NP_417809 KefB	196	e-06	28.2	177	162	0.02	27.3	176	143	e-8	34.4	96
NP_414589 K+ antiporter	175	e-04	25.4	142	141	0.2	24.7	166	131	e-7	25.4	142
NP_415011 transport protein	133	0.03	23.2	142	113	4.4	23.2	142	89	0.005	23.2	142
NP_417748 TrkA	128	0.04	23.7	135	114	2.9	22.2	176	99	e-3	21.8	133
NP_416807 NAD(P) binding	103	0.98	26.1	92					70	0.29	26.1	92

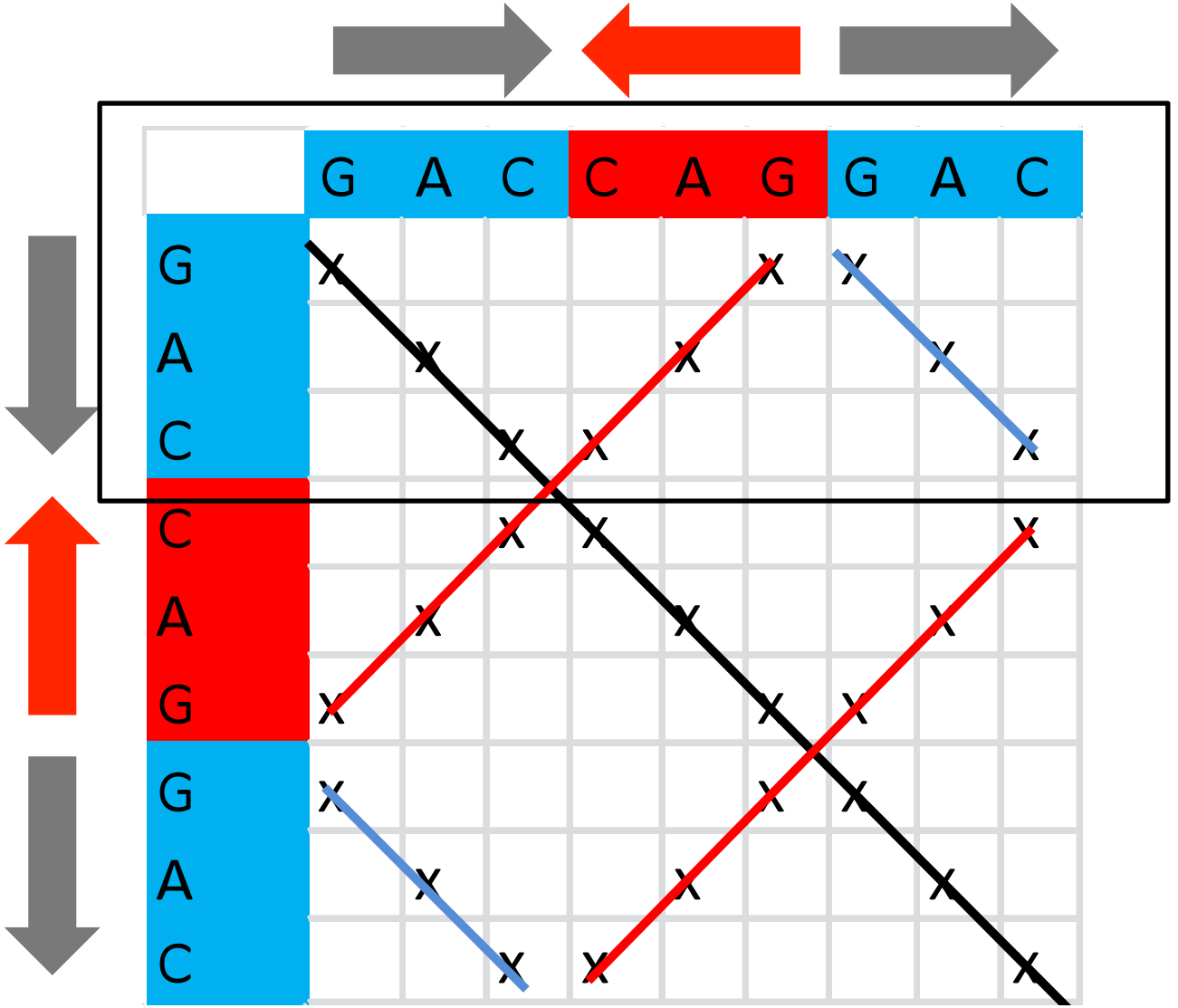
Wichtig im Grenzbereich...!!!

Dot Plot

- manchmal reicht es, Alignments visuell zu vereinfachen...
- Sequenz mit sich selbst vergleichen:
  - > Hauptdiagonale
  - > parallele Diagonalen: direct repeats
  - > orthogonale Diagonalen: inverted repeats
  - > Quadrate mit „noise“: simple repeats



Dot Plot

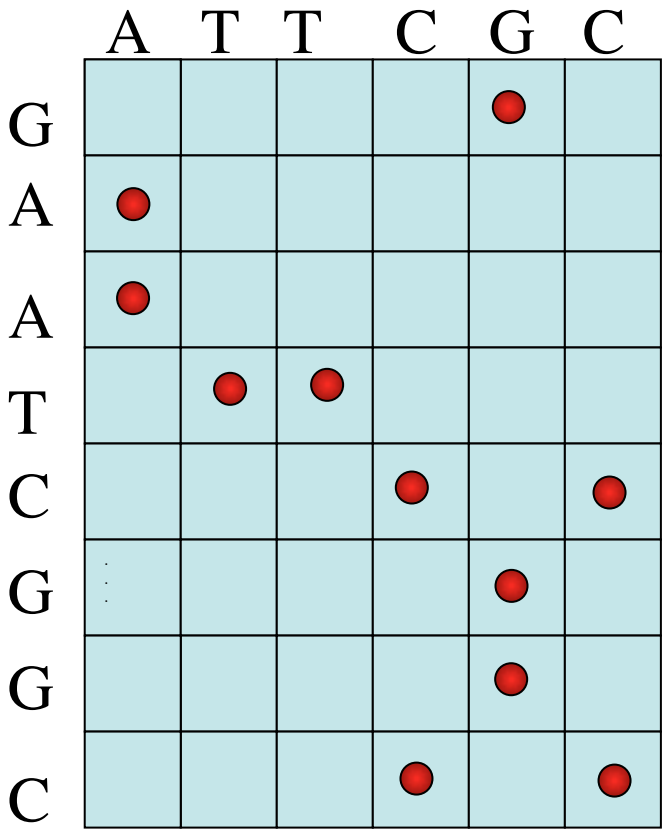


Dot Plot

Mache einen Dotplot..

- window =1
  Stringenz = 1

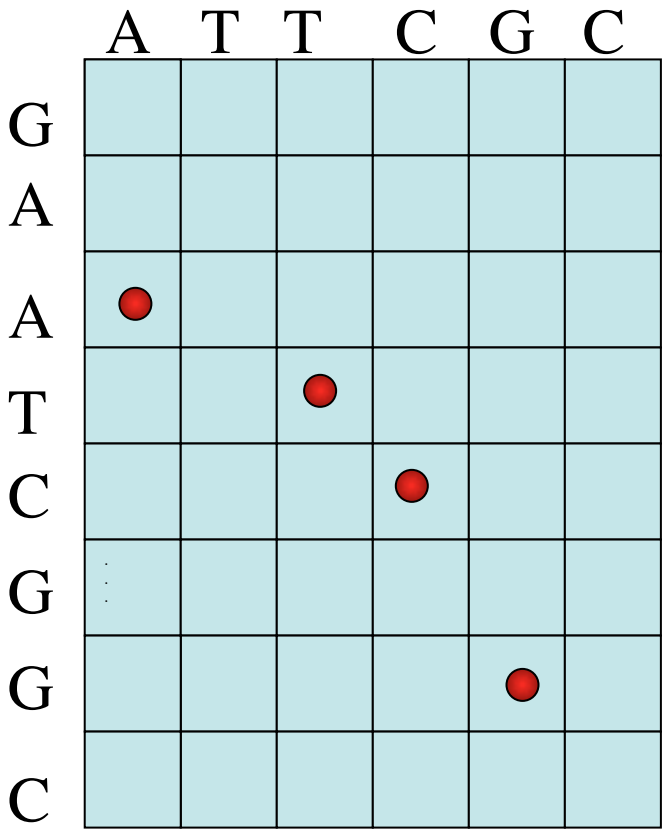
- window = 2
  stringenz = 2



Dot Plot

Mache einen Dotplot..

- window = 2
  stringenz = 2



Alignment für DB-Suchen

Heuristische Algorithmen:
gut & schnell, aber nicht unbedingt optimal

„seed-and extend“:

- Suchsequenz in kurze Abschnitte („words“ bzw. „*k*-tuple“) aufbrechen (Wilbur und Lipman, 1983).
- zunächst sehr schnell nach „word hits“ in der DB suchen
- diese dann erweitern zu längeren Segmenten

FastA

- **Fast-A**ll: Pearson & Lipman 1985

- ist DB-Suchprogramm-Kollektion, aber auch Dateiformat

- Ktup-Suche: aa = 2, Nt = 6
  Achtung: je länger das Ktup oder word, desto schneller, aber auch unsensitiver ist die Suche

- verbindet Ktuples, die nahe zueinander auf der gleichen Diagonale liegen

FastA

KFLVMDEADRLLDEEFGPVL

Tuple	Matching positions
AD	8
DE	6, 13
DR	9
EA	7
EE	14
EF	15
FG	16
FL	2
GP	17
KF	1
LD	12
LL	11
LV	3
MD	5
PV	18
RL	10
VM	4
VL	19

K=2

1. Bereche Index mit den Positionen aller Ktups in der Query-Sequenz

1-> PTGLVPC
2-> APLVGGVVV
3-> LVMDEADRLV

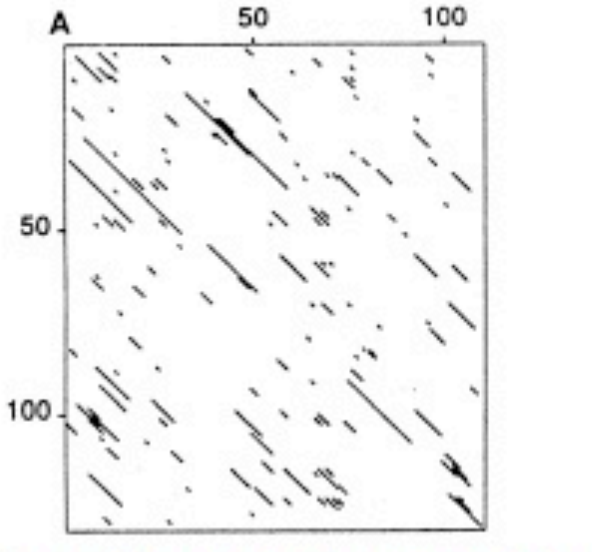
Tuple	Matching sequence	Matching position
LV	1	4
	2	3
	3	1,9

