Einführung in die Angewandte Bioinformatik: Sequenzähnlichkeit, Alignment, BLAST
28.05.2009

Prof. Dr. Sven Rahmann

Sequenzvergleich: Motivation

Hat man bereits
- die DNA-Sequenz eines Gens, oder
- die Aminosäuresequenz eines Proteins (Primärstruktur)

bestimmt, 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 bereits 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 auch zu Irrtümern?

- Wie lässt sich die Ähnlichkeit zwischen einer Sequenz und einer (sehr großen) Datenbank von Sequenzen möglichst effizient (=schnell) berechnen?

Modellierung von Sequenzähnlichkeit

Inwiefern und wo sind folgende Sequenzen ähnlich bzw. nicht ähnlich?
(Menschliches Hämoglobin, Alpha- und Beta-Untereinheit)

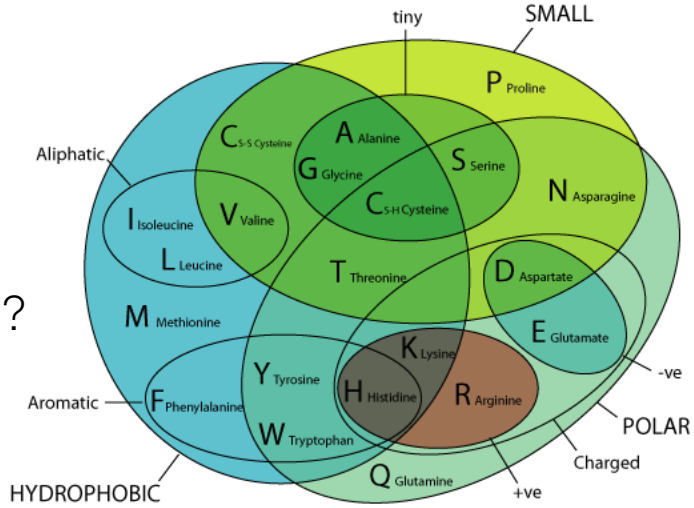
```
>P69905|HBA_HUMAN Hemoglobin subunit alpha - Homo sapiens
MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
KKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTP
AVHASLDKFLASVSTVLTSKYR

>P68871|HBB_HUMAN Hemoglobin subunit beta - Homo sapiens
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK
VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG
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 Score(I,L) = 2 und Score(V,W) = −3,
was sollte dann Score(P,A) sein?

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

