

**Einführung in die Angewandte Bioinformatik:  
Sequenzähnlichkeit, Sequenzalignment, BLAST  
10.06.2010**

Prof. Dr. Sven Rahmann

## Sequenzvergleich: Motivation

Hat man

- die DNA-Sequenz eines Gens,
  - die Aminosäuresequenz eines Proteins (Primärstruktur),
- weiß man noch nichts über seine Struktur oder Funktion.

Evtl. ist aber ein verwandtes Gen in einer Datenbank vorhanden,  
und über dieses Gen / Protein ist Wissen verfügbar.

Wie findet man das heraus?

Man vergleicht

- die neue nicht charakterisierte Sequenz
- mit allen bekannten Sequenzen.

## Sequenzvergleich: Fragen

- Wie kann man Sequenzähnlichkeit messen / quantifizieren?
- **Wieviel** Ähnlichkeit ist notwendig, um Informationen über ein Gen / Protein auf ein anderes zu übertragen?
- Sequenzähnlichkeit und ähnliche Struktur / Funktion sind zwar miteinander korreliert, aber nicht dasselbe. Inwieweit ist die Übertragung von Wissen durch Sequenzvergleich gerechtfertigt? Führt dieser Ansatz zu Irrtümern?
- Wie lässt sich die Ähnlichkeit zwischen einer Sequenz und einer (sehr großen) Datenbank von Sequenzen effizient (=schnell) berechnen?

## Modellierung von Sequenzähnlichkeit

Inwiefern und wo sind folgende Sequenzen ähnlich bzw. nicht ähnlich?  
(Hämoglobin / *Homo Sapiens*, Alpha- und Beta-Untereinheit)

```
>P69905|HBA_HUMAN Hemoglobin subunit alpha - Homo sapiens
MVLSPADKTNVKAAWGKVGAAHAGEYGAELERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
KKVADALTNAVAHVDDMPNALSALSSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTP
AVHASLDKFLASVSTVLTISKYR
```

```
>P68871|HBB_HUMAN Hemoglobin subunit beta - Homo sapiens
MVHLTPEEKSAVTALWGKVNVDVGGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK
VKAHGKKVLGAFSDGLAHLNLDLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG
KEFTPPVQAAYQKVVAGVANALAHKYH
```

Man „sieht“ erst einmal nichts!

Um Ähnlichkeiten zwischen Proteinen (Ketten von Aminosäuren)  
zu quantifizieren, machen wir zunächst Aussagen zu einzelnen Aminosäuren.

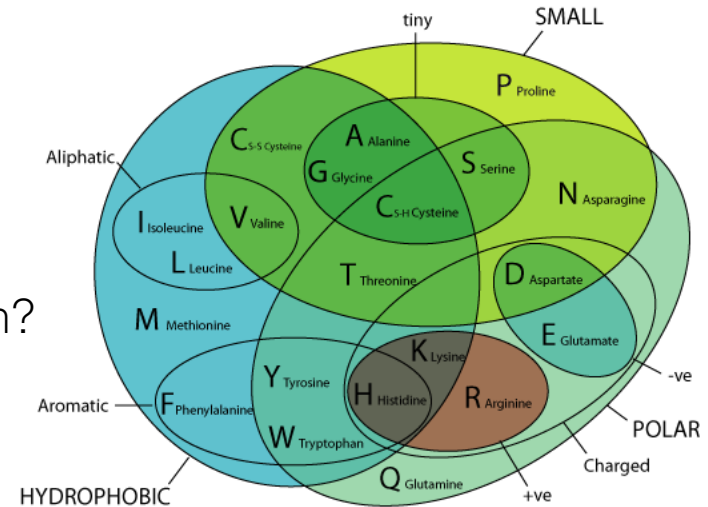
# Modellierung von Sequenzähnlichkeit: Ähnlichkeit von Aminosäuren

Ähnlichkeit =  
gleiche physikalische u. chemische Eigenschaften?

Experte: Messung und Vergleich von  
Eigenschaften (Größe, Ladung, Polarität,...)  
⇒ numerische Ähnlichkeitswerte („Scores“) ??

Wenn  $\text{Score}(I,L) = 2$  und  $\text{Score}(V,W) = -3$ ,  
was sollte dann  $\text{Score}(P,A)$  sein?

Schwierig, alle Score-Verhältnisse  
dieser Art konsistent zu schätzen.



<u>I</u>	<u>L</u>	<u>V</u>	<u>C</u>	<u>A</u>	<u>G</u>	<u>M</u>	<u>F</u>	<u>Y</u>	<u>W</u>	<u>H</u>	<u>K</u>	<u>R</u>	<u>E</u>	<u>Q</u>	<u>D</u>	<u>N</u>	<u>S</u>	<u>T</u>	<u>P</u>	
XXXXXXXXXXXX																			X	
																				X

## Modellierung von Sequenzähnlichkeit: Definition der Ähnlichkeit von Aminosäuren

### Beobachtung:

Während der Evolution werden durch Selektionsdruck ähnliche Aminosäuren häufiger durch einander ersetzt als unähnliche, da ähnliche eher die Struktur und Funktion des Proteins intakt lassen.

### „Geniale“ Idee:

**Definiere** die Häufigkeit, mit der zwei Aminosäuren durch einander ersetzt werden, als Ähnlichkeit.

### Probleme / zu beachten:

- (1) Häufigkeitszählung auf muss auf verwandte Proteine beschränkt werden.
- (2) Mutationshäufigkeit abhängig vom betrachteten Zeitraum.
- (3) natürliche Häufigkeit von Aminosäuren nicht gleich (5%), außerdem ggf. nach Spezies verschieden.

## Scorematrizen und Scorematrix-Familien

Für Aminosäuren  $x, y$  und evolutionäre Zeitspanne  $t$ ,  
definiere die Ähnlichkeit  $\text{Score}(x, y | t)$  als:

$$\text{Score}(x, y | t) := \log [ M(x, y | t) / (f(x) * f(y)) ] .$$

Hierbei ist:

- $M(x, y | t)$ : Relative Häufigkeit, mit der  $x$  und  $y$  ausgetauscht werden,  
wenn man Sequenzpaare mit evolutionärem Abstand  $t$  betrachtet.
- $f(x)$ : Relative Häufigkeit der Aminosäure  $x$
- $f(y)$ : Relative Häufigkeit der Aminosäure  $y$

Die Normalisierung mit  $f(x) * f(y)$  rechnet aus  $M(x, y | t)$   
die a-priori bekannten Häufigkeiten der Aminosäuren heraus.  
Das Häufigkeitsverhältnis wird durch Logarithmieren additiv.  
Name: „**log-odds score**“.