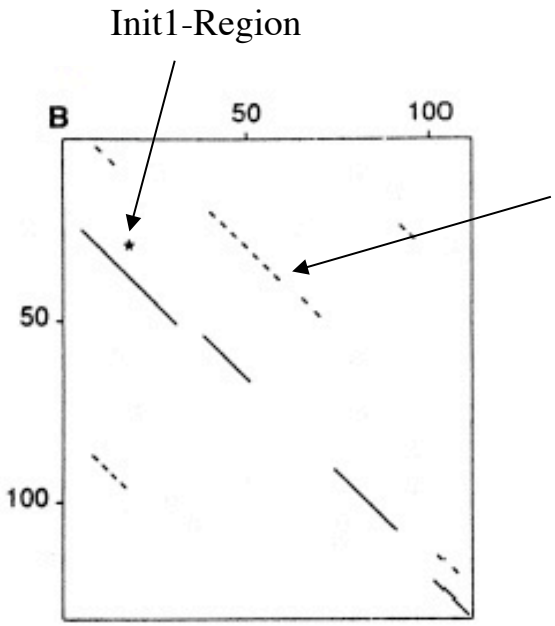
2. Berechne Index für Ktup-Positionen in der DB (einmal nur zu tun)

Folie: Tal Dagan, D'dorf

FastA

Ktup matches zwischen Query
und EINER DB-Sequenz

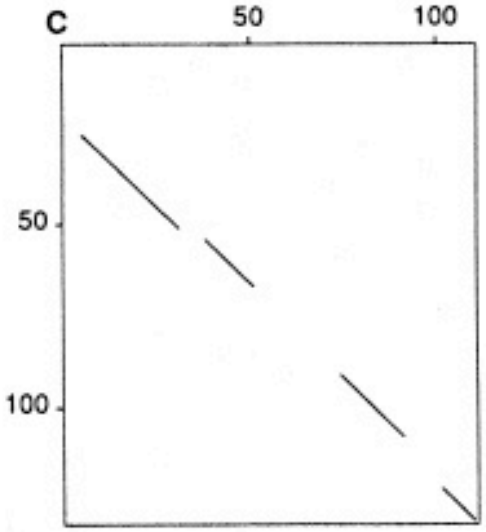




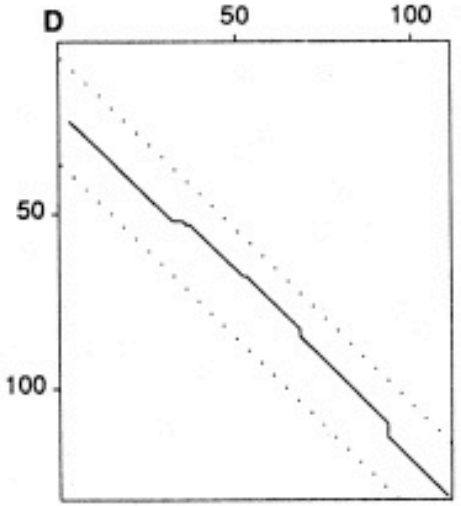
3. Identifiziere Position passender Ktups zwischen Query und allen DB-Einträgen unter Verwendung der Indices

4. Identifiziere die 10 besten Diagonalen durch scoring: Höchster score = „init1“

FastA

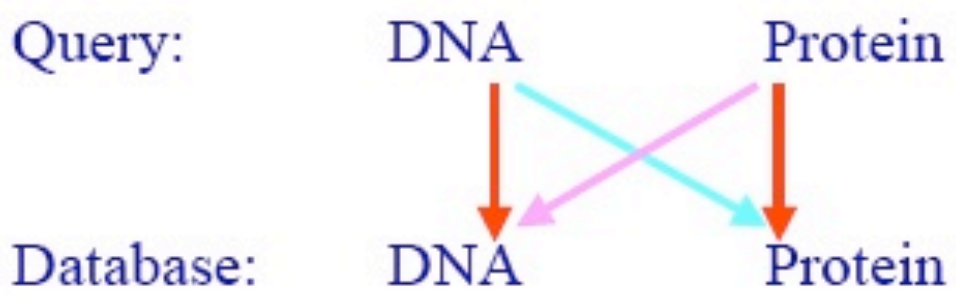


5. Verbinde high-score-Diagonalen
(ergibt initn-score)



6. Für höchste initn-Matches kalkuliere optimales lokales Aligment (SW) in Streifen um Diagonale (Breite meist 32 As)
(ergibt opt-score)

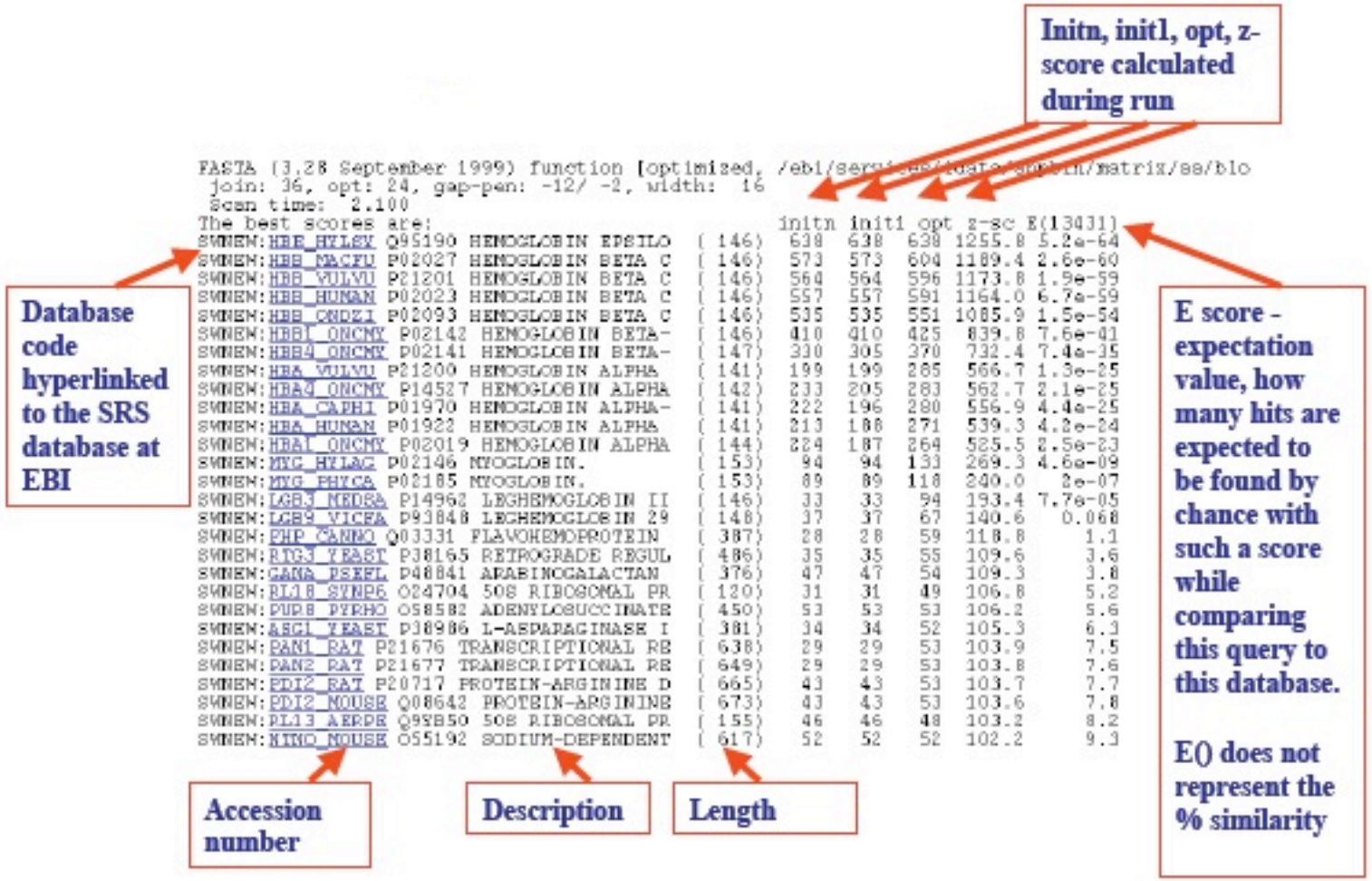
FastA-Programmfamilie



PROGRAM	FUNCTION	
fasta3	scan a protein or DNA sequence library for similar sequences	i
fastx/y3	compare a DNA sequence to a protein sequence database, comparing the translated DNA sequence in forward and reverse frames.	i
tfastx/y3	compares a protein to a translated DNA data bank	i
fasts3	compares linked peptides to a protein databank	i
fastf3	compares mixed peptides to a protein databank	i

Folie: Tal Dagan, D'dorf

FastA-Output




```
FASTA (3.28 September 1999) function [optimized, /ebi/serv
 join: 36, opt: 24, gap-pen: -12/ -2, width:  16
 Scan time:  2.100
The best scores are:                                      initn init1 opt  z-sc E(13431)
SWNEW:HBE_HYLSY Q95190 HEMOGLOBIN EPSILO  ( 146)  638   638   638 1255.8 5.2e-64
SWNEW:HBB_MACFU P02027 HEMOGLOBIN BETA C  ( 146)  573   573   604 1189.4 2.6e-60
SWNEW:HBB_VULVU P21201 HEMOGLOBIN BETA C  ( 146)  564   564   596 1173.8 1.9e-59
SWNEW:HBB_HUMAN P02023 HEMOGLOBIN BETA C  ( 146)  557   557   591 1164.0 6.7e-59
SWNEW:HBB_ONDZI P02093 HEMOGLOBIN BETA C  ( 146)  535   535   551 1085.9 1.5e-54
SWNEW:HBB1_ONCMY P02142 HEMOGLOBIN BETA-  ( 146)  410   410   425  839.8 7.6e-41
SWNEW:HBB4_ONCMY P02141 HEMOGLOBIN BETA-  ( 147)  330   305   370  732.4 7.4e-35
SWNEW:HBA_VULVU P21200 HEMOGLOBIN ALPHA   ( 141)  199   199   285  566.7 1.3e-25
SWNEW:HBA4_ONCMY P14527 HEMOGLOBIN ALPHA  ( 142)  233   205   283  562.7 2.1e-25
SWNEW:HBA_CAPHI P01970 HEMOGLOBIN ALPHA-  ( 141)  222   196   280  556.9 4.4e-25
SWNEW:HBA_HUMAN P01922 HEMOGLOBIN ALPHA   ( 141)  213   188   271  539.3 4.2e-24
SWNEW:HBA1_ONCMY P02019 HEMOGLOBIN ALPHA  ( 144)  224   187   264  525.5 2.5e-23
SWNEW:MYG_HYLAC P02146 MYOGLOBIN.         ( 153)   94    94   133  269.3 4.6e-09
SWNEW:MYG_PHYCA P02185 MYOGLOBIN.         ( 153)   89    89   118  240.0   2e-07
SWNEW:LGB3_MEDSA P14962 LEGHEMOGLOBIN II  ( 146)   33    33    94  193.4 7.7e-05
SWNEW:LGB9_VICFA P93848 LEGHEMOGLOBIN 29  ( 148)   37    37    67  140.6   0.068
SWNEW:FHP_CANNO Q03331 FLAVOHEMOPROTEIN   ( 387)   28    28    59  118.8     1.1
SWNEW:RTG3_YEAST P38165 RETROGRADE REGUL  ( 486)   35    35    55  109.6     3.6
SWNEW:GANA_PSEFL P48841 ARABINOGALACTAN   ( 376)   47    47    54  109.3     3.8
SWNEW:RL18_SYNP6 O24704 50S RIBOSOMAL PR  ( 120)   31    31    49  106.8     5.2
SWNEW:PUR8_PYRHO O58582 ADENYLOSUCCINATE  ( 450)   53    53    53  106.2     5.6
SWNEW:ASG1_YEAST P38986 L-ASPARAGINASE I  ( 381)   34    34    52  105.3     6.3
SWNEW:PAN1_RAT P21676 TRANSCRIPTIONAL RE  ( 638)   29    29    53  103.9     7.5
SWNEW:PAN2_RAT P21677 TRANSCRIPTIONAL RE  ( 649)   29    29    53  103.8     7.6
SWNEW:PDI2_RAT P20717 PROTEIN-ARGININE D  ( 665)   43    43    53  103.7     7.7
SWNEW:PDI2_MOUSE Q08642 PROTEIN-ARGININE  ( 673)   43    43    53  103.6     7.8
SWNEW:RL13_AERPE Q9YB50 50S RIBOSOMAL PR  ( 155)   46    46    48  103.2     8.2
SWNEW:NINO_MOUSE O55192 SODIUM-DEPENDENT  ( 617)   52    52    52  102.2     9.3
```