```
ILVCAGMFYWHKREQDNSTP
XXXXXXXXXXX·····X·  Hydrophob.
·····XXXXXXXXXX·X  Polar
··XXXX·····XXXXX  Small
·····X  Proline
···XX·····X··  Tiny
XXX·····  Aliphatic
·····XXXX·····  Aromatic
·····XXX·····  Positive
·····X·X····  Negative
·····XXXX·X····  Charged
```

Abbildung: Livingstone, C. D. and Barton, G. J. (1993), "Protein Sequence Alignments: A Strategy for the Hierarchical Analysis of Residue Conservation", Comp. Appl. Bio. Sci., 9, 745-756.

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 Score(x,y | t) als:

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

Hierbei ist:

- $M(x,y \mid t)$: 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 $M(x,y \mid t)$
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):
- Score(x,x | t) positiv und groß,
- 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 $t >= 0$.

(„Familie von Score-Matrizen“.)

Zeiteinheit bei Scorematrix-Familien: „Evolutionsmenge“

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.

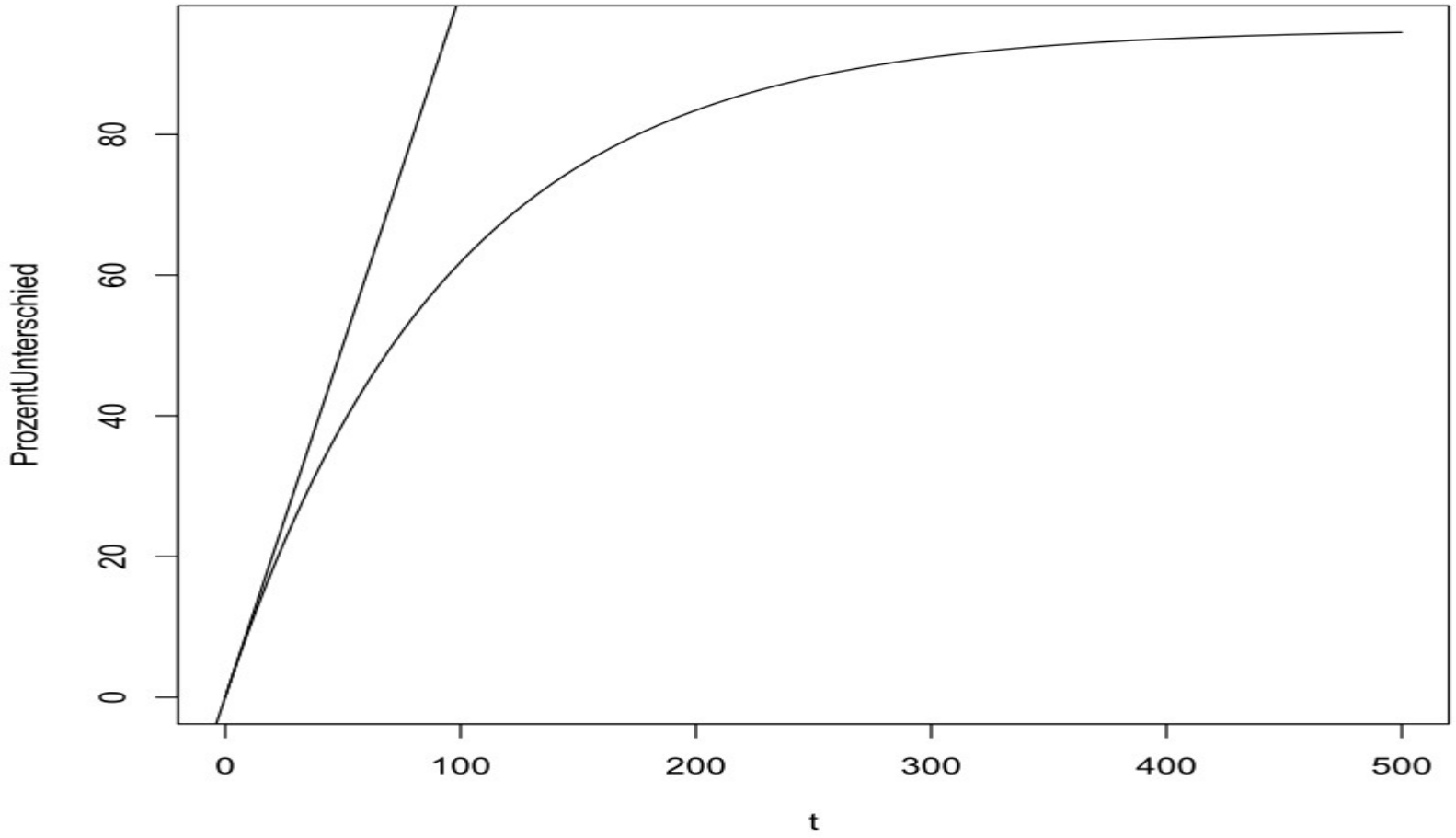
1 PAM und 100 PAM

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

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

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 Austauschen in 71 Familien sehr nah verwandter Proteine.
Für große *t* werden die Austauschhäufigkeiten $M(x,y \mid t)$
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 $\log [M(x,y \mid t) / (f(x)*f(y))]$
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	**9**	-1	-1	-3	0	-3	-3	-3	-4	-3	-3	-3	-3	-1	-1	-1	-1	-2	-2	-2
S	-1	**4**	1	-1	1	0	1	0	0	0	-1	-1	0	-1	-2	-2	-2	-2	-2	-3
T	-1	1	**4**	1	-1	1	0	1	0	0	0	-1	0	-1	-2	-2	-2	-2	-2	-3
P	-3	-1	1	7	-1	-2	-1	-1	-1	-1	-2	-2	-1	-2	-3	-3	-2	-4	-3	-4
A	0	1	-1	-1	**4**	0	-1	-2	-1	-1	-2	-1	-1	-1	-1	-1	-2	-2	-2	-3
G	-3	0	1	-2	0	**6**	-2	-1	-2	-2	-2	-2	-2	-3	-4	-4	0	-3	-3	-2
N	-3	1	0	-2	-2	0	**6**	1	0	0	-1	0	0	-2	-3	-3	-3	-3	-2	-4
