

# Sequenzvergleich und Datenbanksuche

---

## Vorlesung

Einführung in die Angewandte Bioinformatik

Prof. Dr. Sven Rahmann

## Webseite zur Vorlesung

<http://bioinfo.wikidot.com/>

## Sprechstunde

Mo 16-17 in OH14, R214

Sven.Rahmann -at- tu-dortmund.de

# Sequenzvergleiche

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

Möglicherweise ist aber bereits aus einem anderen Genomprojekt ein nahe verwandtes Gen in einer Datenbank abgelegt, und über dieses Gen / Protein ist bereits Wissen vorhanden.

Wie findet man das heraus?

Man vergleicht die (neue, uncharakterisierte) Sequenz mit allen bekannten Sequenzen!

Wenn man eine ähnliche Sequenz findet (und dazu Informationen vorhanden sind), ist man ein gutes Stück weiter!

# Sequenzvergleiche

Es ergeben sich eine Reihe von Fragen:

- Wie kann man Sequenzähnlichkeit messen / quantifizieren?  
[Modellierung, Informatik, Mathematik]
- „Wieviel“ Ähnlichkeit ist notwendig, um Informationen über ein Gen / Protein auf ein anderes zu übertragen?  
[Praxis, Erfahrung, Modellierung, Statistik]
- Sequenzähnlichkeit und ähnliche (Protein)struktur und (Protein)funktion sind zwar miteinander korreliert, aber jeweils nicht dasselbe!  
Inwieweit ist die Übertragung von Wissen durch Sequenzvergleich gerechtfertigt?  
Führt dieser Ansatz auch zu Irrtümern?  
[Praxis, Erfahrung, Modellierung, Statistik]
- Wie lässt sich die Ähnlichkeit zwischen einer Sequenz und einer (sehr großen) Datenbank von Sequenzen möglichst effizient (=schnell) berechnen?  
[Informatik, Algorithmik]

# Problematik: Modellierung von Ähnlichkeit

Wie / wo sind folgende Sequenzen ähnlich / nicht ähnlich?

Es handelt sich um menschliches Hämoglobin, Alpha- und Beta-Untereinheit, hier im FASTA-Format angegeben

```
>P69905|HBA_HUMAN Hemoglobin subunit alpha - Homo sapiens
MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMF1LSFPTTKTYFP2PHFDLSHGSAQVKGHG
KKVADALTNVAHVDDMPNALSALS3DLHAH4KLRVDPVNFKLLSHCLLVTLAAHLPAEFTP
AVHASLDKFLASVSTVLT5SKYR
```

```
>P68871|HBB_HUMAN Hemoglobin subunit beta - Homo sapiens
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK
VKAHGKKV1LGAFSDGLAHL2DNLKGT3FATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG
KEFTPPVQAAYQKVVAGVANALAHKYH
```

Man „sieht“ erst einmal nichts!

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



# Score-Matrizen

Während der Evolution werden durch Selektionsdruck ähnliche Aminosäuren häufiger durcheinander ersetzt als unähnliche, da ähnliche mit höherer Wahrscheinlichkeit die Struktur und damit die Funktion des Proteins intakt lassen.

Also kann man (in als evolutionär verwandt bekannten Proteinen!) die Häufigkeit, mit der zwei Aminosäuren (durch DNA-Mutation) durcheinander ersetzt werden, als Ähnlichkeitswert nehmen.

## Problem dabei:

Die natürliche Häufigkeit von Aminosäuren ist bereits sehr verschieden: siehe z.B.

<http://www.tiem.utk.edu/bioed/webmodules/aminoacid.htm>

## Weiteres Problem (?):

Die Mutationshäufigkeit ist abhängig vom betrachteten evolutionären Zeitraum.

Sollte ein Ähnlichkeitswert von Aminosäuren auch davon abhängen?

Amino Acids	Codons	Observed Frequency in Vertebrates
Alanine	GCU, GCA, GCC, GCG	7.4 %
Arginine	CGU, CGA, CGC, CGG, AGA, AGG	4.2 %
Asparagine	AAU, AAC	4.4 %
Aspartic Acid	GAU, GAC	5.9 %
Cysteine	UGU, UGC	3.3 %
Glutamic Acid	GAA, GAG	5.8 %
Glutamine	CAA, CAG	3.7 %
Glycine	GGU, GGA, GGC, GGG	7.4 %
Histidine	CAU, CAC	2.9 %
Isoleucine	AUU, AUA, AUC	3.8 %
Leucine	CUU, CUA, CUC, CUG, UUA, UUG	7.6 %