Folie: Tal Dagan, D‘dorf

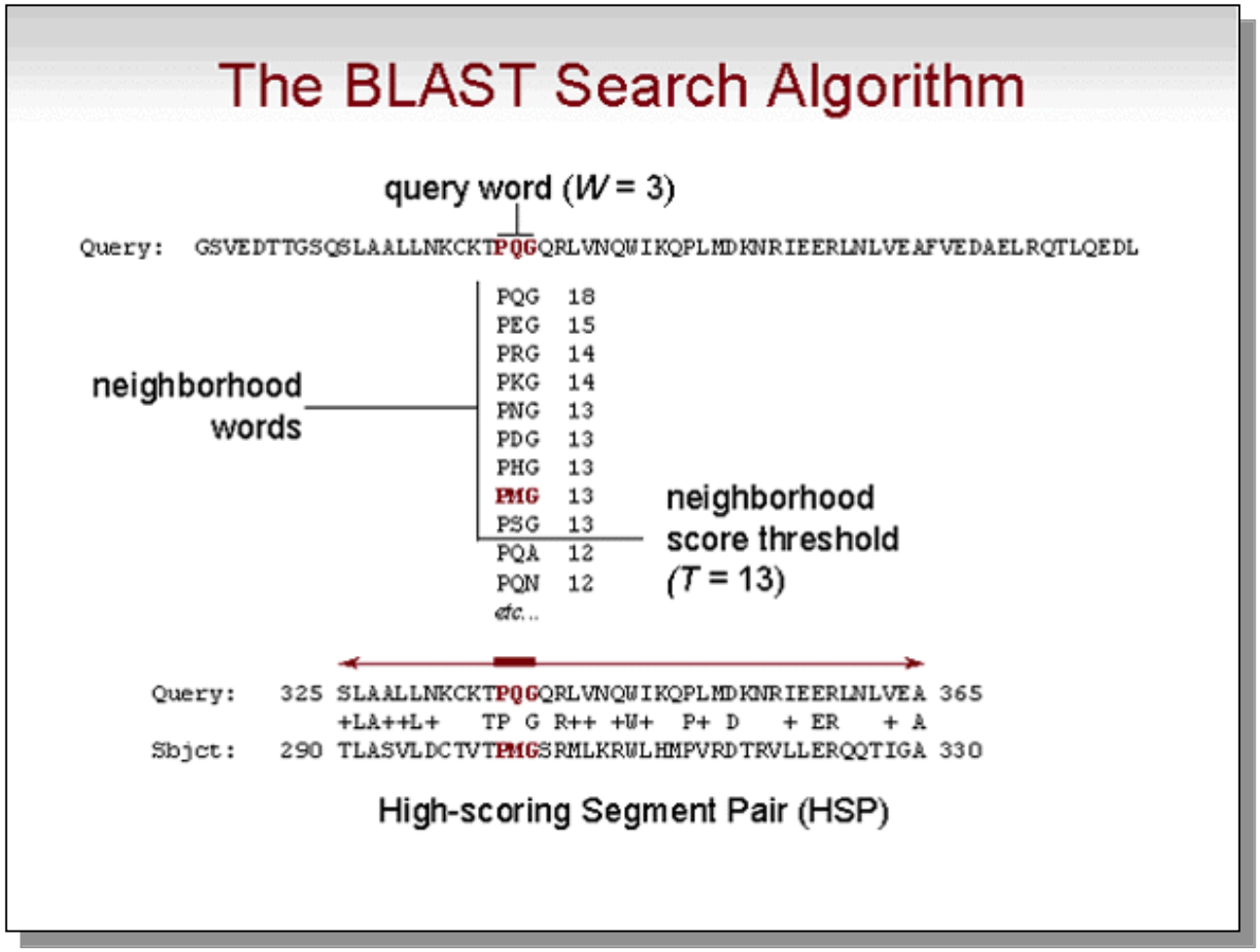
Bill Pearson says...

Which program when?

Problem	Program	Explanation	Alternate
Identify unknown protein	(1) `fasta3`	General protein comparison. Use *ktup=2* (the unknown default) for speed; *ktup=1* for a more sensitive search. Search first against the smallest library likely to contain a homolog (i.e. SwissProt rather than Genpept).	`blastp/`
	(2)`ssearch3`	10-50-fold slower than `fasta3` faster on Macs, but provides maximum sensitivity. No advantage for DNA comparisons.	`fasta3/ blastp`
	(3)`tfastx3/ tfasty3`	If a homolog cannot be found in the protein databases, check the DNA databases with `tfastx3` or `tfasty3`. `tfasty3` provides more accurate alignments, but is about 33% slower.	`tblastn/ tfasta`[a]
Identify structural DNA sequence	`fasta3`	If the DNA sequence encodes a protein, use protein sequence comparison first, then try translated protein sequence comparison (`fastx3/fasty3`). For repeated DNA sequences or structural RNAs, search first with *ktup=6* (the default), then *ktup=3*. Search with *ktup< 3* only for very short sequences (PCR primers).	`blastn`
Identify EST sequence	`fastx3/ fasty3`	Protein sequence comparison is far more sensitive than DNA comparison, so check first to see if the EST encodes a product homologous to a known protein. Current version searches forward strand only, so use `fastx3 -i` as well.	`fasta3/ blastx/ tblastx`
Confirm statistical significance	`prss3`	Use 500-2000 shuffles, and remember to normalize the statistical significance to the size of the database originally searched (typically 10,000 - 100,000 sequences).	

[a]No longer recommended.

BLAST

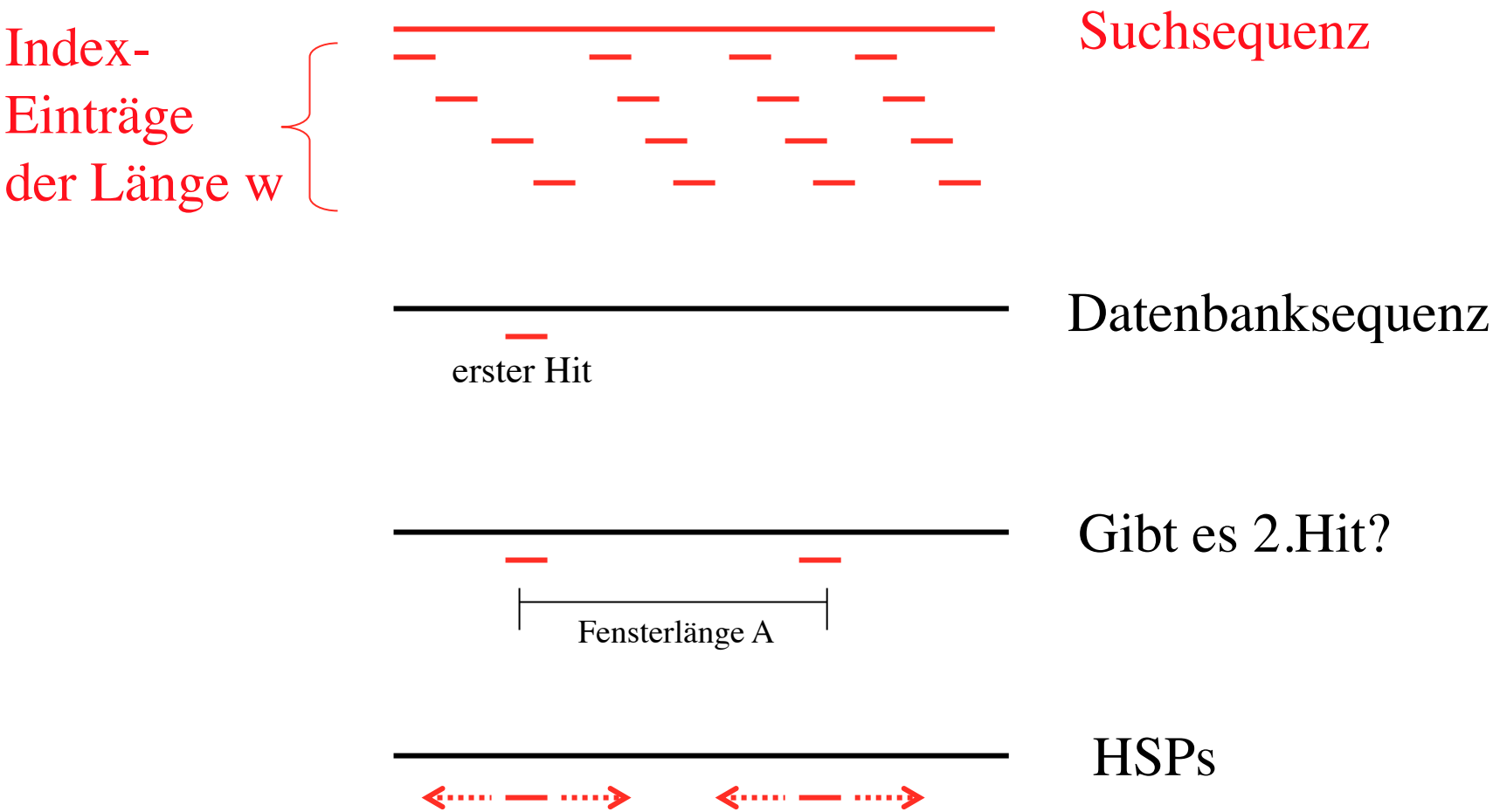


(word size W = 11 bei DNA)

- zunächst wird nach kurzen lokal passenden Abschnitten („words“) gesucht,

- dann versucht BLAST2.0, die Bereiche neben den „matching words“ unter Einbeziehung von Lücken zu optimieren

BLAST



zwei lokale Alignments,
Verknüpfung über Lücken falls möglich erlaubt

BLAST Parameter

- E-value: Wahrscheinlichkeit, dass ein solcher match zufällig in einer DB derselben Größe gefunden wird
- Filters: entfernt repetitive (low-complexity) Regionen
- Matrix: PAM, BLOSUM (machen manchmal den Unterschied!)
- Datenbanken: noch entscheidender!!! Vielfalt!
- Limits: z. B. nur in bestimmten Taxa suchen...
- Alignments und Descriptions: können bis zu Tausend angezeigt werden

Altschul et al. 1990, 1997

BLAST

- Programmfamilie
- schneller als FASTA!
- sucht lokale Alignments in DB
- ausgefeilte Such-Statistik (Karlin & Altschul 1990)
- Words können *ähnlich* sein (ungleich FastA)
- Low-complexity-Regionen werden entfernt

BLAST :
Endecke die Möglichkeiten...

blastn	DNA-Sequenz ÷ DNA-DB > nur nahe Verwandtschaft; beide Stränge verglichen
blastp	As-Sequenz ÷ Protein-DB > entfernte Verwandtschaft (default:BLOSUM62)
blastx	DNA-Seq > in 6 Leserahmen translatiert ÷ Protein-DB > findet mögliche Proteine in einer nicht-charakterisierten DNA-Sequenz (z.B. EST)!