D	-3	0	1	-1	-2	-1	1	**6**	2	0	-1	-2	-1	-3	-3	-4	-3	-3	-3	-4
E	-4	0	0	-1	-1	-2	0	2	**5**	2	0	0	1	-2	-3	-3	-3	-3	-2	-3
Q	-3	0	0	-1	-1	-2	0	0	2	**5**	0	1	1	0	-3	-2	-2	-3	-1	-2
H	-3	-1	0	-2	-2	-2	1	1	0	0	**8**	0	-1	-2	-3	-3	-2	-1	2	-2
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	**5**	2	-1	-3	-2	-3	-3	-2	-3
K	-3	0	0	-1	-1	-2	0	-1	1	1	-1	2	**5**	-1	-3	-2	-3	-3	-2	-3
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	**5**	1	2	-2	0	-1	-1
I	-1	-2	-2	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	**4**	2	1	0	-1	-3
L	-1	-2	-2	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	**4**	3	0	-1	-2
V	-1	-2	-2	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	**4**	-1	-1	-3
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	**6**	3	1
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7	2
W	-2	-3	-3	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	**11**

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
Score(M,M) + Score(V,V) + Score(L,H) + 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.

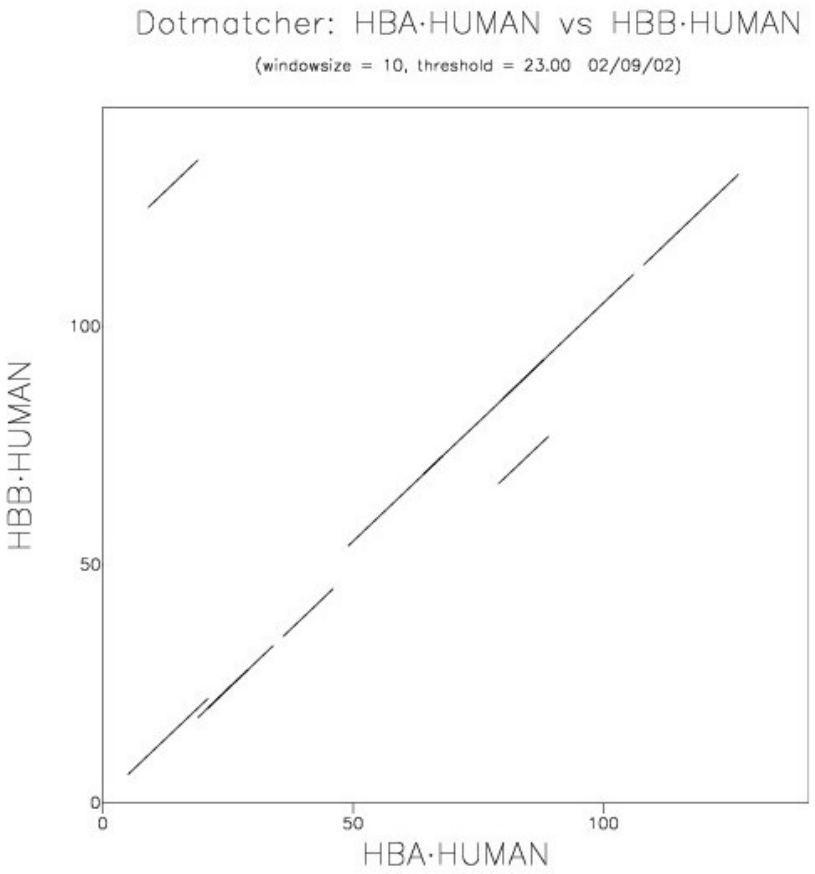
Jedes (x,y)-Positionspaar,
für das der Peptid-Score einen Schwellenwert überschreitet,
markiert man mit einem Punkt („dot“).
Alternativ kann der Peptid-Score auch in Graustufen dargestellt werden
(z.B. Schwarz für sehr hoch, grau für hoch, weiß für niedrig).

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

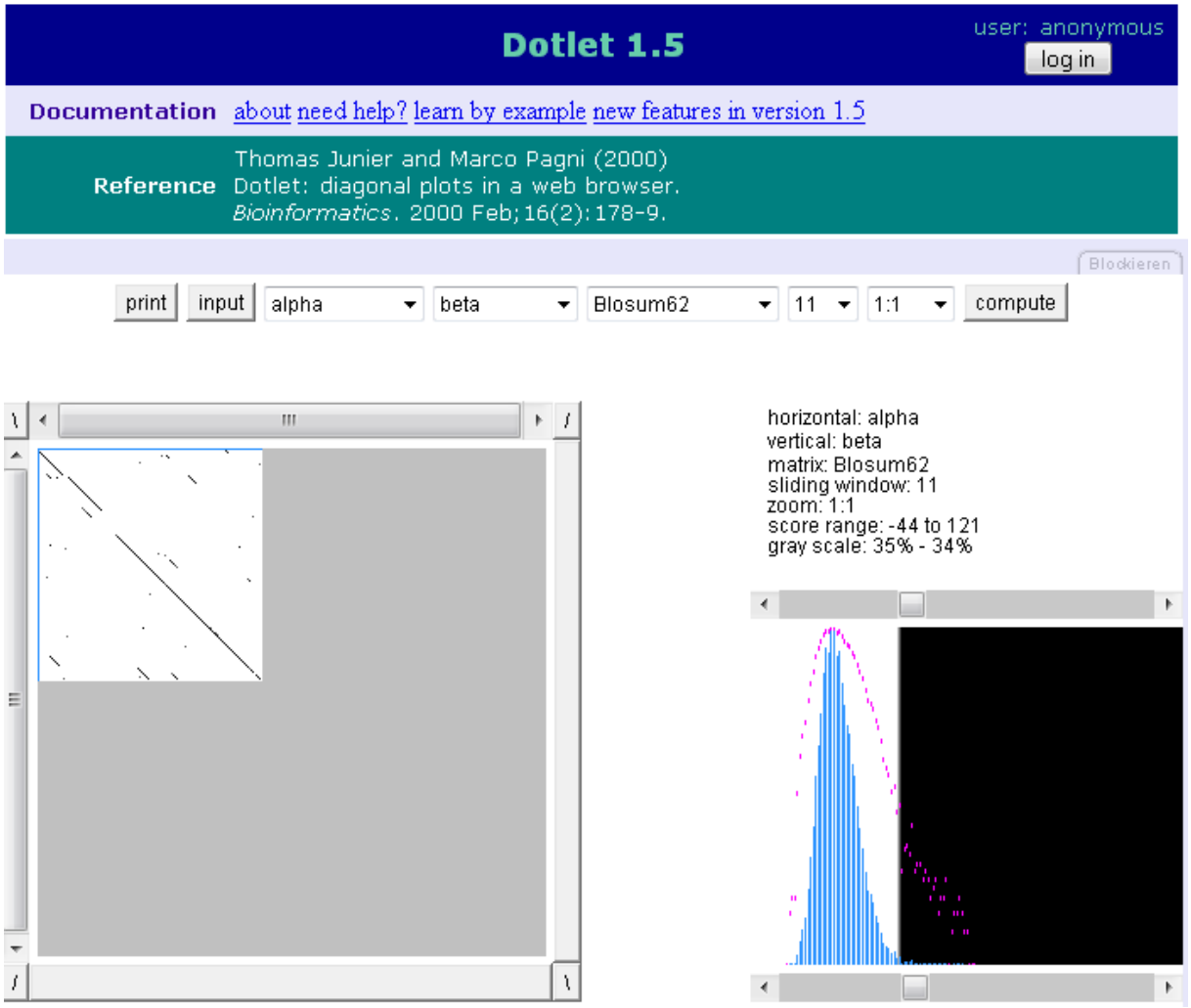
Dot Plots: 2D-Visualisierung von Protein-Ähnlichkeiten

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



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
MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
KKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTP
AVHASLDKFLASVSTVLTSKYR

>P68871|HBB_HUMAN Hemoglobin subunit beta - Homo sapiens
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK
VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG
KEFTPPVQAAYQKVVAGVANALAHKYH
```