# Score-Matrizen

Diese Überlegungen führen auf folgende Formel:

Für Aminosäuren  $x, y$   
und eine gegebene evolutionäre Zeitspanne  $t$   
wird die Ähnlichkeit  $\text{Score}(x, y | t)$  definiert als

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

Hierbei ist:

- $M(x, y | t)$ : Relative Häufigkeit, mit der  $x$  und  $y$  gegeneinander ausgetauscht werden, wenn man Sequenzpaare betrachtet, die evolutionär  $t$  Zeiteinheiten auseinanderliegen
- $f(x)$ : Relative Häufigkeit der Aminosäure  $x$
- $f(y)$ : Relative Häufigkeit der Aminosäure  $y$

Die „Normalisierung“ mit  $f(x) \cdot f(y)$  rechnet aus  $M(x, y | t)$  die a-priori bekannten verschiedenen Häufigkeiten der Aminosäuren heraus.

Das Häufigkeitsverhältnis wird durch Logarithmieren additiv.

# Der Zeit-Parameter bei Score-Matrizen

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

Für kleine  $t$  sollte beispielsweise  $\text{Score}(x,x | t)$  positiv und groß sein,  
aber  $\text{Score}(x,y | t)$  stark negativ für  $x \neq y$ ,  
denn in einer kurzen Zeitspanne ist es nicht plausibel,  
dass viele Mutationen auftreten.

Für große  $t$  sollten diese Unterschiede weniger gravierend sein.

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

Man spricht von einer **Familie von Score-Matrizen**.

# Der Zeit-Parameter bei Score-Matrizen

In welcher **Zeiteinheit** misst man den Zeit-Parameter  $t$ ?

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

Also legt man fest:  $t = 1$  entspricht der Zeit, in der sich im Schnitt 1% der Aminosäuren einer Proteinsequenz ändern. Diese Einheit nennt man **PAM (percent accepted mutations)**.

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

Nein! Einige Aminosäuren ändern sich mehrfach, aber man „sieht“ im direkten Vergleich nur die letzte Änderung. Eine Mehrfachänderung kann auch zu einer sog. Rücksubstitution führen, so dass man gar keine Änderung mehr sieht (obwohl z.B. 3 stattgefunden haben).

Auch nach „unendlich“ viel Zeit sehen ca. 5% (1/20) der Positionen unverändert aus; selbst in Sequenzen, die überhaupt nichts miteinander zu tun haben, stimmt im Schnitt jede 20. Aminosäure überein!

# Beispiele für Familien von Score-Matrizen

Welche Zahlen man konkret als Ähnlichkeitswerte bekommt, hängt davon ab,

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

Hier gibt es sehr viele Möglichkeiten. Bekannt sind

- die **PAM-Familie** von M. Dayhoff (1978, Atlas of Protein Sequence and Structure, vol.5) [nicht mit der PAM-Zeiteinheit verwechseln!]
- die **BLOSUM-Familie** von Henikoff & Henikoff (1992)

## PAM- $t$ :

Die PAM-Familie von Score-Matrizen beruht auf 1,572 Austauschungen in 71 Familien sehr nah verwandter Proteine.

Für große Zeiten  $t$  werden die Austauschhäufigkeiten durch Extrapolation geschätzt.

## BLOSUM- $s$ :

Der Parameter  $s$  (zwischen 0 und 100) bei BLOSUM ist eine Art „inverse Zeit“: Großes  $s$  bedeutet, die Matrix enthält Werte für nah verwandte Sequenzen, kleines  $s$  bedeutet, die Matrix enthält Werte für entfernt verwandte Sequenzen.

Die BLOSUM-Matrizen basieren auf einer größeren Datenbasis als die PAM-Matrizen.

# Die PAM250-Matrix

Die PAM250-Matrix wird häufig für evolutionär relativ weit entfernte, aber noch eindeutig verwandte Proteine verwendet.

Um keine Fließkomma-Zahlen in der Matrix zu verwenden, wurden die Logarithmen der Häufigkeitsverhältnisse (log odds) mit 10 multipliziert.

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 BLOSUM62 Matrix

Die folgende Matrix („BLOSUM62“) ist heute die Standard-Score-Matrix für Proteinsequenz-Vergleiche.

	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>

# Peptid-Ähnlichkeiten

Wir können die Ähnlichkeit von zwei Aminosäuren quantifizieren.  
Wie messen wir die nun Ähnlichkeit zwischen zwei Peptiden  
(kurze Abfolgen von Aminosäuren) gleicher Länge?