## Der Zeit-Parameter bei Scorematrix-Familien

Es macht Sinn, den evolutionären Abstand  $t$  als Parameter zu berücksichtigen.

Kleine Werte von  $t$  (kurze Zeitspanne):

- $\text{Score}(x,x | t)$  positiv und groß,
- $\text{Score}(x,y | t)$  stark negativ für  $x \neq y$ .

Unplausibel, dass Mutationen auftreten.

Große Werte von  $t$  (längere Zeitspannen):

- weniger gravierende Unterschiede.

Daher:

- nicht eine allgemeine Score-Matrix für Aminosäure-Ähnlichkeiten,
- sondern eine Score-Matrix für jeden Zeitparameter  $t \geq 0$ .

(„Familie von Score-Matrizen“.)



## Zeiteinheit bei Scorematrix-Familien: „Evolutionsmenge“ PAM

In welcher Einheit beschreibt man den Zeit-Parameter  $t$ ?

Man würde gerne Realzeit (z.B. Millionen Jahre) nehmen, aber man weiß nicht genau, wie schnell Proteinsequenzen mutieren, bzw. dies hängt von vielen Einflussfaktoren ab und ist variabel, und man kann sich damalige Proteinsequenzen nicht beschaffen.

Also legt man fest:

$t = 1$  entspricht der Zeit,

in der sich im Durchschnitt 1% der Aminosäuren einer Proteinsequenz ändern.

Name dieser Einheit: PAM (percent accepted mutations).

Sie beschreibt eine „Menge an Evolution“.

Umrechnung in Realzeit ist möglich, wenn die Evolutionsrate bekannt ist.

## 1 PAM und 100 PAM

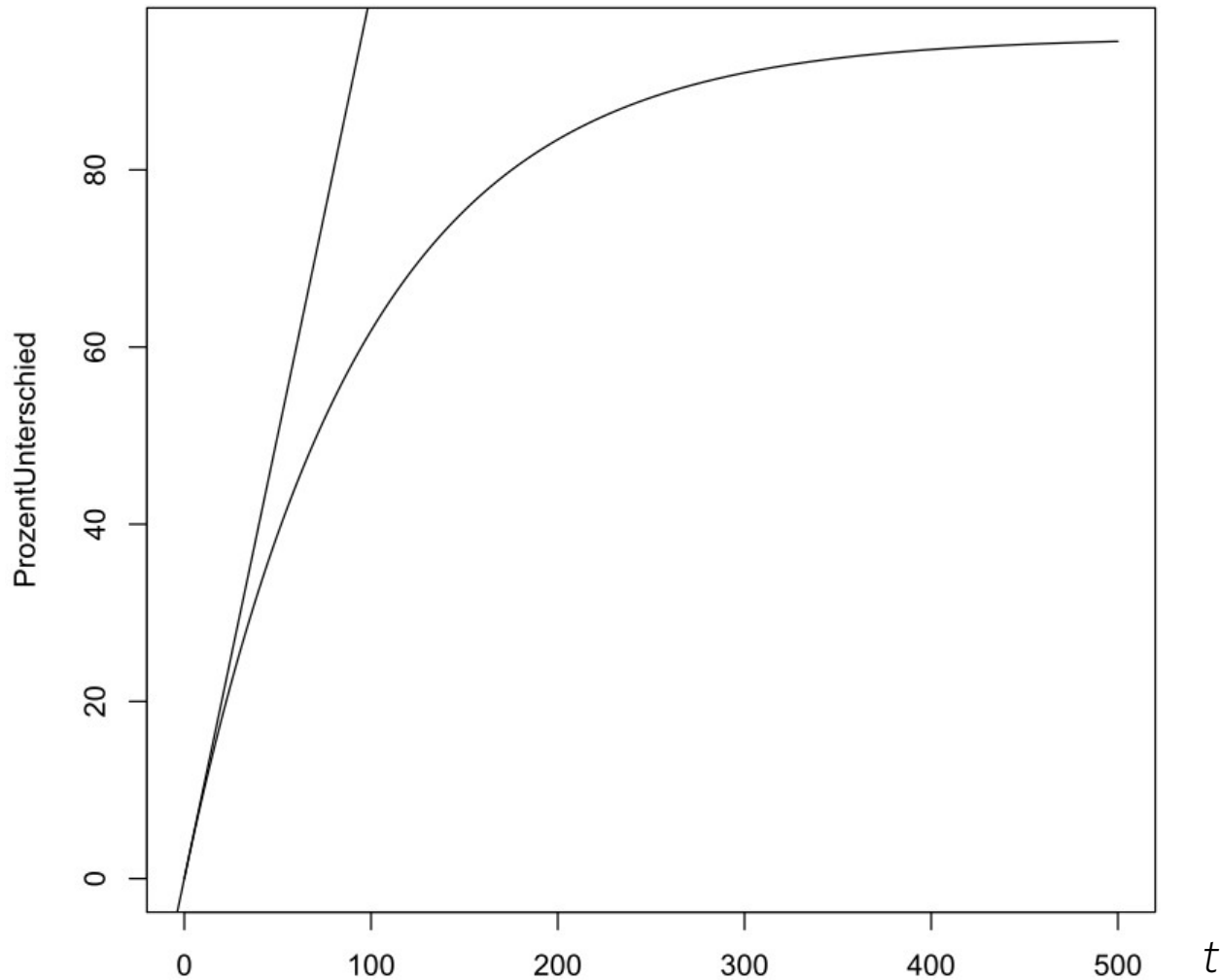
100 PAM entsprechen der 100-fachen Zeitdauer von 1 PAM.  
Heißt das, dass sich jede Aminosäure (100%) ändert?