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
||  |:|.:|:.|.|.||||.. ..|
MVHLTPEEKSAVTALWGKVNVDEVG-
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 ≠ 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
Score(Gap der Länge k) = k*Score(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?

```
MVLSPADKTNVKAAWGKVGAHAGE-      MV-LSP----ADKTNV--KAAWGKVGAHAGE
MVHLTPEEKSAVTALWGKVNVDEVG      MVHLTPEEKSAV-TALWGKVNVDEVG-----
```

?

```
MV-LSPADKTNVKAAWGKVGA-HAGE
MVHLTPEEKSAVTALWGKVNVDEVG-
```

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?

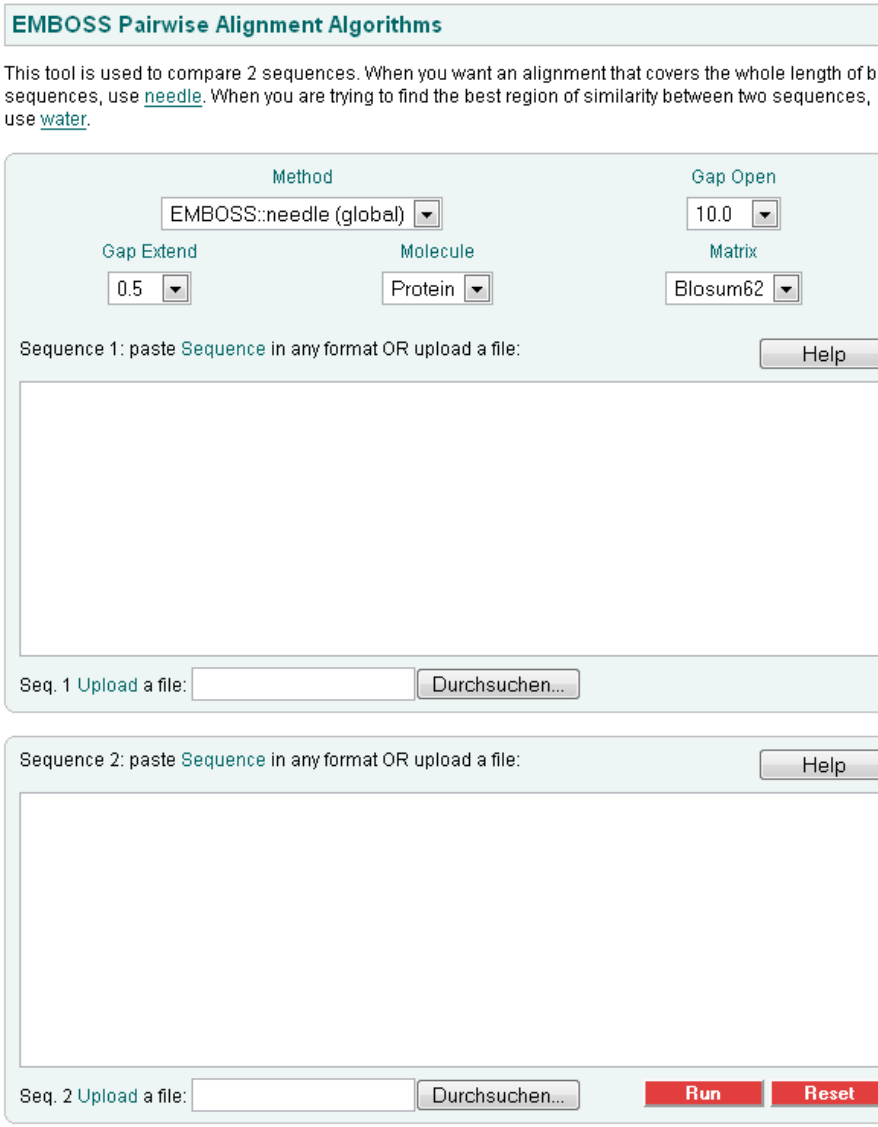
- Alle möglichen Alignments ausprobieren und Score berechnen
  - ineffizient: sehr große Anzahl möglicher Alignments zu bewerten
- Alignment-Algorithmen
  - Needleman-Wunsch [Needleman & Wunsch, J. Mol. Biol. 48; 443-453 (1970)]
  - Smith-Waterman [Smith & Waterman, J. Mol. Biol 147(1); 195-197 (1981)]
  - 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/.



Berechnung optimaler Alignments

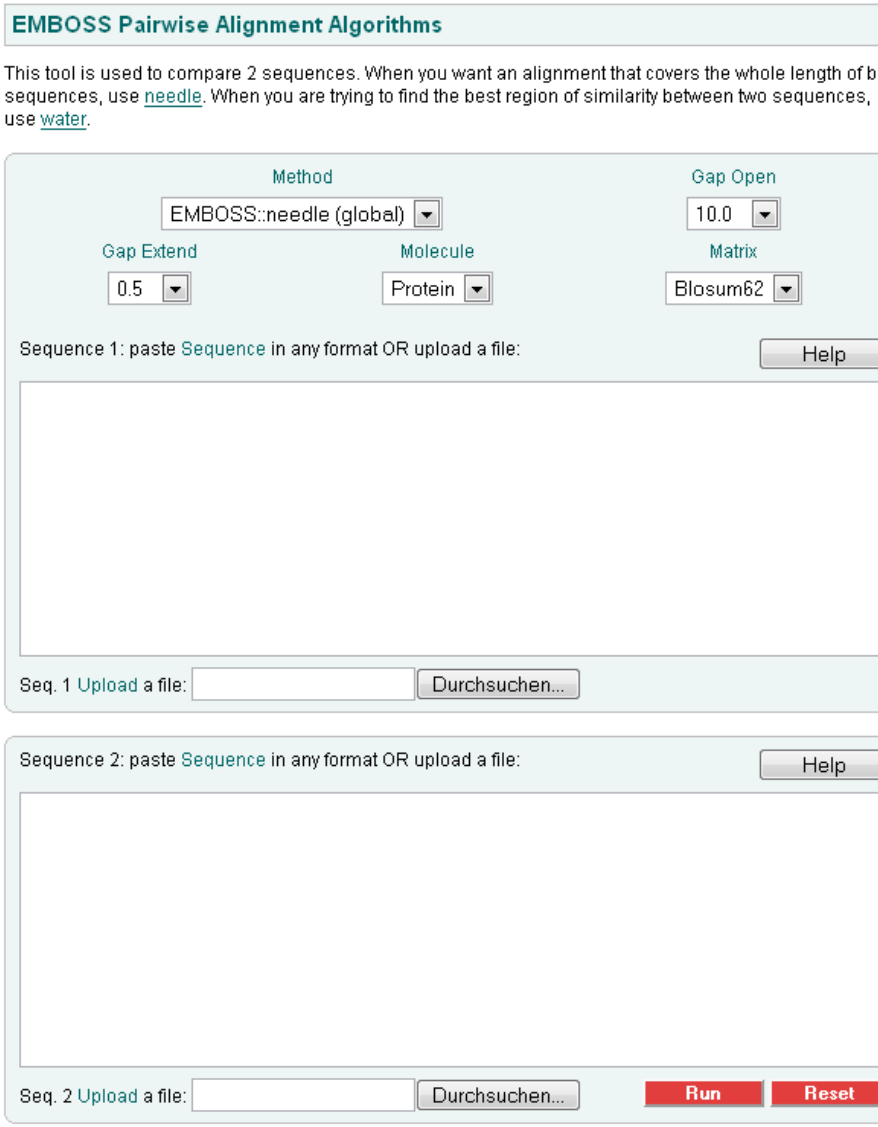
Eingabeparameter

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



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;