BLAST :
Endecke die Möglichkeiten...

tblastn	As-Seq gegen DNA-DB (6-frame translatiert!) > findet nicht-annotierte Genregionen in DNA-DB-Sequenzen
tblastx	6-frame-Translation einer DNA-Seq ÷ 6-frame-Translation einer DNA-DB > Analyse von ESTs auf Proteinebene zur Detektion entfernter Verwandtschaft > kann nicht mit nr-DB benutzt werden (zu aufwendig)

MegaBLAST

- sehr schnelle Nt-Suche (10x schneller als BLASTN)
- für sehr ähnliche Sequenzen
- word size : default 28
- nicht-affine gap penalty: schneller, weniger Memory erforderlich
  > mehr, aber kürzere gaps

Anwendung: schnelles Alignment zwischen ähnlichen Sequenzen, z. B. menschliche cDNA an menschliches Genom alignen

Ma et al. 2002

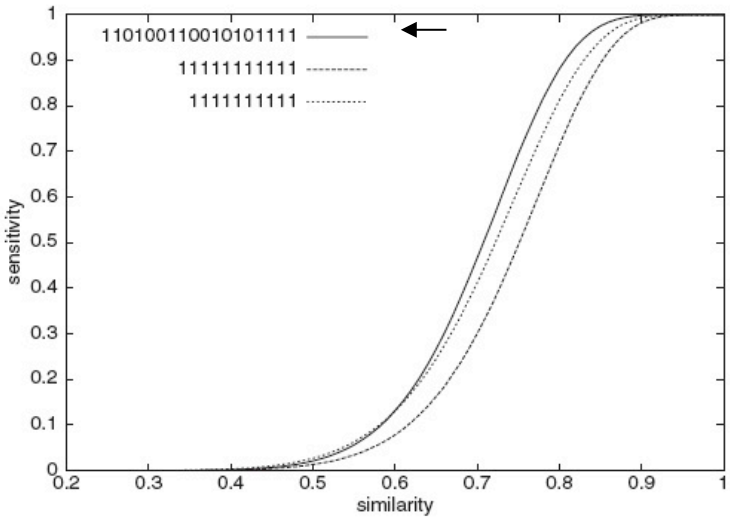
Discontiguous MegaBLAST

- schaut nicht -wie BLASTN und MegaBLAST- nach exakten word-matches (wird unproduktiv bei < 80 % id)
- Suche nach ‚discontiguous words' besser (weniger, aber signifikantere hits):

random hits which slow down the computation. We use a new idea that allows us to have a higher probability of a hit in a homologous region, even while having somewhat lower expected number of random hits.

Blast looks for matches of k (default $k = 11$ in Blastn and $k = 28$ in MegaBlast) consecutive letters as seeds. Instead we propose to use *non*consecutive k letters as seeds. We call the relative positions of the k letters a *model*, and k its *weight*.

This seemingly simple change has a surprisingly large effect on sensitivity. An appropriately chosen model can have a significantly higher probability of having at least one hit in a homologous region, compared to Blast's consecutive seed model, even while having a lower *expected* number of hits†. For example, in a region of



Anwendung: schnelle Nt- Suche bei entfernt verwandten DNA-Sequenzen auf EST und Genomebene (z.B. mit Datenbank ‚TRACES')

BLAT

„BLAST-like alignment tool“

- DNA-BLAT findet 40 Bp (>95% id) oder länger extrem schnell (500xBLAST)
- Protein-BLAT findet 20 aa (>80%id)
- **Index (DNA) enthält alle nicht-überlappenden 11-mere des Genoms (1 Gb RAM)**
- Index wird gebraucht um passende Regionen im Genom schnell zu identifizieren, die dann für genaueren Vergleich „hochgeladen“ werden

Anwendung: lokales Alignment zwischen längeren Sequenzen, z. B. cDNA an menschliches Genom alignen

Score-Statistik

- BLAST berechnet Signifikanz aus Simulationen mit „normalen“, d. h. durchschnittlichen Sequenzen

- FASTA erstellt Verteilung von similarity scores während der DB-Suche (selektiert 60000 scores aus DB mit realen Sequenzen)

- PRSS (aus FASTA-Paket) berechnet Signifikanz durch Erstellen Hunderter von „shuffled (random) sequences“ gleicher Länge und Zusammensetzung

Score-Statistik

- Reale unverwandte Sequenzen haben similarity scores wie zufällige Sequenzen

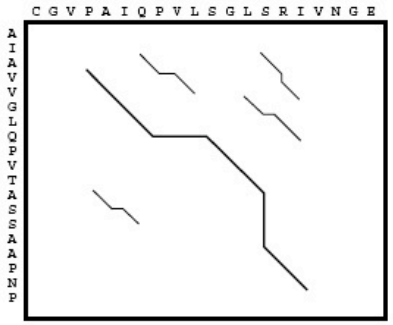
- Wenn die Similarität statistisch signifikant nicht ZUFÄLLIG ist, muss sie daher auf VERWANDSCHAFT schließen lassen

- E-Values < 0.001 sind erfahrungsgemäß Treffer

DNA vs. Protein

The best scores are:		DNA E(188,018)	tfastx3 E(187,524)	prot. E(331,956)
DMGST	D.melanogaster GST1-1	1.3e-164	4.1e-109	1.0e-109
MDGST1	M.domestica GST-1 gene	2e-77	3.0e-95	1.9e-76
LUCGLTR	Lucilia cuprina GST	1.5e-72	5.2e-91	3.3e-73
MDGST2A	M.domesticus GST-2 mRNA	9.3e-53	1.4e-77	1.6e-62
MDNF1	M.domestica nf1 gene. 10	4.6e-51	2.8e-77	2.2e-62
MDNF6	M.domestica nf6 gene. 10	2.8e-51	4.2e-77	3.1e-62
MDNF7	M.domestica nf7 gene. 10	6.1e-47	9.2e-77	6.7e-62
AGGST15	A.gambiae GST mRNA	3.1e-58	4.2e-76	4.3e-61
CVU87958	Culicoides GST	1.8e-41	4.0e-73	3.6e-58
AGG3GST11	A.gambiae GST1-1 mRNA	1.5e-46	2.8e-55	1.1e-43
BMO6502	Bombyx mori GST mRNA	1.1e-23	8.8e-50	5.7e-40
AGSUGST12	A.gambiae GST1-1 gene	2.3e-16	4.5e-46	5.1e-37
MOTGLUSTRA	Manduca sexta GST	5.7e-07	2.5e-30	8.0e-25
RLGSTARGN	R.legominosarum *gstA*	0.0029	3.2e-13	1.4e-10
HUMGSTT2A	H. sapiens GSTT2	0.32	3.3e-10	2.0e-09
HSGSTT1	H.sapiens GSTT1 mRNA	7.2	8.4e-13	3.6e-10
ECAE000319	E. coli hypothet. prot.	—	4.7e-10	1.1e-09
MYMDCMA	Methyl. dichlorometh. DH	—	1.1e-09	6.9e-07
BCU19883	Burkholderia maleylacetate red.	—	1.2e-09	1.1e-08
NFU43126	Naegleria fowleri GST	—	3.2e-07	0.0056
SP505GST	Sphingomonas paucim	—	1.8e-06	0.0002
EN1838	H. sapiens maleylaceto. iso.	—	2.1e-06	5.9e-06
HSU86529	Human GSTZ1	—	3.0e-06	8.0e-06
SYCCPNC	Synechocystis GST	—	1.2e-05	9.5e-06
HSEF1GMR	H.sapiens EF1g mRNA	—	9.0e-05	0.00065

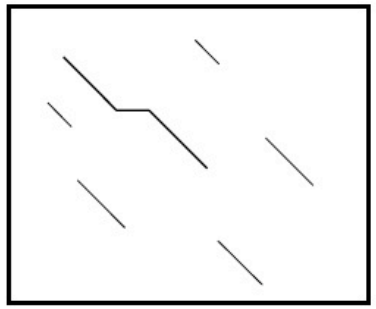
Was ist besser?



Smith-Waterman

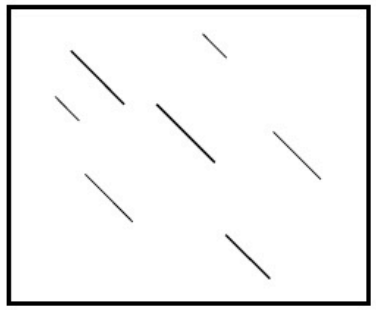
time: 10:00 min

ssearch



FASTA

time: 2:00 min



BLAST

time: 20 sec

Sensitivität

Speed

Bill Pearson says...

BLAST and *FASTA*
Which program when?

Blast for proteins
Blast for speed
FASTA for DNA
FASTA for frameshifts
FASTA for accurate statistics
(protein and coding DNA)
SSEARCH for optimal
(be careful with PSI-BLAST)

Bill Pearson says...

Program		
BLAST	FASTA	Functio n
`blastp`	`fasta 3`	General protein sequence similarity searches. **`blastp`** is faster and can show alignments between several domains in the same sequence. **`fasta3`** displays a Smith-Waterman final alignment and produces more accurate statistical estimates in some cases.
`blastn`	`fasta 3`	DNA sequence comparison. **`blastn`** is highly optimized for speed; it uses a fixed word size (11 nucleotides) and scoring matrix that are inappropriate for some problems (e.g. searching for PCR primer matches).
`blastx`	`fastx3/` `fasty3`	Compare a translated DNA to a protein sequence database. While **`blastx`** does six independent searches (one for each of the six frames), **`fastx3`** and **`fasty3`** effectively does a single forward (or backward) search, which allows frameshifts in computing the similarity score and alignments. As a result, **`fastx3`** and **`fasty3`** are more sensitive and can produce much better alignments than **`blastx`** when the DNA sequence has frameshift errors.
`tblastn`	`tfastx3/` `tfasty3`	Compare a protein sequence to a DNA sequence database, translating in the three forward and reverse frames. Again, **`tfastx3`** and **`tfasty3`** provide more accurate alignments than **`tblastn`** when the DNA sequences have frameshift errors.
	`tblastx`	Compare a DNA query sequence to a DNA library, translating both sequences in all six frames and scoring using a protein substitution matrix (BLOSUM62). **`fasta3`** with *ktup=6* (the default) provides a similar function, but does not use a protein scoring matrix.