Nein!

Einige Aminosäuren ändern sich mehrfach,  
aber man „sieht“ im vorher-nachher-Vergleich nur die letzte Änderung.  
Auch möglich: Rücksubstitution; man sieht keine Änderung.

Auch nach unendlich vielen PAM sehen ~5% der Positionen unverändert aus.  
Auch in Sequenzen, die nicht verwandt sind,  
stimmt im Schnitt jede 20. Aminosäure überein.

## Zusammenhang: Zeit $t$ [PAM] gegen Prozent Unterschied



## Die Score-Werte einer Score-Matrix-Familie

Ähnlichkeiten

$$\text{Score}(x,y | t) := \log [ M(x,y | t) / (f(x)*f(y)) ] .$$

hängen davon ab,

- welche Sequenzpaare man betrachtet, um Austauschhäufigkeiten zu zählen,
- mit welcher Methode man den Zeitparameter  $t$  den Sequenzpaaren zuordnet,
- wie man Informationen aus verschiedenen Zeitabständen  $t$  integriert.

Es gibt viele Möglichkeiten und zwei bekannte Familien:

- **PAM-Familie** von M. Dayhoff (1978)  
[nicht mit der PAM-Zeiteinheit verwechseln!]
- **BLOSUM-Familie** von Henikoff & Henikoff (1992)

## Die PAM-Familie

PAM- $t$ :

beruht auf 1572 Austausch in 71 Familien sehr nah verwandter Proteine.  
Für große  $t$  werden die Austauschhäufigkeiten  $M(x,y | t)$   
nicht beobachtet, sondern durch Extrapolation geschätzt.

PAM250-Matrix:

für evolutionär relativ weit entfernte, aber noch eindeutig verwandte Proteine.

Um nur ganze Zahlen zu verwenden,  
werden die log-odds scores  $\log [ M(x,y | t) / (f(x)*f(y)) ]$   
mit 10 multipliziert und gerundet.

## Beispiel: Die PAM-250-Matrix

Table 1 - The log odds matrix for 250 PAMs (multiplied by 10)

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

## Die BLOSUM-Familie

BLOSUM-s:

beruht auf zahlreichen Proteinen aus verschiedenen Familien mit verschiedenem evolutionären Abstand (größere Datenbasis als bei PAM-Matrizen).

Aber: Abstand wird nicht in Zeit  $t$  [PAM] gemessen, sondern „invers“ mit „Verwandtschaftsmaß“  $s$  (zwischen 0 und 100). Sequenzen mit Verwandtschaft größer als  $s$  werden zusammengefasst.

Kleines  $s$  bedeutet, BLOSUM-s enthält Werte für entfernt verwandte Sequenzen. Großes  $s$  bedeutet, BLOSUM-s enthält Werte für nah verwandte Sequenzen,

BLOSUM62 ist heute die Standard-Matrix für Proteinsequenz-Vergleiche.

## Beispiel: Die BLOSUM-62-Matrix

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	<b>9</b>	-1	-1	-3	0	-3	-3	-3	-4	-3	-3	-3	-3	-1	-1	-1	-1	-2	-2	-2
S	-1	<b>4</b>	1	-1	1	0	1	0	0	0	-1	-1	0	-1	-2	-2	-2	-2	-2	-3
T	-1	1	<b>4</b>	1	-1	1	0	1	0	0	0	-1	0	-1	-2	-2	-2	-2	-2	-3
P	-3	-1	1	<b>7</b>	-1	-2	-1	-1	-1	-1	-2	-2	-1	-2	-3	-3	-2	-4	-3	-4
A	0	1	-1	-1	<b>4</b>	0	-1	-2	-1	-1	-2	-1	-1	-1	-1	-1	-2	-2	-2	-3
G	-3	0	1	-2	0	<b>6</b>	-2	-1	-2	-2	-2	-2	-2	-3	-4	-4	0	-3	-3	-2
N	-3	1	0	-2	-2	0	<b>6</b>	1	0	0	-1	0	0	-2	-3	-3	-3	-3	-2	-4
D	-3	0	1	-1	-2	-1	1	<b>6</b>	2	0	-1	-2	-1	-3	-3	-4	-3	-3	-3	-4
E	-4	0	0	-1	-1	-2	0	2	<b>5</b>	2	0	0	1	-2	-3	-3	-3	-3	-2	-3
Q	-3	0	0	-1	-1	-2	0	0	2	<b>5</b>	0	1	1	0	-3	-2	-2	-3	-1	-2
H	-3	-1	0	-2	-2	-2	1	1	0	0	<b>8</b>	0	-1	-2	-3	-3	-2	-1	2	-2
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	<b>5</b>	2	-1	-3	-2	-3	-3	-2	-3
K	-3	0	0	-1	-1	-2	0	-1	1	1	-1	2	<b>5</b>	-1	-3	-2	-3	-3	-2	-3
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	<b>5</b>	1	2	-2	0	-1	-1
I	-1	-2	-2	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	<b>4</b>	2	1	0	-1	-3
L	-1	-2	-2	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	<b>4</b>	3	0	-1	-2
V	-1	-2	-2	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	<b>4</b>	-1	-1	-3
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	<b>6</b>	3	1
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	<b>7</b>	2
W	-2	-3	-3	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	<b>11</b>



## Von Aminosäure-Ähnlichkeiten zu Peptid-Ähnlichkeiten

Bisher: Ähnlichkeit von zwei Aminosäuren quantifizieren.

Jetzt: Ähnlichkeit zwischen zwei Peptiden (kurze Abfolgen von Aminosäuren) gleicher Länge?

Idee: Summe der Aminosäure-Ähnlichkeiten

Beispiel: MVLS, MVHL

Mit BLOSUM62 ist die Summe

$$\text{Score}(M,M) + \text{Score}(V,V) + \text{Score}(L,H) + \text{Score}(S,L) = 5 + 4 + (-3) + (-2) = 4,$$

Daraus ergibt sich eine Visualisierungsmöglichkeit: Dot-Plots.

## Dot Plots: 2D-Visualisierung von Protein-Ähnlichkeiten

Dot-Plot-Achsen:

x-Achse: Positionen der einen Sequenz,

y-Achse: Positionen der anderen Sequenz.