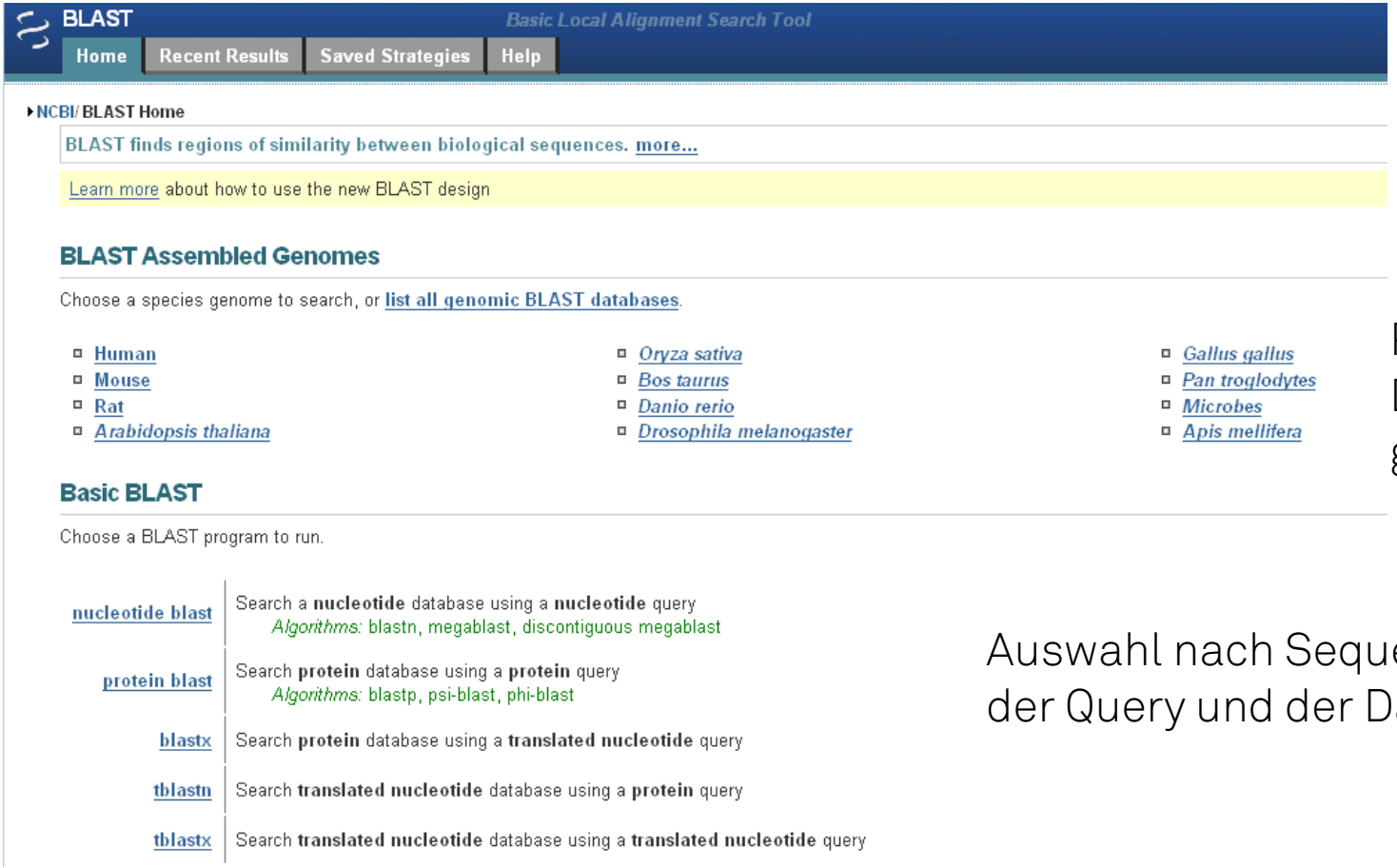
geben Liste der hinreichend ähnlichen Datenbank-Sequenzen aus,
bei Bedarf auch mit Alignments.

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)

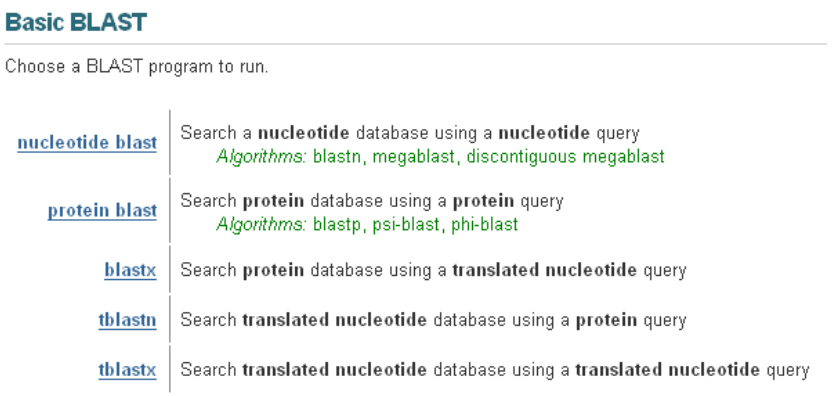
BLAST: Programmauswahl [http://blast.ncbi.nlm.nih.gov/]