Anhang

Dotlet

Figure 1: The `dotlet` menu bar

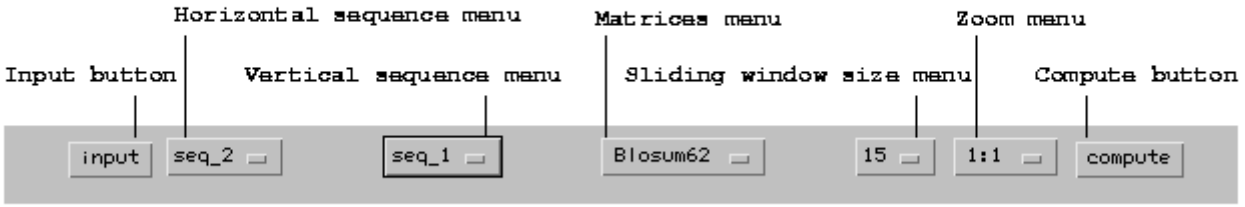


Figure 3: The Dots window, with no grayscale adjustments

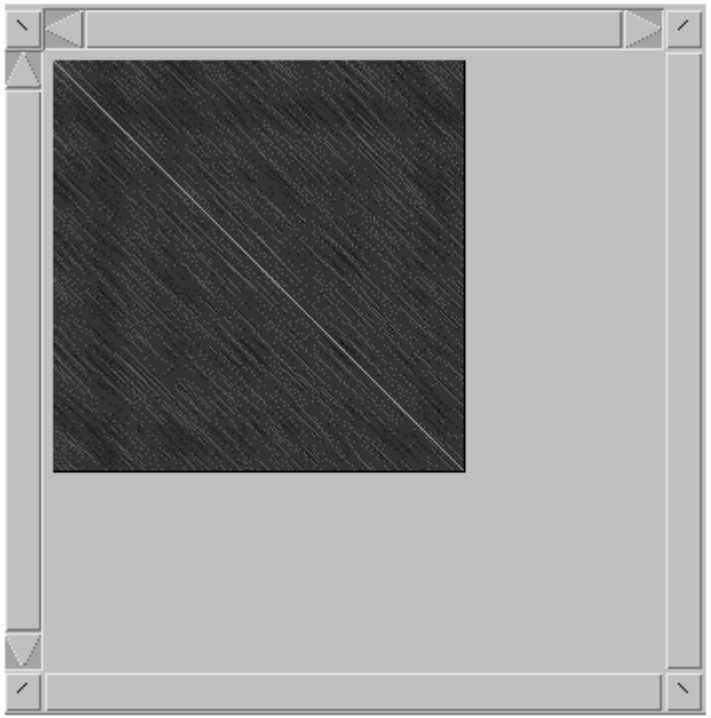


Figure 4: The Histogram window, unadjusted

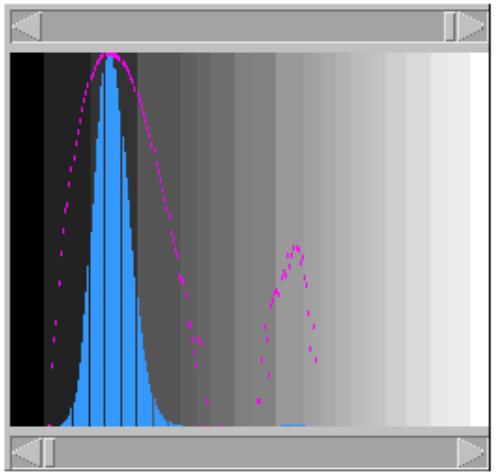
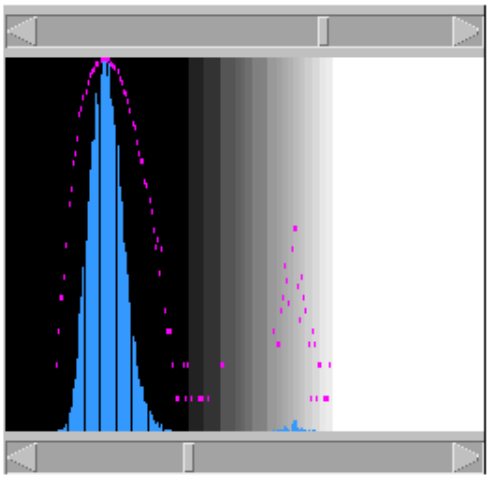
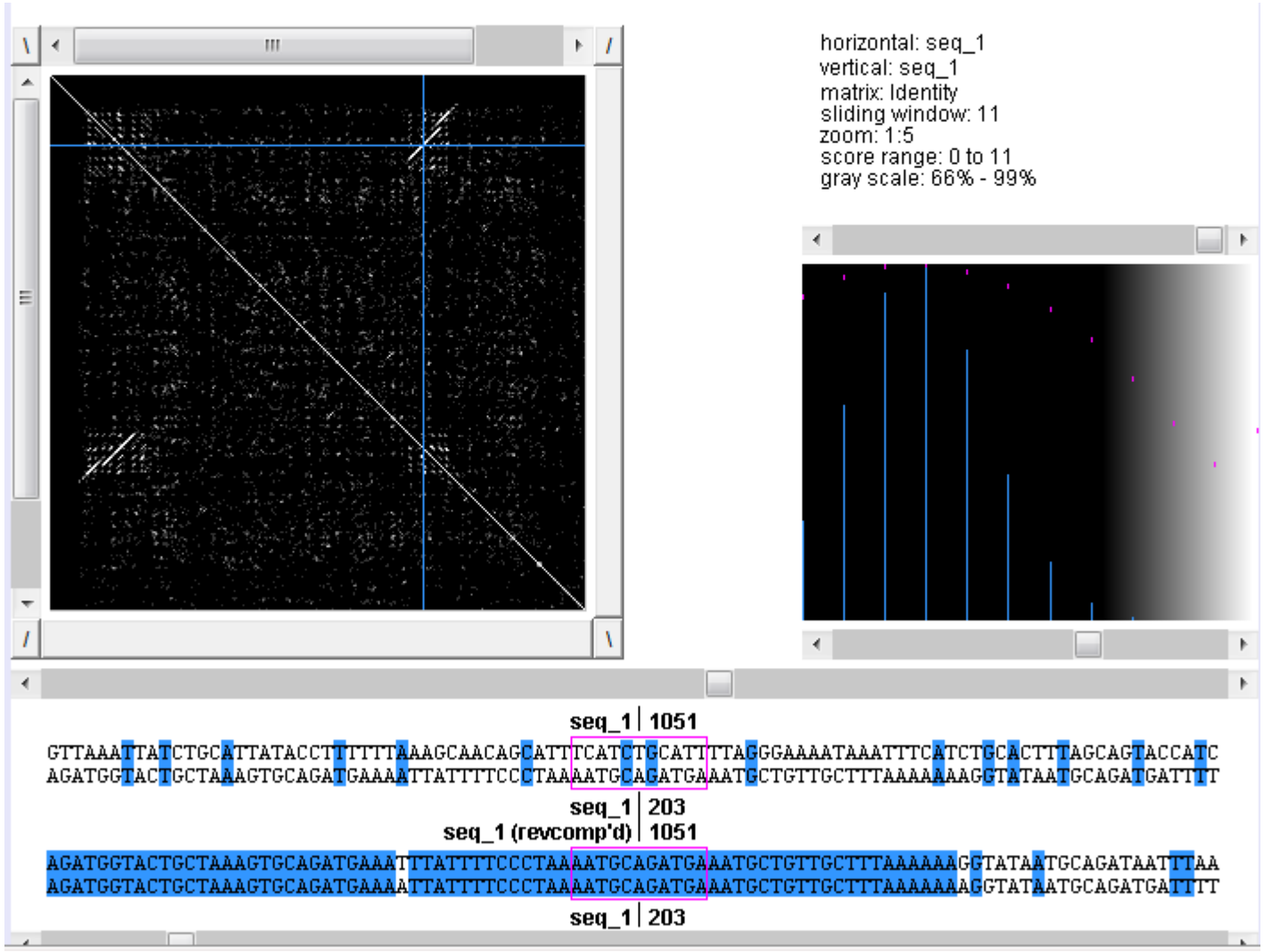