Wähle eine Peptidlänge (z.B. 10) als Fenstergröße,  
berechne für jede Kombination von Startpositionen in beiden Sequenzen  
den Peptid-Ähnlichkeitsscore.

Jedes (x,y)-Positionspaar,

für das der Peptid-Score einen Schwellenwert überschreitet,  
markiert man mit einem Punkt („dot“).

Alternativ kann der Peptid-Score auch in Graustufen dargestellt werden  
(z.B. Schwarz für sehr hoch, grau für hoch, weiß für niedrig).

**Beispiel:** Hämoglobin Alpha/Beta, Peptidlänge 10, Schwellenwert 23.

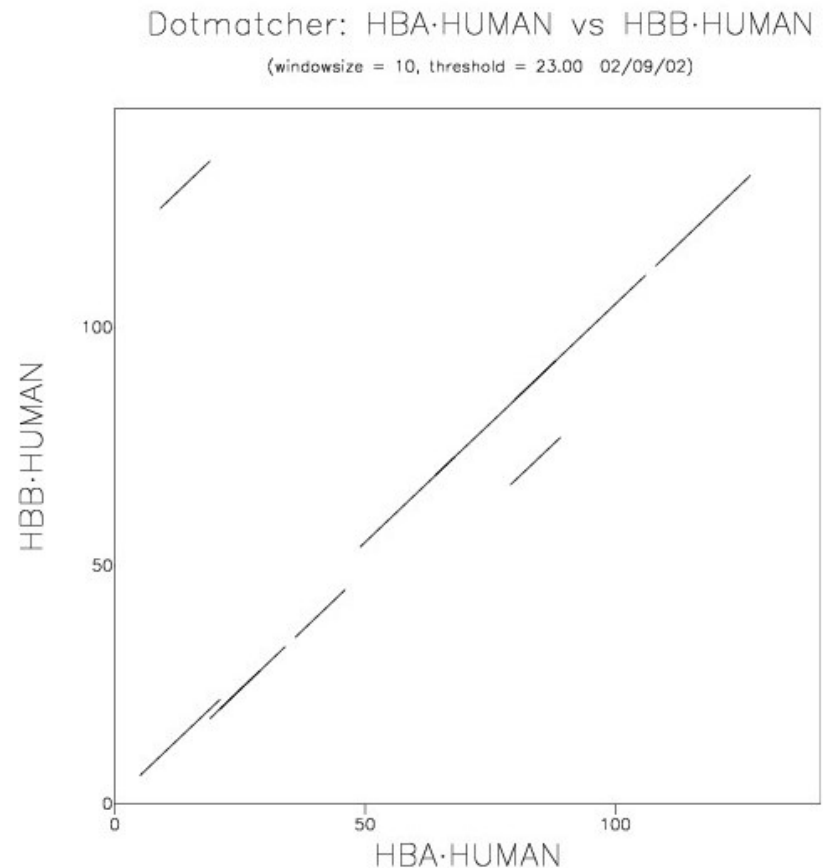
## Dot Plots: 2D-Visualisierung von Protein-Ähnlichkeiten

### Beispiel:

Hämoglobin Alpha/Beta,  
Peptidlänge 10, Schwellenwert 23.

### Zu erkennen:

- Gute globale Ähnlichkeit.
- In der Alpha-Kette fehlt bei Pos. ~50 ein Stück im Vergleich zur Beta-Kette.
- Wiederholte Peptide, Repeats ?



## Dot Plots mit Dotlet

The screenshot displays the Dotlet 1.5 web interface. At the top, it shows the user is 'anonymous' with a 'log in' button. Below this is a navigation bar with links for 'Documentation', 'about', 'need help?', 'learn by example', and 'new features in version 1.5'. A 'Reference' section cites Thomas Junier and Marco Pagni (2000) from *Bioinformatics*. The main control area includes buttons for 'print', 'input', and 'compute', along with dropdown menus for 'alpha', 'beta', 'Blosum62', '11', and '1:1'. A 'Blockieren' button is in the top right. The interface is split into two main panels. The left panel shows a dot plot with a diagonal line and a zoomed-in view of the upper-left corner. The right panel shows a histogram with a blue bar chart and a pink dashed curve, with a text box listing parameters: horizontal: alpha, vertical: beta, matrix: Blosum62, sliding window: 11, zoom: 1:1, score range: -44 to 121, and gray scale: 35% - 34%.

<http://myhits.isb-sib.ch/cgi-bin/dotlet>.

## Dot Plots mit Dotlet

Demonstration mit den Hämoglobin-Untereinheiten

```
>P69905|HBA_HUMAN Hemoglobin subunit alpha - Homo sapiens  
MVLSPADKTNVKAAWGKVGAHAGEYGAELERMFLSFPTTKTYFPHFDLSHGSAQVKGHG  
KKVADALTNVAHVDDMPNALSALSSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTP  
AVHASLDKFLASVSTVLTSKYR
```

```
>P68871|HBB_HUMAN Hemoglobin subunit beta - Homo sapiens  
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK  
VKAHGKKVLGAFSDGLAHLAHDNLKGTFFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG  
KEFTTPPVQAAVQKVVAGVANALAHKYH
```

URL: <http://myhits.isb-sib.ch/cgi-bin/dotlet>.

## Von Peptid-Ähnlichkeiten zu Sequenz-Ähnlichkeiten: Paarweise Sequenzalignments

Bisher: Ähnlichkeit von zwei Peptiden gleicher Länge quantifizieren.

Jetzt: Ähnlichkeit zwischen zwei Protein (verschiedener Länge)?

Berücksichtigen, dass es nicht nur Mutationen,  
sondern auch Einfügungen / Löschungen in einer Sequenz geben kann.

### Definition: (Paarweises) Protein-Sequenzalignment

- besteht aus zwei Zeilen und mehreren Spalten.
- In jeder Spalte stehen entweder
  - zwei Aminosäuren (identisch oder verschieden), oder
  - ein „Gap“ (Auslassung) und eine Aminosäure (Einfügung).

„Gap“: geschrieben als –, auch als indel (insertion/deletion) bezeichnet.