Protein- oder DNA-Query gegen Genome

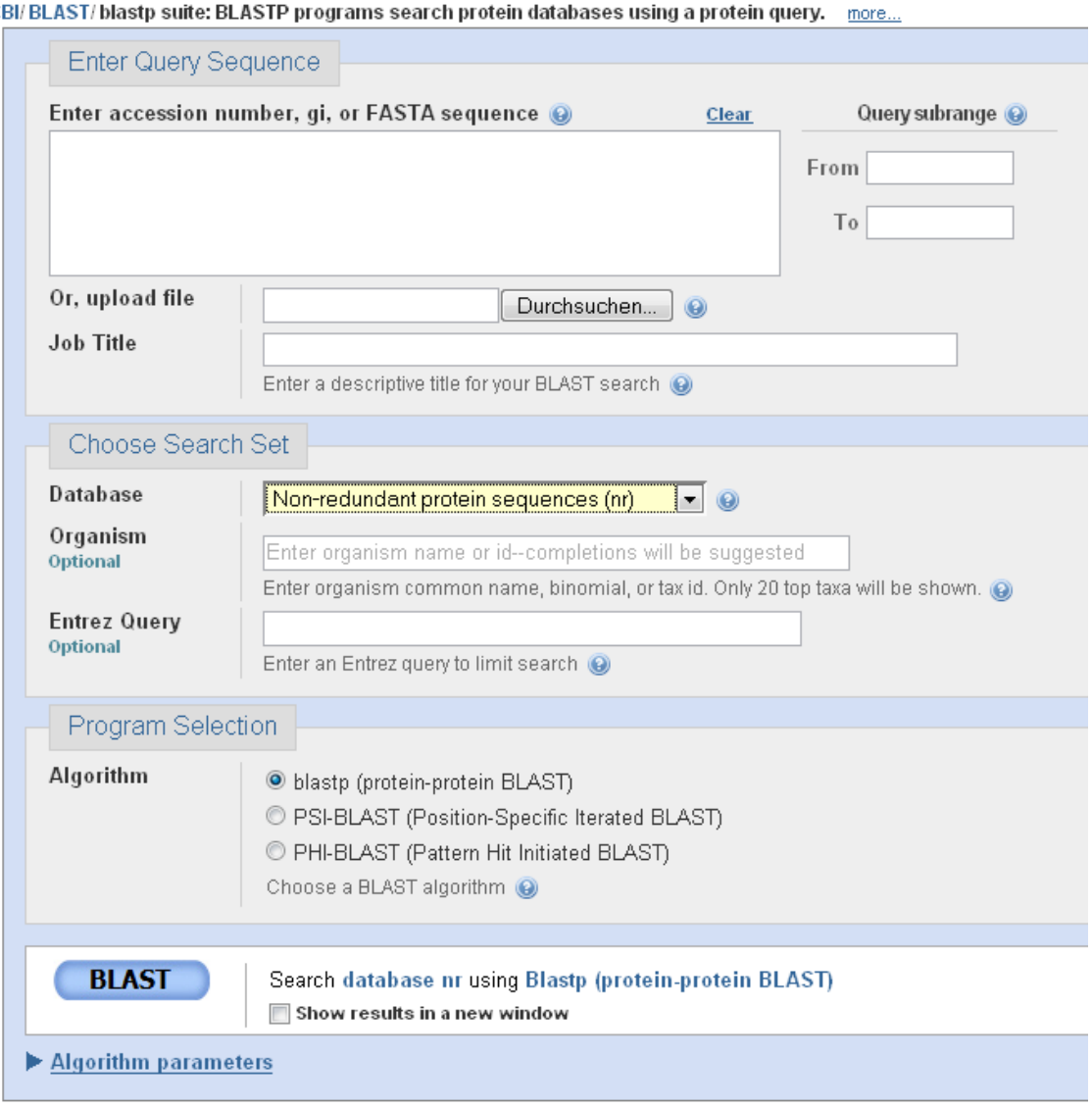
Auswahl nach Sequenztyp der Query und der Datenbank

BLAST-Varianten



Programm	Query-Typ	DB-Typ		
blastn	Nukleotid	Nukleotid		Standard, auch entfernte Sequenzen
megablast	Nukleotid	Nukleotid		sehr ähnliche Sequenzen, sehr schnell
disc. megabl.	Nukleotid	Nukleotid		auch weiter entfernte Sequenzen
blastp	Protein	Protein		Standard
psi-blast	Protein	Protein		langsamer; entfernte Verwandte
phi-blast	Protein	Protein		Einschränkung auf Sequenz-Motive
blastx	Nukelotid (T)	Protein	(*)	(*): Vergleich auf Protein-Ebene.
tbalstn	Protein	Nukleotid (T)	(*)	(T): Nukleotid-Sequenz wird
tblastx	Nukleotid (T)	Nukleotid (T)	(*)	mit genetischem Code in Proteinsequenzen übersetzt.

Protein BLAST (blastp)



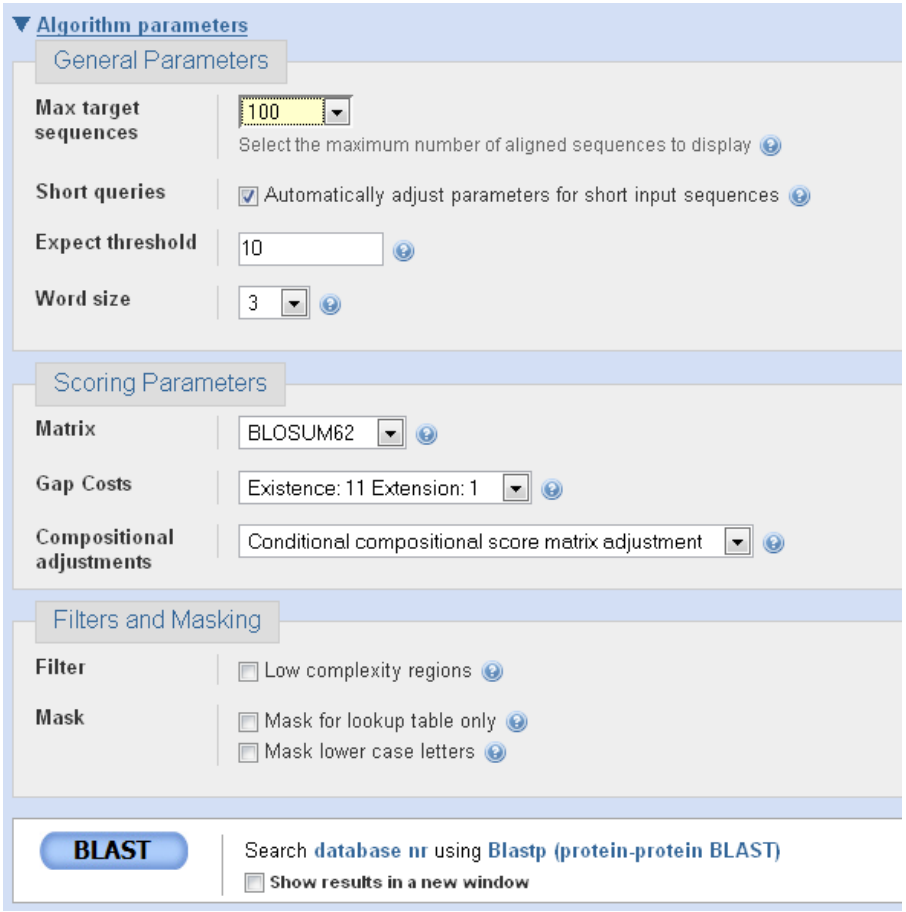
Hier wird die Sequenz eingegeben oder hochgeladen, bzw. ein Teil davon ausgewählt (subrange).