Einfache Idee: durch die Summe der Aminosäure-Ähnlichkeiten

**Beispiel:** MVLS, MVHL

Die Summe ist

$\text{Score}(M,M) + \text{Score}(V,V) + \text{Score}(L,H) + \text{Score}(S,L) = 5 + 4 + (-3) + (-2) = 4$ ,  
wenn man BLOSUM62 zugrunde legt.

Daraus ergibt sich eine Visualisierungsmöglichkeit: Dot-Plots.

# Dot Plots: Visueller Vergleich von zwei Sequenzen

Ein Dot-Plot stellt die Positionen der einen Sequenz auf der x-Achse dar, die der anderen Sequenz auf der y-Achse.

Man wählt eine Fenstergrösse (Peptidlänge) und berechnet für jede Kombination von Startpositionen in beiden Sequenzen den Peptid-Ähnlichkeitsscore.

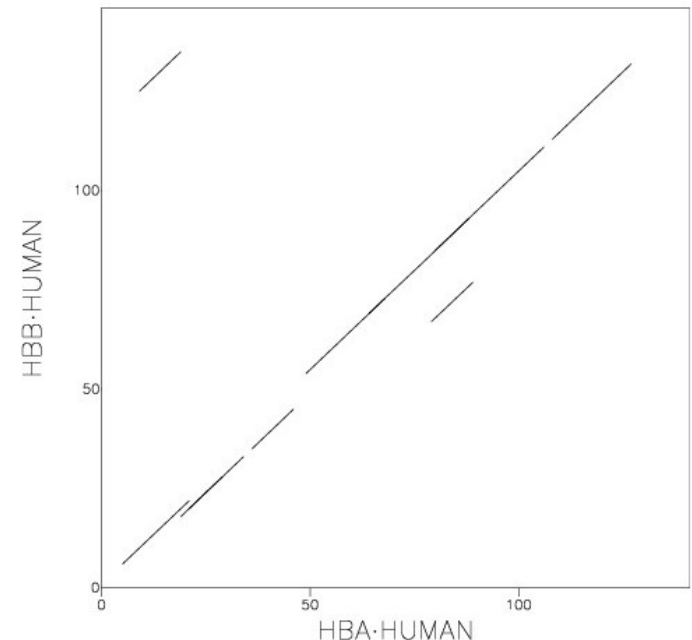
Jedes Positionspaar, für das der Peptid-Score einen definierten Schwellenwert überschreitet, markiert man mit einem Punkt (dot).

**Beispiel:** Hämoglobin Alpha/Beta,  
Peptidlänge 10, Score-Schwelle 23.

Was kann man erkennen?

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

Dotmatcher: HBA·HUMAN vs HBB·HUMAN  
(windowsize = 10, threshold = 23.00 02/09/02)



# Dotlet

Alternativ zur Definition eines Schwellenwertes kann der Peptid-Ähnlichkeitswert auch in Graustufen dargestellt werden (z.B. Schwarz für niedrig, weiß für hoch).

Selbst Dotplots erstellen kann man z.B. mit Dotlet unter <http://myhits.isb-sib.ch/cgi-bin/dotlet>

The screenshot displays the Dotlet 1.5 web interface. At the top right, it shows the user is 'anonymous' with a 'log in' button. Below the header, there are 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) Dotlet: diagonal plots in a web browser. *Bioinformatics*. 2000 Feb;16(2):178-9.' The interface includes a control bar with buttons for 'print', 'input', dropdown menus for 'alpha' and 'beta', a 'Blosum62' matrix dropdown, a '11' sliding window dropdown, a '1:1' zoom dropdown, and a 'compute' button. The main area shows a dotplot window with a zoomed-in view of the diagonal. To the right, a histogram window displays a distribution of scores, with a blue bar chart and a pink dashed curve. A text box above the histogram lists parameters: 'horizontal: alpha', 'vertical: beta', 'matrix: Blosum62', 'sliding window: 11', 'zoom: 1:1', 'score range: -44 to 121', and 'gray scale: 35% - 34%'.

# Dotlet - Demo

Demonstration mit den Hämoglobin-Sequenzen: <http://myhits.isb-sib.ch/cgi-bin/dotlet>

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

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

# Von Aminosäuren und Peptiden zu Proteinen

Wir können jetzt die Ähnlichkeit von zwei Aminosäuren und Peptiden quantifizieren.  
Wie messen wir die Ähnlichkeit zwischen zwei Proteinen (Ketten von Aminosäuren)?  
Es gibt (theoretisch) viele Möglichkeiten.  
Wir müssen berücksichtigen, dass es nicht nur Mutationen, sondern auch Einfügungen / Löschungen in einer Sequenz geben kann.