Liest man die erste Zeile ohne Gaps, ergibt sich die erste Proteinsequenz.

Liest man die zweite Zeile ohne Gaps, ergibt sich die zweite Proteinsequenz.

## Paarweise Sequenzalignments

Scores von Sequenzalignments: Summe der Spalten-Scores

**Aminosäure + Aminosäure:** nach Score-Matrix (z.B. BLOSUM62)

**Aminosäure + Gap:** noch festzulegen (negativer Score, z.B. -8)

**Beispiel** (BLOSUM62, Gap-Score = -8)

```
MV-LSPADKTNVKAAWGKVGA-HAGE
|| |:|.:.|.|.||||.. ..|
MVHLTPEEKSAVTALWGKVNVDVVG-
54 41 ...
-8
```

Die mittlere Zeile dient der Darstellung von Identität (|), hoher Ähnlichkeit (:), und Ähnlichkeit (.).

Erste 5 Spalten: Score von  $5+4+(-8)+4+1 = 6$ .

Zum Rechnen: Welcher Score ergibt sich insgesamt?

## Gapkosten

Einfügen von Gaps in Alignments erlaubt,

- Sequenzen unterschiedlicher Länge überhaupt zu alignieren;
- mehr hoch-scorende ähnliche Aminosäuren miteinander zu alignieren, indem man an geeigneten Stellen die Sequenzen gegeneinander verschiebt.

Gaps modellieren, dass während der Evolution in einer der Sequenzen Teile gelöscht wurden bzw. hinzugekommen sind.

Dies ist bei nah verwandten Sequenzen selten, so dass ein Gap mit einem stark negativen Score bewertet wird.

Es wird nicht unbedingt nur eine Aminosäure gelöscht, sondern evtl. mehrere. Daher: zusammenhängender Gap der Länge 3  $\neq$  drei Gaps der Länge 1.



## Lineare und affine Gapkosten

Seien  $g(k)$  die Kosten eines zusammenhängenden Gaps der Länge  $k$ .

Lineare Gapkosten:  $g(k) = -ck$  mit Konstante  $c > 0$ .  
 $c$ : Gapkosten  
 $\text{Score}(\text{Gap der Länge } k) = k * \text{Score}(\text{Gap der Länge } 1)$

Affine Gapkosten:  $g(k) = -o - e(k-1)$  mit Konstanten  $o > 0, e > 0$   
 $o$ : Gap-open-Kosten  
 $e$ : Gap-extend-Kosten

(Es sind noch kompliziertere Modelle denkbar.)

## Optimale paarweise Sequenzalignments

Viele Möglichkeiten, Alignments zu bilden (mit unterschiedlichen Scores).  
Welches ist „richtig“, um die Sequenzähnlichkeit auszuwerten?

MVLSPADKTNVKAAWGKVG AHAGE-  
MVHLTPEEKSAVTALWGKVN VDEVG

MV-LSP----ADKTNV--KAAWGKVG AHAGE  
MVHLTPEEKSAV-TALWGKVN VDEVG-----

?

MV-LSPADKTNVKAAWGKVG A-HAGE  
MVHLTPEEKSAVTALWGKVN VDEVG-

Idee: Alignment mit dem höchsten Score („**optimales Alignment**“).  
Mehr Ähnlichkeit geben die Sequenzen nicht her.  
Weniger nur, wenn man ihre Ähnlichkeiten nicht optimal herausarbeitet.

Der **Alignment-Score** zweier Sequenzen ist der **maximal mögliche** Score unter allen paarweisen Alignments dieser beiden Sequenzen.

## Globale und lokale Alignments

### Globales Alignment:

Alignment der gesamten Sequenzen.

### Lokales Alignment:

Alignment der am besten passenden Teilstücke der Sequenzen  
(höchster Score über alle Paare von Start- und Endpositionen).

## Wann global, wann lokal?

Global:

wenn beide Sequenzen einander über ihre gesamte Länge ähnlich sind  
(und damit auch in etwa gleich lang).

Lokal:

wenn die Sequenzen nicht über ihre ganze Länge homolog sind  
(homolog = evolutionär verwandt), sondern nur in Teilen.

## Berechnung optimaler Alignments

### Alignment-Problem

**Gegeben:** zwei Sequenzen, Score-Matrix, Gapkosten, Alignment-Typ (lokal/global).

**Gesucht:** Alignment-Score und zugehöriges optimales Alignment

### Wie findet man das optimale Alignment?

- Alle möglichen Alignments ausprobieren und Score berechnen
  - ineffizient: sehr große Anzahl möglicher Alignments zu bewerten
- Alignment-Algorithmen
  - Needleman-Wunsch [Needleman & Wunsch, J. Mol. Biol. 48; 443-453 (1970)]
  - Smith-Waterman [Smith & Waterman, J. Mol. Biol. 147(1); 195-197 (1981)]
  - NW („needle“) ist für globale Alignments, SW („water“) für lokale Alignments.

## Berechnung optimaler Alignments

European Bioinformatics Institute (EBI)  
bietet Zugriff auf EMBOSS  
(European Molecular Biology Open Software Suite)  
per Web-Formular.

Alignment-Web-Formular:  
<http://www.ebi.ac.uk/emboss/align/>.

EMBOSS enthält weitere Programme:  
<http://emboss.sourceforge.net/>.

### EMBOSS Pairwise Alignment Algorithms

This tool is used to compare 2 sequences. When you want an alignment that covers the whole length of both sequences, use [needle](#). When you are trying to find the best region of similarity between two sequences, use [water](#).

Method: EMBOSS::needle (global) | Gap Open: 10.0 | Gap Extend: 0.5 | Molecule: Protein | Matrix: Blosum62

Sequence 1: paste [Sequence](#) in any format OR upload a file:

Sequence 2: paste [Sequence](#) in any format OR upload a file:

# Berechnung optimaler Alignments

## Eingabeparameter

**Method:** „needle“ (global); „water“ (lokal)

**Gap open:** Gap-open-Kosten  $o$

**Gap extend:** Gap-extend-Kosten  $e$

**Molecule:** Protein oder DNA

**Matrix:** Scorematrix (z.B. BLOSUM62)

**Sequence 1+2:** Sequenzen

### EMBOSS Pairwise Alignment Algorithms

This tool is used to compare 2 sequences. When you want an alignment that covers the whole length of both sequences, use [needle](#). When you are trying to find the best region of similarity between two sequences, use [water](#).