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

Welche Variante des Algorithmus soll verwendet werden?

weitere Parameter (nächste Folie)

Paramter 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)

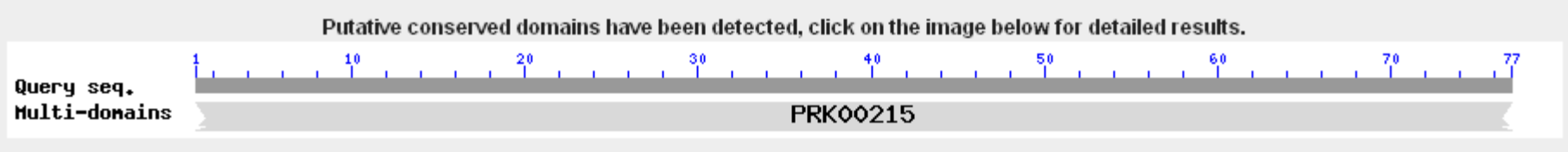
- Welche Scorematrix und Gapkosten sollen verwendet werden?
- Adjustierung: (*)

- Query filtern oder maskieren, z.B. bekannte Repeats?

Zu (*): erfordert Theorie zur BLAST-Statistik

BLAST-Ausgabe

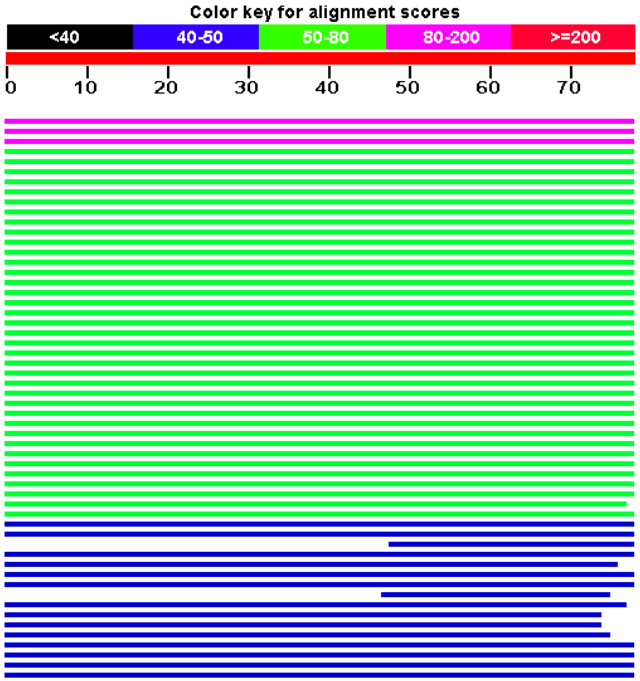
1. Während der Suche: Conserved Domains (Proteindomänen) in der Query



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

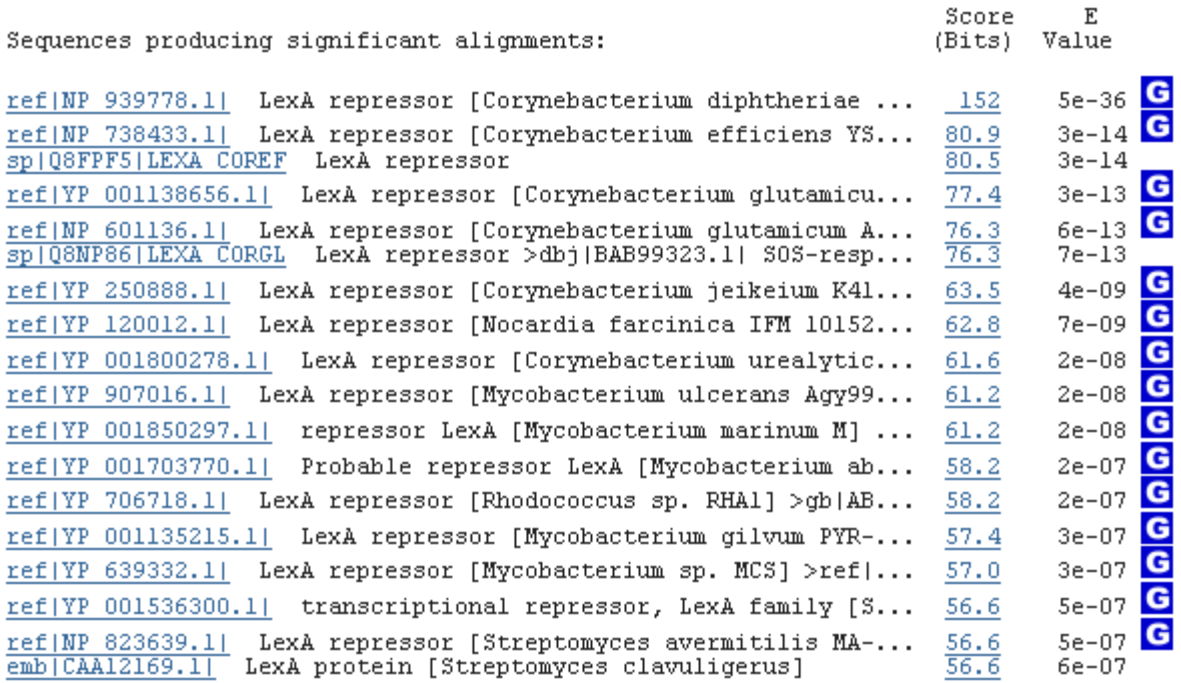
Ausdehnung des Balkens zeigt,
wo das ähnlichste Segment
der DB-Sequenz in der Query liegt
(lokales Alignment).