Figure 5: The Histogram window, after adjusting the grayscale

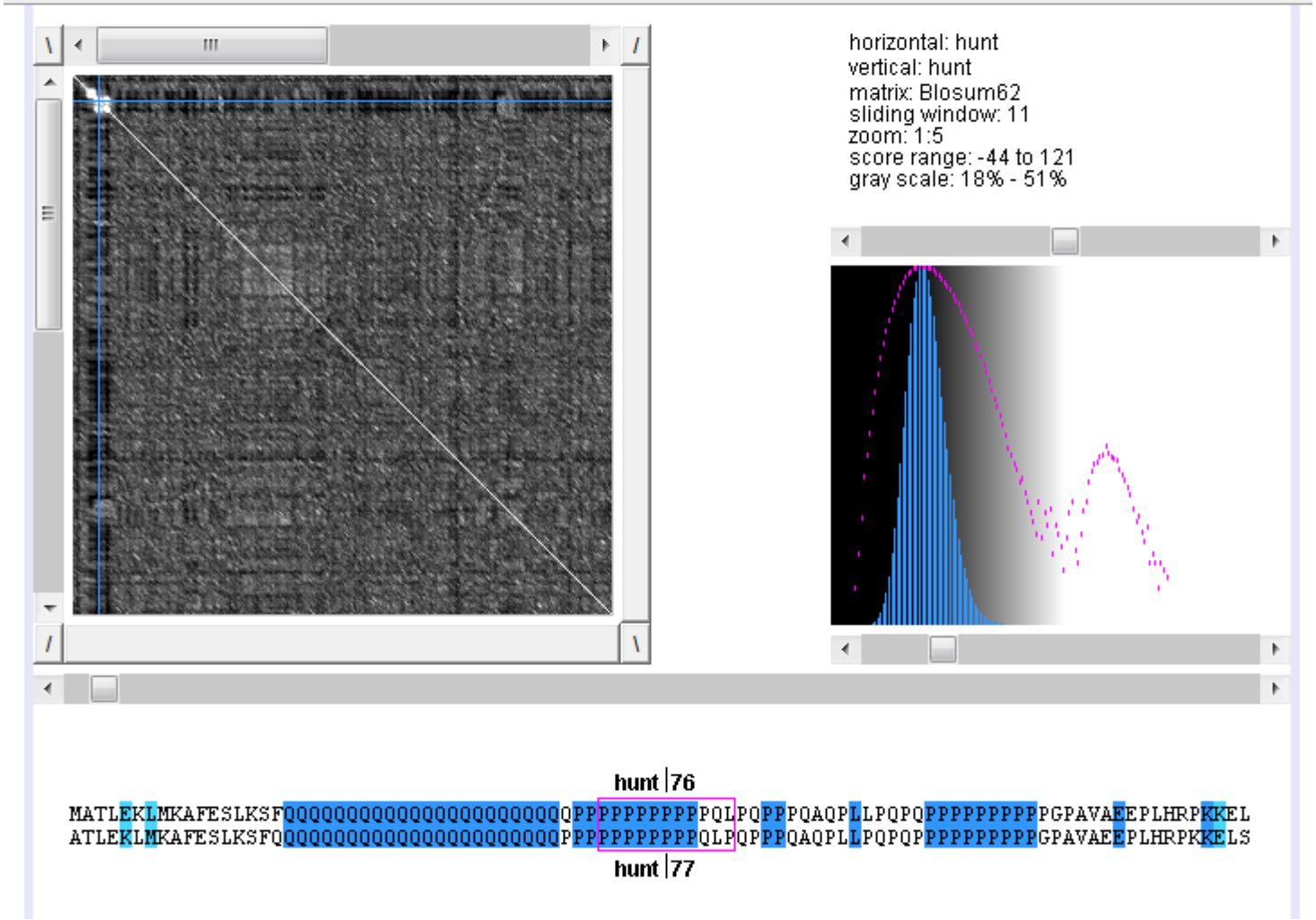


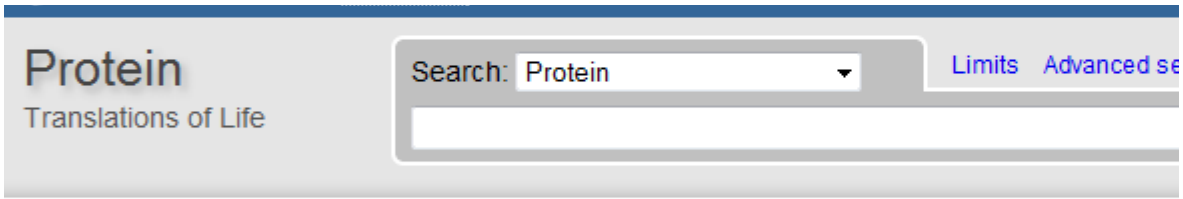
Score

X56335 - Inverted Repeat (Foldback Transposon)



Huntingtin


horizontal: hunt
vertical: hunt
matrix: Blosum62
sliding window: 11
zoom: 1:5
score range: -44 to 121
gray scale: 18% - 51%
hunt 76
MATLEKLMKAFESLKSFQQQQQQQQQQQQQQQQQQQQQQQQQQQPPPPPPPPPPQLPQPPPQAQPLLPQPQPPPPPPPPPPGPAVAEEPLHRPKKEL
ATLEKLMKAFESLKSFQQQQQQQQQQQQQQQQQQQQQQQQQQQPPPPPPPPPPQLPQPPPQAQPLLPQPQPPPPPPPPPPGPAVAEEPLHRPKKELS
hunt 77



Display Settings: FASTA

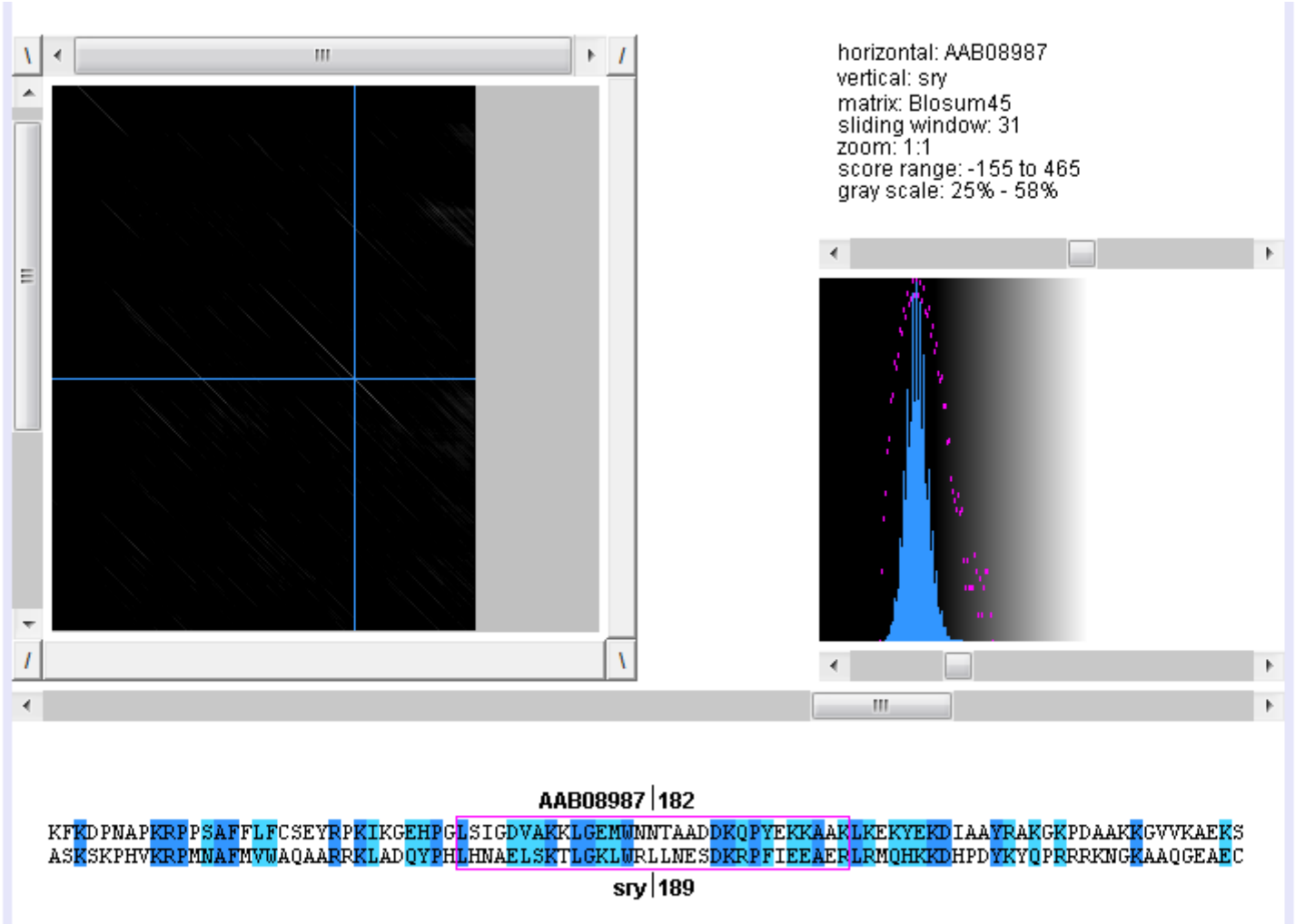
huntingtin [Homo sapiens]

NCBI Reference Sequence: NP_002102.4

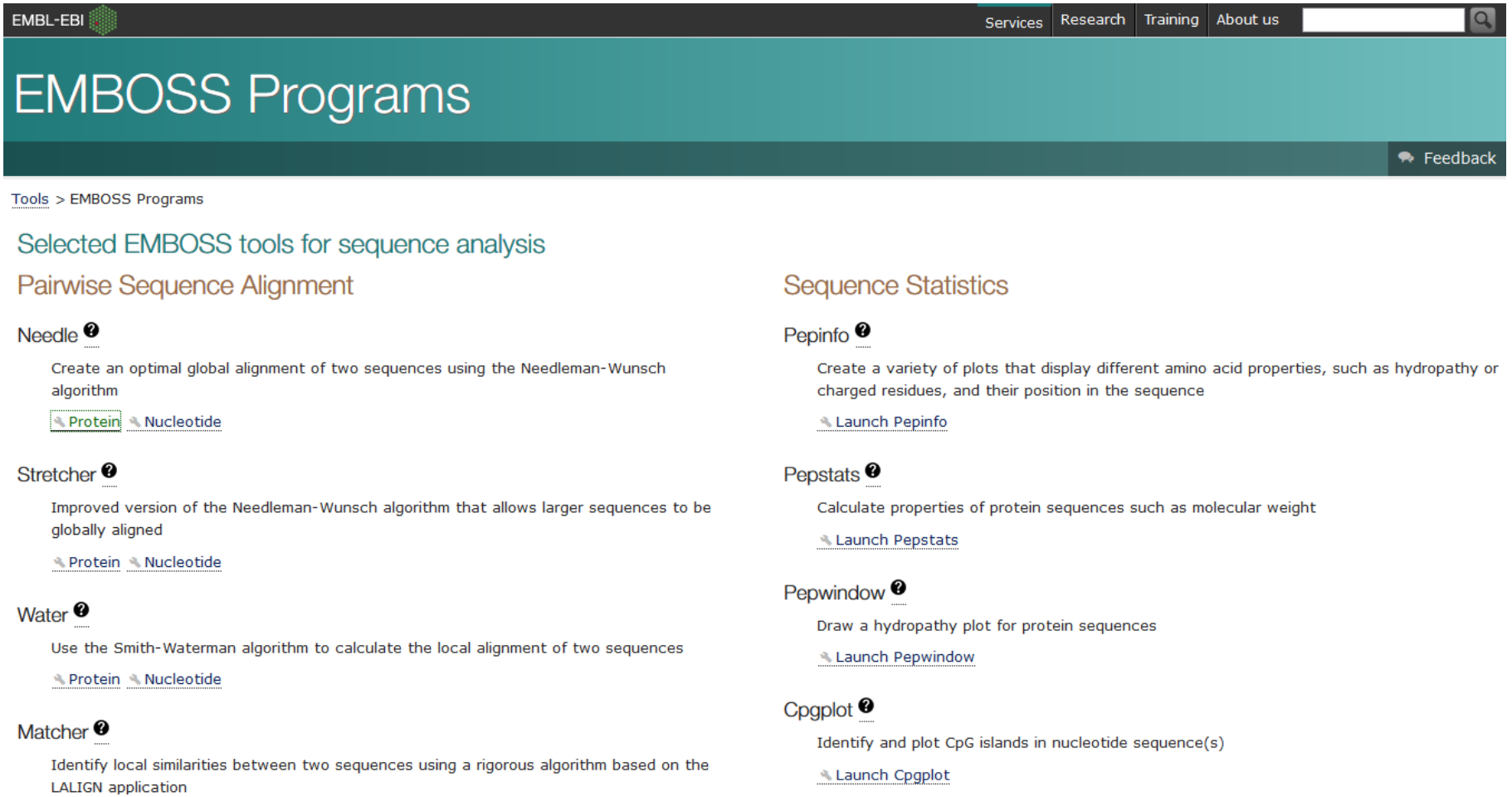
GenPept Graphics

```
>gi|90903231|ref|NP_002102.4| huntingtin [Homo sapiens]
MATLEKLMKAFESLKSFQQQQQQQQQQQQQQQQQQQQQQPPPPPPPPPPQLPQPPPQAQPLLPQPQPP
PPPPPPPPGPAVAEEPLHRPKKELSATKKDRVNHCLTICENIVAQSVRNSPEFQKLLGIAMELFLLCSDD
AESDVRMVADECLNKVIKALMDSNLPRLQLELYKEIKKNGAPRSLRAALWRFAELAHLVRPQKCRPYLVN
LLPCLTRTSKRPEESVQETLAAAVPKIMASFGNFANDNEIKVLLKAFIANLKSSSPTIRRTAAGSAVSIC
QHSRRTQYFYSWLLNVLLGLLVPVEDEHSTLLILGVLLTLRYLVPLLQQQVKDTSLKGSFGVTRKEMEVS
PSAEQLVQVYELTLHHTQHQDHNVVTGALELLQQLFRTPPPELLQTLTAVGGIGQLTAAKEESGGRSRSG
SIVELIAGGGSSCSPVLSRKQKGKVLLGEEEALEDDSESRSDVSSSALTASVKDEISGELAASSGVSTPG
SAGHDIITEQPRSQHTLQADSVDLASCDLTSSATDGDEEDILSHSSSQVSAVPSDPAMDLNDGTQASSPI
SDSSQTTTEGPDSAVTPSDSSEIVLDGTDNQYLGLQIGQPQDEDEEATGILPDEASEAFRNSSMALQQAH
LLKNMSHCRQPSDSSVDKFVLRDEATEPGDQENKPCRIKGDIGQSTDDDSAPLVHCVRLLSASFLLTGGK
NVLVPDRDVRVSVKALALSCVGAAVALHPESFFSKLYKVPLDTTEYPEEQYVSDILNYIDHGDPQVRGAT
AILCGTLICSILSRSRFHVGDWMGTIRTLTGNTFSLADCIPLLRKTLKDESSVTCKLACTAVRNCVMSLC
SSSYSELGLQLIIDVLTLRNSSYWLVRTELLETLAEIDFRLVSFLEAKAENLHRGAHHYTGLLKLQERVL
NNVVIHLLGDEDPRVRHVAAASLIRLVPKLFYKCDQGQADPVVAVARDQSSVYLKLLMHETQPPSHFSVS
TITRIYRGYNLLPSITDVTMENNLSRVIAAVSHELITSTTRALTFGCCEALCLLSTAFPVCIWSLGWHCG
```