Method: EMBOSS::needle (global) | Gap Open: 10.0 | Gap Extend: 0.5 | Molecule: Protein | Matrix: Blosum62

Sequence 1: paste [Sequence](#) in any format OR upload a file:

Seq. 1 Upload a file:  

Sequence 2: paste [Sequence](#) in any format OR upload a file:

Seq. 2 Upload a file:

## Datenbanksuche nach ähnlichen Sequenzen

### Häufige Situation:

Man hat ein „Stück“ neue/unbekannte Sequenz (DNA, RNA, oder Protein), möchte etwas darüber herausfinden.

### Idee:

Suche nach ähnlichen Sequenzen, über die etwas bekannt ist (z.B. in GenBank).  
Aber: Lästig, jede GenBank-Sequenz z.B. in das EMBOSS-Formular einzugeben.

### Datenbanksuchprogramme:

vergleichen (d.h. berechnen Alignment-Score)

- **eine** Sequenz (Anfrage, **Query**)
  - mit **jeder** Sequenz einer Sequenzdatenbank;
- geben Liste der hinreichend ähnlichen Datenbank-Sequenzen aus,  
bei Bedarf auch mit Alignments.

## Datenbanksuche nach ähnlichen Sequenzen: BLAST

### NCBI BLAST (Basic Local Alignment Search Tool):

Programm(sammlung) zur Datenbanksuche nach ähnlichen Sequenzen

URL: <http://blast.ncbi.nlm.nih.gov/>

#### Merkmale:

- findet (mehrere) lokale Alignments zwischen Query und DB-Sequenzen
- schnell (auch auf großen Datenbanken)
- berechnet (zunächst) keine exakten optimalen Alignments mit allen Sequenzen, sondern sucht nach Übereinstimmenden Regionen mit hohem Score  
erweitert diese im Erfolgsfall zu Alignments
- verschiedene Varianten (nach Molekültyp; für spezielle Anwendungen)



# BLAST: Programmauswahl [<http://blast.ncbi.nlm.nih.gov/>]

**BLAST** Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

► NCBI/BLAST Home

BLAST finds regions of similarity between biological sequences. [more...](#)

[Learn more](#) about how to use the new BLAST design

**BLAST Assembled Genomes**

Choose a species genome to search, or [list all genomic BLAST databases](#).

- [Human](#)
- [Mouse](#)
- [Rat](#)
- [Arabidopsis thaliana](#)
- [Oryza sativa](#)
- [Bos taurus](#)
- [Danio rerio](#)
- [Drosophila melanogaster](#)
- [Gallus gallus](#)
- [Pan troglodytes](#)
- [Microbes](#)
- [Apis mellifera](#)

**Basic BLAST**

Choose a BLAST program to run.

<a href="#">nucleotide blast</a>	Search a <b>nucleotide</b> database using a <b>nucleotide</b> query <i>Algorithms:</i> blastn, megablast, discontinuous megablast
<a href="#">protein blast</a>	Search <b>protein</b> database using a <b>protein</b> query <i>Algorithms:</i> blastp, psi-blast, phi-blast
<a href="#">blastx</a>	Search <b>protein</b> database using a <b>translated nucleotide</b> query
<a href="#">tblastn</a>	Search <b>translated nucleotide</b> database using a <b>protein</b> query
<a href="#">tblastx</a>	Search <b>translated nucleotide</b> database using a <b>translated nucleotide</b> query

Protein- oder  
DNA-Query  
gegen Genome

Auswahl nach Sequenztyp  
der Query und der Datenbank

## BLAST-Varianten

Programm	Query-Typ	DB-Typ	
blastn	Nukleotid	Nukleotid	Standard, auch entfernte Sequenzen sehr ähnliche Sequenzen, sehr schnell auch weiter entfernte Sequenzen
megablast	Nukleotid	Nukleotid	
disc. megabl.	Nukleotid	Nukleotid	
blastp	Protein	Protein	Standard langsamer; entfernte Verwandte Einschränkung auf Sequenz-Motive
psi-blast	Protein	Protein	
phi-blast	Protein	Protein	
blastx	Nukleotid (T)	Protein	(*) (*)
tblastn	Protein	Nukleotid (T)	(*) (T)
tblastx	Nukleotid (T)	Nukleotid (T)	(*)

(\*) (\*): Vergleich auf Protein-Ebene.  
(\*) (T): Nukleotid-Sequenz wird mit genetischem Code in Proteinsequenzen übersetzt.

### Basic BLAST

Choose a BLAST program to run.

#### nucleotide blast

Search a **nucleotide** database using a **nucleotide** query  
*Algorithms: blastn, megablast, discontinuous megablast*

#### protein blast

Search **protein** database using a **protein** query  
*Algorithms: blastp, psi-blast, phi-blast*

#### blastx

Search **protein** database using a **translated nucleotide** query

#### tblastn

Search **translated nucleotide** database using a **protein** query

#### tblastx

Search **translated nucleotide** database using a **translated nucleotide** query

## Protein BLAST (blastp)

BI/BLAST/blastp suite: BLASTP programs search protein databases using a protein query. [more...](#)

Enter Query Sequence

Enter accession number, gi, or FASTA sequence [?](#) [Clear](#)

Query subrange [?](#)

From

To

Or, upload file  [Durchsuchen...](#) [?](#)

Job Title

Enter a descriptive title for your BLAST search [?](#)

Choose Search Set

Database  [?](#)

Organism Optional

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown. [?](#)

Entrez Query Optional

Enter an Entrez query to limit search [?](#)

Program Selection

Algorithm

blastp (protein-protein BLAST)

PSI-BLAST (Position-Specific Iterated BLAST)

PHI-BLAST (Pattern Hit Initiated BLAST)

Choose a BLAST algorithm [?](#)

**BLAST** Search **database nr** using **Blastp (protein-protein BLAST)**

Show results in a new window

[Algorithm parameters](#)

Hier wird die Sequenz eingegeben oder hochgeladen, bzw. ein Teil davon ausgewählt (subrange).