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

(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 $5*10^{-36}$ 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, $3*10^{-14}$)
- 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    RDPNKPRAVDVRALPDPIPSKPGRKPGPKKS---SVAISPDPAETSPTSFVPIVGSIAAG  57
            RDPNKPRAVDVR LP+   +    K GPK     +   SP P   S TSF+P+VG IAAG
Sbjct  103  RDPNKPRAVDVRHLPE---TDNRTKAGPKAKARPTAGASPQPELASSTSFIPVVGKIAAG  159

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

BLAST-Statistik

```
>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    RDPNKPRAVDVRALPDPIPSKPGRKPGPKKS---SVAISPDPAETSPTSFVPIVGSIAAG  57
            RDPNKPRAVDVR LP+   +    K GPK     +   SP P   S TSF+P+VG IAAG
Sbjct  103  RDPNKPRAVDVRHLPE---TDNRTKAGPKAKARPTAGASPQPELASSTSFIPVVGKIAAG  159

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

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 ist?

E-Wert ($3*10^{-14}$):

Lösung der Interpretationsprobleme

Blast-Statistik: E-Wert

Definition (E-Wert):
Wie viele Treffer **mit diesem oder höherem** Score
erwartet man aufgrund zufälliger Sequenzähnlichkeiten?
(mit zufälliger Query gleicher Länge in zufälliger Datenbank gleicher Größe)

Kleiner E-Wert (z.B. $< 10^{-10}$) heißt:
- unwahrscheinlich, dass beobachtete Ähnlichkeit
  auf einem nur zufällig ähnlichen Stück Sequenz beruht.
- plausibel, dass eine evolutionäre Verwandtschaft vorliegt.

Hoher E-Wert (nahe bei oder größer als 1) heißt:
- Such-Treffer lässt sich durch Zufall erklären,
  sollte nicht notwendigerweise als biologisch relevant gelten.

Berechnung des E-Werts

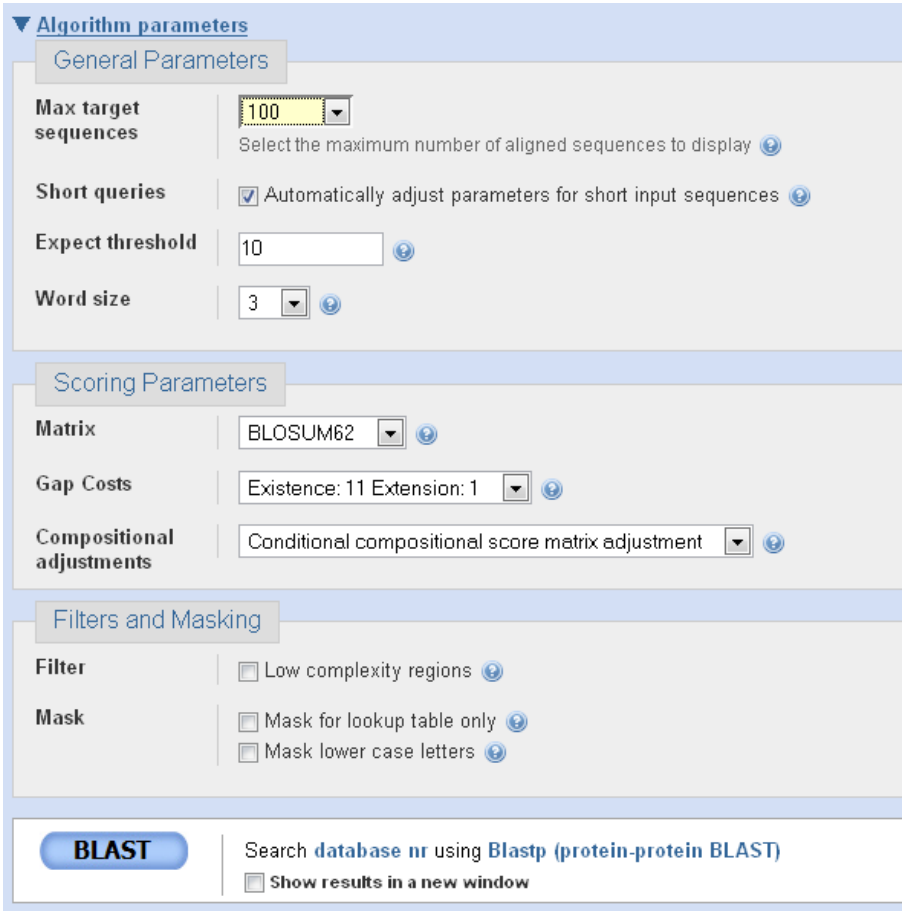
E-Wert (erwartete Anzahl Treffer mit diesem oder besserem Score) hängt ab von:
- Score s
- Querylänge m
- Datenbanklänge n (Gesamtlänge)
- Scorematrix und Gapkosten

E-Wert lässt sich mathematisch berechnen und approximieren durch

$$E(s) \approx Kmn\, e^{-\lambda s}.$$

Dabei hängen $K>0$, $\lambda>0$ von der Scorematrix und den Gapkosten ab.

Verbleibende Paramter von blastp



- **Erwartungswert-Schwellenwert:**
  Ausgabe nur von Treffern mit Score s,
  so dass $E(s) <$ Schwellenwert.
  Kleinerer Schwellenwert:
  schneller, nur nah verwandte Treffer,

- **Adjustierung:**
  E-Wert Berechnung geht von Zufälligkeit von Query und Datenbank aus.
  Query ist aber bei der Suche fest.
  Wenn die Aminosäure-Zusammensetzung der Query von der „typischen“ abweicht, sind die E-Werte unpassend.
  Daher: Adjustiere Parameter K, λ
  im Hinblick auf Query-Zusammensetzung.

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
- Scores und E-Werte

Anhang: E und e

- E-Wert: E heißt Erwartung
- 5E-02 = 5e-02 = 0.05: E=e heißt Exponent
  (3.2e+18 = $3.2 \cdot 10^{18}$, **nicht** $3.2 \cdot e^{18}$)
- e in $e^{-\lambda s}$: e heißt Eulersche Zahl (ca. 2.71)