HMG-SRY


horizontal: AAB08987
vertical: sry
matrix: Blosum45
sliding window: 31
zoom: 1:1
score range: -155 to 465
gray scale: 25% - 58%
AAB08987|182
KFKDPNAPKRPPSAFFLFCSEYRPKIKGEHPGLSIGDVAKKLGEMWNNTAADDKQPYEKKAAKLKEKYEKDIAAYRAKGKPDAAKKGVVKAEKS
ASKSKPHVKRPMNAFMVWAQAARRKLADQYPHLHNAELSKTLGKLWRLLNESDKRPFIEEAERLRMQHKKDHPDYKYQPRRRKNGKAAQGEAEC
sry|189

Welches Tool auf der EMBOSS-Seite halten Sie intuitiv für geeignet?



Needleman-Wunsch Alignment
Standard Einstellungen

```
Aligned_sequences: 2
# 1: EMBOSS_001
# 2: EMBOSS_001
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 158
# Identity:      68/158 (43.0%)
# Similarity:    91/158 (57.6%)
# Gaps:          19/158 (12.0%)
# Score: 328.5
#
#
#=======================================

EMBOSS_001         1 -------------APLSADQASLVKSTWAQVRNSEVEILAAVFTAYPDIQ     37
                                  :.||.||..||:||:.:|:...|.||.|||.|.|.||
EMBOSS_001         1 MKFIILALCVAAASALSGDQIGLVQSTYGKVKGDSVGILYAVFKADPTIQ     50

EMBOSS_001        38 ARFPQFAGKDVASIKDTGAFATHAGRIVGFVSEIIALIGNESNAPAVQTL     87
                     |.||||.|||:.:||....|:||||||||:..:|      .:.|.:...
EMBOSS_001        51 AAFPQFVGKDLDAIKGGAEFSTHAGRIVGFLGGVI------DDLPNIGKH     94

EMBOSS_001        88 VGQLAASHKARGISQAQFNEFRAGLVSYVSSNVAWNAAAESAWTAGLDNI    137
                     |..|.|:||.||::.||||.|||..::|:..:|.:.||.|:||.|..|..
EMBOSS_001        95 VDALVATHKPRGVTHAQFNNFRAAFIAYLKGHVDYTAAVEAAWGATFDAF    144

EMBOSS_001       138 FGLLFAAL    145
                     ||.:||.:
EMBOSS_001       145 FGAVFAKM    152
```

Erniedrigen Sie einmal drastisch die gap penalty-Werte (z.B. von 10 auf 1)
Was passiert?

```
# Aligned_sequences: 2
# 1: EMBOSS_001
# 2: EMBOSS_001
# Matrix: EBLOSUM62
# Gap_penalty: 1.0
# Extend_penalty: 0.0
#
# Length: 168
# Identity:      77/168 (45.8%)
# Similarity:    99/168 (58.9%)
# Gaps:          39/168 (23.2%)
# Score: 380.992
#
#
#=======================================

EMBOSS_001         1 ------APL-----SA---DQASLVKSTWAQVR-NSEVEILAAVFTAY-P     34
                       | |     ||   ||..||:||:.:|: :| |.||.|||.|  |
EMBOSS_001         1 MKFIILA-LCVAAASALSGDQIGLVQSTYGKVKGDS-VGILYAVFKA-DP     47

EMBOSS_001        35 DIQARFPQFAGKDV-ASIKDTGA-FATHAGRIVGFVSEIIA-L--IGNES     79
                     .|||.||||.|||: | ||. || |:||||||||:..:|. |  || :
EMBOSS_001        48 TIQAAFPQFVGKDLDA-IKG-GAEFSTHAGRIVGFLGGVIDDLPNIG-K-     93

EMBOSS_001        80 NAPAVQTLVGQLAASHKARGISQAQFNEFRAGLVSYVSSNVAWNAAAESA    129
                     :   |..||    |:||.||::.||||.|||..::|:..:|.:.||.|:|
EMBOSS_001        94 H---VDALV----ATHKPRGVTHAQFNNFRAAFIAYLKGHVDYTAAVEAA    136

EMBOSS_001       130 WTAG--LDNIFGLLFAAL    145
                     |  |  .|..||.:||.:
EMBOSS_001       137 W--GATFDAFFGAVFAKM    152
```

Probieren Sie den alternativ angebotenen Alignment-Algorithmus aus: was ändert sich? → lokales S-W Alignment

```
# Aligned_sequences: 2
# 1: EMBOSS_001
# 2: EMBOSS_001
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 143
# Identity:      68/143 (47.6%)
# Similarity:    90/143 (62.9%)
# Gaps:           6/143 ( 4.2%)
# Score: 328.5
#
#
#=======================================

EMBOSS_001         3 LSADQASLVKSTWAQVRNSEVEILAAVFTAYPDIQARFPQFAGKDVASIK     52
                     ||.||..||:||:.:|:...|.||.|||.|.|.|||.||||.|||:.:||
EMBOSS_001        16 LSGDQIGLVQSTYGKVKGDSVGILYAVFKADPTIQAAFPQFVGKDLDAIK     65

EMBOSS_001        53 DTGAFATHAGRIVGFVSEIIALIGNESNAPAVQTLVGQLAASHKARGISQ    102
                     ....|:||||||||||:..:|      .:.|.:...|..|.|:||.||:.
EMBOSS_001        66 GGAEFSTHAGRIVGFLGGVI------DDLPNIGKHVDALVATHKPRGVTH    109

EMBOSS_001       103 AQFNEFRAGLVSYVSSNVAWNAAAESAWTAGLDNIFGLLFAAL    145
                     ||||.|||..::|:..:|.:.||.|:||.|..|..||.:||.:
EMBOSS_001       110 AQFNNFRAAFIAYLKGHVDYTAAVEAAWGATFDAFFGAVFAKM    152
```

Hämoglobin-Untereinheiten

Standard Einstellungen Globales Needleman-Wunsch Alignment

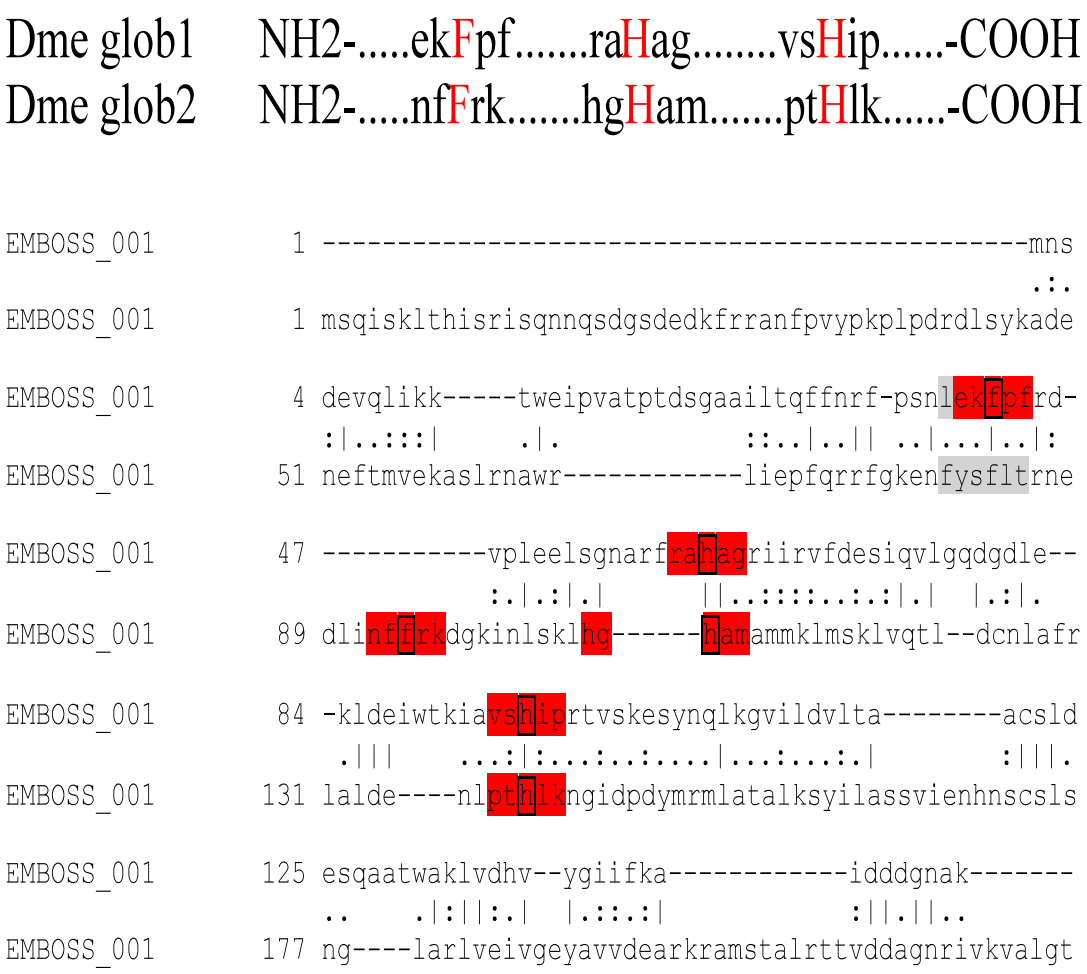
```
Aligned_sequences: 2
# 1: EMBOSS_001
# 2: EMBOSS_001
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 149
# Identity:      65/149 (43.6%)
# Similarity:    90/149 (60.4%)
# Gaps:           9/149 ( 6.0%)
# Score: 292.5
#
#
#=======================================

EMBOSS_001         1 MV-LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-D     48
                     || |:|.:|:.|.|.||||  :..|.|.|||.|:.:.:|.|:.:|..| |
EMBOSS_001         1 MVHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGD     48

EMBOSS_001        49 LS-----HGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLR     93
                     ||     .|:.:||.|||||..|.::.:||:|::....:.||:||..||.
EMBOSS_001        49 LSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLH     98

EMBOSS_001        94 VDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR    142
                     |||.||:||.:.|:..||.|...||||.|.|:..|.:|.|:..|..||.
EMBOSS_001        99 VDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH    14
```