Welche Protein-Datenbank soll durchsucht werden?  
Wie werden die Ergebnisse eingeschränkt?

Welche Variante des Algorithmus soll verwendet werden?

weitere Parameter (nächste Folie)

## Parameter von blastp

- Wie viele Treffer maximal anzeigen?
- Für kurze Queries anpassen?
- Erwartungswert-Schwellenwert: (\*)
- Wortlänge: Übereinstimmung dieser Länge mit Query ist notwendig (größer=schneller, aber findet nicht entferntere Sequenzen)
- Welche Scorematrix und Gapkosten sollen verwendet werden?
- Adjustierung: (\*)
- Query filtern oder maskieren, z.B. bekannte Repeats?

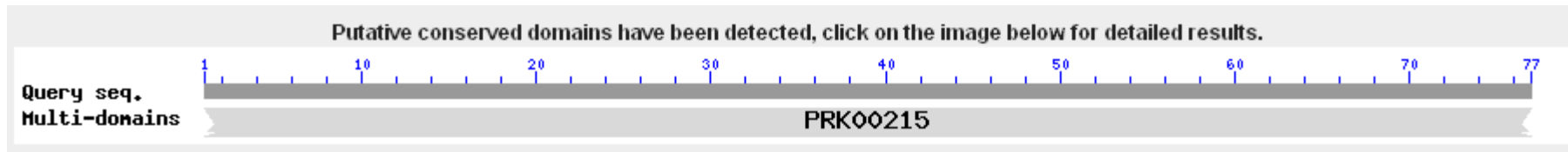
Zu (\*): erfordert Theorie zur BLAST-Statistik

The screenshot shows the NCBI BLAST web interface with the following sections:

- Algorithm parameters**
  - General Parameters**
    - Max target sequences: 100 (dropdown)
    - Short queries:  Automatically adjust parameters for short input sequences
    - Expect threshold: 10 (input field)
    - Word size: 3 (dropdown)
  - Scoring Parameters**
    - Matrix: BLOSUM62 (dropdown)
    - Gap Costs: Existence: 11 Extension: 1 (dropdown)
    - Compositional adjustments: Conditional compositional score matrix adjustment (dropdown)
  - Filters and Masking**
    - Filter:  Low complexity regions
    - Mask:  Mask for lookup table only,  Mask lower case letters
- BLAST**
  - Search database nr using Blastp (protein-protein BLAST)
  - Show results in a new window

## BLAST-Ausgabe

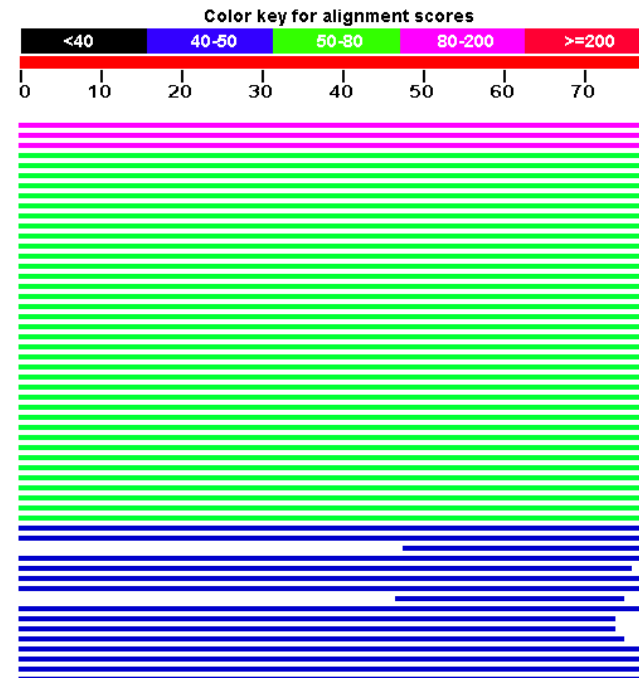
1. Während der Suche: Conserved Domains (Proteindomänen) in der Query



2. Ergebnisliste (graphisch):  
beste Treffer in der Datenbank.

Ausdehnung des Balkens zeigt,  
wo das ähnlichste Segment  
der DB-Sequenz in der Query liegt  
(lokales Alignment).

Farbe zeigt Grad der Ähnlichkeit an  
(hier gemessen in Prozent Identität).



## BLAST-Ausgabe

3. Ergebnisliste (Text):  
einzeilige Beschreibungen mit 5 Spalten

Sequences producing significant alignments:		Score (Bits)	E Value	
<a href="#">ref NP_939778.1 </a>	LexA repressor [Corynebacterium diphtheriae ...	<a href="#">152</a>	5e-36	<a href="#">G</a>
<a href="#">ref NP_738433.1 </a>	LexA repressor [Corynebacterium efficiens YS...	<a href="#">80.9</a>	3e-14	<a href="#">G</a>
<a href="#">sp Q8FPF5 LEXA_COREF</a>	LexA repressor	<a href="#">80.5</a>	3e-14	
<a href="#">ref YP_001138656.1 </a>	LexA repressor [Corynebacterium glutamicu...	<a href="#">77.4</a>	3e-13	<a href="#">G</a>
<a href="#">ref NP_601136.1 </a>	LexA repressor [Corynebacterium glutamicum A...	<a href="#">76.3</a>	6e-13	<a href="#">G</a>
<a href="#">sp Q8NP86 LEXA_CORGL</a>	LexA repressor >dbj BAB99323.1  SOS-resp...	<a href="#">76.3</a>	7e-13	
<a href="#">ref YP_250888.1 </a>	LexA repressor [Corynebacterium jeikeium K41...	<a href="#">63.5</a>	4e-09	<a href="#">G</a>
<a href="#">ref YP_120012.1 </a>	LexA repressor [Mycobacterium farcinica IFM 10152...	<a href="#">62.8</a>	7e-09	<a href="#">G</a>
<a href="#">ref YP_001800278.1 </a>	LexA repressor [Corynebacterium urealytic...	<a href="#">61.6</a>	2e-08	<a href="#">G</a>
<a href="#">ref YP_907016.1 </a>	LexA repressor [Mycobacterium ulcerans Agy99...	<a href="#">61.2</a>	2e-08	<a href="#">G</a>
<a href="#">ref YP_001850297.1 </a>	repressor LexA [Mycobacterium marinum M] ...	<a href="#">61.2</a>	2e-08	<a href="#">G</a>
<a href="#">ref YP_001703770.1 </a>	Probable repressor LexA [Mycobacterium ab...	<a href="#">58.2</a>	2e-07	<a href="#">G</a>
<a href="#">ref YP_706718.1 </a>	LexA repressor [Rhodococcus sp. RHA1] >gb AB...	<a href="#">58.2</a>	2e-07	<a href="#">G</a>
<a href="#">ref YP_001135215.1 </a>	LexA repressor [Mycobacterium gilvum PYR-...	<a href="#">57.4</a>	3e-07	<a href="#">G</a>
<a href="#">ref YP_639332.1 </a>	LexA repressor [Mycobacterium sp. MCS] >ref ...	<a href="#">57.0</a>	3e-07	<a href="#">G</a>
<a href="#">ref YP_001536300.1 </a>	transcriptional repressor, LexA family [S...	<a href="#">56.6</a>	5e-07	<a href="#">G</a>
<a href="#">ref NP_823639.1 </a>	LexA repressor [Streptomyces avermitilis MA-...	<a href="#">56.6</a>	5e-07	<a href="#">G</a>
<a href="#">emb CAA12169.1 </a>	LexA protein [Streptomyces clavuligerus]	<a href="#">56.6</a>	6e-07	<a href="#">G</a>

- (1) Link + Schlüssel der gefundenen DB-Sequenz
- (2) Name (+ Kurzbeschreibung) der gefundenen DB-Sequenz
- (3) Alignment-Score (in Bits [d.h. Logarithmus zur Basis 2 wurde verwendet])
- (4) der E-Wert (5e-36 bedeutet  $5 \cdot 10^{-36}$  und gibt die Anzahl der Treffer an, die man rein zufällig mit diesem oder besserem Score **erwarten** würde)
- (5) Links zu anderen NCBI-Datenbanken (z.B. G: Gene, S: Structures)

## BLAST-Ausgabe

4. Detaillierte Beschreibungen:  
insbesondere Alignments & deren Qualität:

- identities, positives, gaps
- Alignment-Score in bits (80.9)
- Alignment-Score in Rohform (198)
- E-Wert (Expect-Wert,  $3 \cdot 10^{-14}$ )
- Methode, mit der der E-Wert berechnet wurde

```
>[ref|NP_738433.1] LexA repressor [Corynebacterium efficiens YS-314]
[dbj|BAC18633.1] putative SOS response repressor LexA [Corynebacterium efficiens
YS-314]
Length=269
```

```
GENE ID: 1034469 CE1823 | LexA repressor [Corynebacterium efficiens YS-314]
(10 or fewer PubMed links)
```

```
Score = 80.9 bits (198), Expect = 3e-14, Method: Compositional matrix adjust.
Identities = 43/80 (53%), Positives = 55/80 (68%), Gaps = 6/80 (7%)
```

```
Query 1 RDPNKPRAVDVRLPDPPIPSKPGRKPGPKS---SWAISPDPAETSPTS FVPIVGSIAAG 57
RDPNKPRAVDVR LP+ + K GPK + SP P S TSF+P+VG IAAG
Sbjct 103 RDPNKPRAVDVRHLPE---TDNRTKAGPKAKARPTAGAS PQPELASSTSFIPVVGKIAAG 159

Query 58 NPILAEENVVDGYFFPFPSEIV 77
+PILAE+N++ Y+P P++IV
Sbjct 160 SPILAEQNIIEEYYPADIV 179
```

```
>[ref|NP_738433.1] [G] LexA repressor [Corynebacterium efficiens YS-314]
[dbj|BAC18633.1] [G] putative SOS response repressor LexA [Corynebacterium efficiens
YS-314]
Length=269

GENE ID: 1034469 CE1823 | LexA repressor [Corynebacterium efficiens YS-314]
(10 or fewer PubMed links)

Score = 80.9 bits (198), Expect = 3e-14, Method: Compositional matrix adjust.
Identities = 43/80 (53%), Positives = 55/80 (68%), Gaps = 6/80 (7%)

Query 1 RDPNKPRAVDVRALPDPIPSKPGRKPGPKKS---SVAISPDPAETSPTS FVPIVGSIAAG 57
RDPNKPRAVDVR LP+ + K GPK + SP P S TSF+P+VG IAAG
Sbjct 103 RDPNKPRAVDVRHLPE---TDNRTKAGPKAKARPTAGASPOPELASSTS FIPVVGKIAAG 159

Query 58 NPILAEENV DGYFFPFPSEIV 77
+PILAE+N++ Y+P P++IV
Sbjct 160 SPILAEQNIEEYYPLPADIV 179
```

## BLAST-Statistik

### Score in Rohform (198):

Summe der Scores über Alignment-Spalten

Interpretationsproblem der Roh-Scores:

- Längere Queries haben höhere Chance, gutes lokales(!) Alignment zu bekommen.
- Je länger die Sequenzen, desto länger „zufällige“ lokale Ähnlichkeiten.
- Unklar, auf welcher Skala der Roh-Score angegeben wird.

### Bit-Score (80.9):

längennormalisiert, genormt (Log-odds zur Basis 2)

Interpretationsproblem des Bit-Scores:

- Wann ist ein Bit-Score wirklich (sehr) hoch ?

### E-Wert ( $3 \cdot 10^{-14}$ ):

Lösung der Interpretationsprobleme



## Zusammenfassung

- Ähnlichkeit von Aminosäuren, Peptiden, (Protein-)Sequenzen
- Score-Matrizen und Gapkosten
- globale und lokale Sequenz-Alignments
- optimale Alignments, Alignment-Score
- Datenbanksuche nach ähnlichen Sequenzen mit BLAST
- Varianten von BLAST
- im Detail: Parameter von blastp (Protein-BLAST); andere ähnlich
  
- Nächste Woche: Erklärung der E-Werte