Alignen Sie die Sequenzen der beiden Untereinheiten des Hämoglobins und erhöhen Sie einmal drastisch die gap penalty-Werte. Was ändert sich?

```
Aligned_sequences: 2
# 1: EMBOSS_001
# 2: EMBOSS_001
# Matrix: EBLOSUM62
# Gap_penalty: 100.0
# Extend_penalty: 10.0
#
# Length: 147
# Identity:      44/147 (29.9%)
# Similarity:    62/147 (42.2%)
# Gaps:           5/147 ( 3.4%)
# Score: 146.0
#
#
#=======================================

EMBOSS_001          1 -----MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFP     45
                        .....|.....................:......|..:.....
EMBOSS_001          1 MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLS     50

EMBOSS_001         46 HFDLSHGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVD     95
                      ..|...|:.:||.|||||..|.:::.:||:|::....:.||:||..||.||
EMBOSS_001         51 TPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVD    100

EMBOSS_001         96 PVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR    142
                      |.||:||.:.|:..||.|...||||.|.|:..|.:|.|:..|..||.
EMBOSS_001        101 PENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH    147
```

Probieren Sie ein paarweises Alignment mit den folgenden zwei Sequenzen, die eine mRNA und eine zu ihr passende microRNA (miRNA) darstellen.

```
# Aligned_sequences: 2
# 1: EMBOSS_001
# 2: EMBOSS_001
# Matrix: EDNAFULL
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 24
# Identity:      18/24 (75.0%)
# Similarity:    18/24 (75.0%)
# Gaps:           3/24 (12.5%)
# Score: 48.0
#
#
#=======================================

EMBOSS_001         1 ACUA-CCUGCACUGUA-AGCACUU     22
                     |||| |.|.||| .|| |||||||
EMBOSS_001        24 ACUAGCAUCCAC-AUAGAGCACUU     46
```




```
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 250
# Identity:      37/250 (14.8%)
# Similarity:    67/250 (26.8%)
# Gaps:         125/250 (50.0%)
# Score: 39.5
```

• Welche Substitutionsmatrix würden Sie anstatt der default-Matrix wählen? Wechseln Sie auf eine besser geeignete Matrix! Notieren Sie die Werte! Wird das Alignment besser?

```
# Matrix: EBLOSUM45
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 239
# Identity:      37/239 (15.5%)
# Similarity:    69/239 (28.9%)
# Gaps:         103/239 (43.1%)
# Score: 100.0
#
#
#=======================================
```

Dme glob1 NH2-.....ekFpf.....raHag.....vsHip.....-COOH
Dme glob2 NH2-.....nfFrk.....hgHam.....ptHlk.....-COOH

```
EMBOSS_001         1 ----------------------------------------------mns      3
                                                                   .:.
EMBOSS_001         1 msqisklthisrisqnnqsdgsdedkfrranfpvypkplpdrdlsykade     50

EMBOSS_001         4 devqlikk-----tweipvatptdsgaailtqffnrfpsnlekfpfrd--     46
                     :|..:::|     .|.:........|...:..|:.| ..:|..|..:|
EMBOSS_001        51 neftmvekaslrnawrliepfqrrfgkenfysfltr-nedlihffrkdgk     99

EMBOSS_001        47 vpleelsgnarfrahagriirvfdesiqvlgqdgdle---kldeiwtkia     93
                     :.|.:|.|      ||..::::..:.:|.|  |.:|.   .|||...
EMBOSS_001       100 inlsklhg------hamammklmsklvqtl--dcnlafrlaldenlp---    138

EMBOSS_001        94 vshiprtvskesynqlkgvildvlta--------acsldesqaatwaklv    135
                      :|:...:..:....|...:...:.|        :|||...    .|:||
EMBOSS_001       139 -thlkngidpdymrmlatalksyilassvienhnscslsng----larlv    183

EMBOSS_001       136 dhv--ygiifka------------idddgnak-------    153
                     :.|  |.::.:|            :||.||..
EMBOSS_001       184 eivgeyavvdearkramstalrttvddagnrivkvalgt    222
```

• Ändern Sie nun die gap extension penalty schrittweise zunächst auf 1.0, dann 5.0. Betrachten Sie den letzten Fall: Erfüllt das alignment nun besser die strukturbiologischen Vorgaben?

```
# Aligned_sequences: 2
# 1: EMBOSS_001
# 2: EMBOSS_001
# Matrix: EBLOSUM45
# Gap_penalty: 10.0
# Extend_penalty: 5.0
#
# Length: 230
# Identity:      23/230 (10.0%)
# Similarity:    59/230 (25.7%)
# Gaps:          85/230 (37.0%)
# Score: 44.0
#
#
#=======================================

EMBOSS_001         1 --------------------------------------------------      0

EMBOSS_001         1 msqisklthisrisqnnqsdgsdedkfrranfpvypkplpdrdlsykade     50

EMBOSS_001         1 --mnsdevqlikktweipvatptdsgaailtqffnrfpsnlekfpfrdvp     48
                       ....|...::..|.:........|...:..|:.| ..:|..| ||...
EMBOSS_001        51 neftmvekaslrnawrliepfqrrfgkenfysfltr-nedlinf-frkdg     98

EMBOSS_001        49 leelsgnarfrahagriirvfdesiqvlgqdgdlekldeiwtkiavship     98
                     ...||   ::..||..::::..:.:|.|  |.:| .:....:...:|:.
EMBOSS_001        99 kinls---klhghamammklmsklvqtl--dcnl-afrlaldenlpthlk    142

EMBOSS_001        99 rtvskesynqlkgvildvltaacsldesqaatwaklvdhvygiifk-aid    147
                     ..:..:....|...:..:.|:..::...:.:.:..:..:..|:.: |:.
EMBOSS_001       143 ngidpdymrmlatalksyilassvienhnscslsnglarlveivgeyavv    192

EMBOSS_001       148 ddgnak------------------------    153
                     |:...:
EMBOSS_001       193 dearkramstalrttvddagnrivkvalgt    222
```

BLAST für’s Laborleben

Die Sequenz ist im kodierenden Bereich fehlerhaft: wo vermuten Sie Fehler?

Human calmodulin mRNA, complete cds
Sequence ID: gb|M19311.1|HUMCAMLength: 1126Number of Matches: 1

Alignment statistics for match #1

Score	Expect	Identities	Gaps	Strand
819 bits(443)	0.0	448/450(99%)	2/450(0%)	Plus/Plus

```
Query  173  GCTGACCAACTGACTGAAGAGCAGATTGCAGAATTCAAAGAAGCTTTTTCATTATTTGAC  232
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  56   GCTGACCAACTGACTGAAGAGCAGATTGCAGAATTCAAAGAAGCTTTTTCATTATTTGAC  115

Query  233  AAAGATGGTGATGGCACTATAACAACAAAGGAACTTGGGACTGTAATGAGATCTCTTGGG  292
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  116  AAAGATGGTGATGGCACTATAACAACAAAGGAACTTGGGACTGTAATGAGATCTCTTGGG  175

Query  293  CAGAATCCCACAGAAGCAGAGTTACAGGACATGATTAATGAAGTAGATGCTGATGGTAAT  352
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  176  CAGAATCCCACAGAAGCAGAGTTACAGGACATGATTAATGAAGTAGATGCTGATGGTAAT  235

Query  353  GGCACAATTGACTTTCCTGAATTTCTGACAATGATGGCAAG-AAAATGAAAGACACAGAC  411 Del>framshift
            |||||||||||||||||||||||||||||||||||||||| ||||||||||||||||||
Sbjct  236  GGCACAATTGACTTTCCTGAATTTCTGACAATGATGGCAAGAAAAATGAAAGACACAGAC  295

Query  412  AGTGAAGAAGAAATTAGAGAAGCATTCCGTGTGTTTGACAAGGATGGCAATGGCTATATT  471
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  296  AGTGAAGAAGAAATTAGAGAAGCATTCCGTGTGTTTGACAAGGATGGCAATGGCTATATT  355

Query  472  AGTGCTGCAGAACTTCGCCATGTGATGACAAACCTTGGAGAGAAGTTAACAGATGAAGAA  531
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  356  AGTGCTGCAGAACTTCGCCATGTGATGACAAACCTTGGAGAGAAGTTAACAGATGAAGAA  415

Query  532  GTTGATGAAAATGATCAGGGAAGCAGATATTGATGGTGATGGTCAAGTAAACTATGAAGA  591 1ns>Frameshift weg
            ||||||| ||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  416  GTTGATG-AAATGATCAGGGAAGCAGATATTGATGGTGATGGTCAAGTAAACTATGAAGA  474

Query  592  GTTTGTACAAATGATGACAGCAAAGTGAAG  621
            |||||||||||||||||||||||||||||
Sbjct  475  GTTTGTACAAATGATGACAGCAAAGTGAAG  504
```

BLAST für's Laborleben

Die Sequenz ist im kodierenden Bereich fehlerhaft: wo vermuten Sie Fehler?

RecName: Full=Calmodulin; Short=CaM [Rattus norvegicus]
Sequence ID: sp|P62161.2|CALM_RATLength: 149Number of Matches: 1

Alignment statistics for match #1

Score	Expect	Method	Identities	Positives	Gaps	Frame
194 bits(493)	1e-60	Compositional matrix adjust.	102/148(69%)	113/148(76%)	0/148(0%)	+2

```
Query  173  ADQLTEEQIAEFKEAFSLFDKDGDGTITTKELGTVMRSLGQNPTEAELQDMINEVDADGN  352
            ADQLTEEQIAEFKEAFSLFDKDGDGTITTKELGTVMRSLGQNPTEAELQDMINEVDADGN
Sbjct  2    ADQLTEEQIAEFKEAFSLFDKDGDGTITTKELGTVMRSLGQNPTEAELQDMINEVDADGN  61

Query  353  GTIDFPEFLTMMARK*KTQTVKKKLEKHSVCLTRMAMAILVLQNFAM**QTLERS*QMKK  532 del>frameshift
            GTIDFPEFLTMMARK K    ++++ +      +     +           L      ++
Sbjct  62   GTIDFPEFLTMMARKMKDTDSEEEIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEE  121

Query  533  LMKMIREADIDGDGQVNYEEFVQMMTAK  616>ins>framshift weg
            + +MIREADIDGDGQVNYEEFVQMMTAK
Sbjct  122  VDEMIREADIDGDGQVNYEEFVQMMTAK  149
```