**Definition:** Ein (paarweises) Protein-**Sequenzalignment** besteht aus zwei Zeilen und mehreren Spalten.

In jeder Spalte stehen entweder zwei Aminosäuren oder ein „Gap“ und eine Aminosäure.  
Ein „Gap“ kennzeichnet eine Auslassung bzw. Einfügung in einer der Sequenzen.  
Ein „Gap“ wird als – geschrieben und 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.

**Beispiel:**

```
MV-LSPADKTNVKA AWGKVGA-HAGE  
|| |:|.:.|:.|.|.||||.. ..|  
MVHLLTPEEKSAVTALWGKVNVD EVG-
```

Die mittlere Zeile gehört nicht zum Alignment und dient nur der besseren Darstellung von Identität (|), hoher Ähnlichkeit (:), und Ähnlichkeit (.).

# Sequenzalignments

Wenn man noch festlegt, welchen Score eine Spalte mit Gap erhält, kann man jeder Alignment-Spalte einen Ähnlichkeitswert (Score) zuordnen, und diese Werte über alle Spalten summieren.

**Beispiel:** Sei Gap-Score = -8, verwende ansonsten BLOSUM62

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

In den ersten 5 Spalten ergibt sich ein Score von  $5+4+(-8)+4+1 = 6$ .

Zum Selber Rechnen: Welcher Score ergibt sich insgesamt?

Auf diese Weise erhält man den Score eines Alignments.

Es gibt aber viele Möglichkeiten (mit unterschiedlichen Scores), Alignments zu bilden, zum Beispiel hier auch

```
MVLSPADKTNVKAAWGKVGGAHAGE-
MVHLTPEEKSAVTALWGKVNVDVVG
```

```
MV-LSP----ADKTNV--KAAWGKVGGAHAGE
MVHLTPEEKSAV-TALWGKVNVDVVG-----
```

Welches Alignment ist das richtige, um die Sequenzähnlichkeit auszuwerten?

# Alignment-Score

?

MV-LSPADKTNVKAAWGKVGGA-HAGE  
MVHLTPEEKSAVTALWGKVNVDVVG-

MVLSPADKTNVKAAWGKVGGAHAGE-  
MVHLTPEEKSAVTALWGKVNVDVVG

MV-LSP----ADKTNV--KAAWGKVGGAHAGE  
MVHLTPEEKSAV-TALWGKVNVDVVG-----

Welches Alignment ist das richtige, um die Sequenzähnlichkeit auszuwerten?

Idee: Das mit dem höchsten Score!

Mehr Ähnlichkeit geben die Sequenzen nicht her.

Weniger nur, wenn man ihre Ähnlichkeiten nicht optimal herausarbeitet.

## Definition:

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

Ein Alignment mit maximalem Score wird **optimales (paarweises) Alignment** genannt.

Wie bekommt man das Alignment mit dem höchsten Score?

Nicht unser Problem (heute)!

Dafür gibt's Algorithmen, Programme, Webseiten: <http://www.ebi.ac.uk/emboss/align/>

# Berechnung optimaler Alignments

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

EMBOSS enthält verschiedene Programme, unter anderem solche zum Berechnen optimaler Alignments. [<http://emboss.sourceforge.net/>]

Das Alignment-Web-Formular am EBI ist abrufbar unter <http://www.ebi.ac.uk/emboss/align/>.

## 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

## Erklärungen:

**Method:** needle(global) oder water(local) -  
Sollen die gesamten Sequenzen aligniert werden, oder soll das am besten passende Stück gesucht und nur das aligniert werden?

**Gap open:** Kosten(!) für eine einzelne Gap-Spalte im Alignment (der Score ist das Negative davon)

**Gap extend:** Kosten für weitere konsekutive Gap-Spalten (z.B. bezahlt man hier für drei aufeinanderfolgende Gap-Spalten nur  $10+0.5+0.5 = 11$  statt  $3*10 = 30$ )

Man kann sowohl eine Sequenz in das Formular hineinkopieren, also auch eine auf dem Computer gespeicherte Datei hochladen.

### 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:

# Globale vs. Lokale Alignments

*Method*, needle(global) oder water(local):

Sollen die gesamten Sequenzen aligniert werden,  
oder soll das am besten passende Stück gesucht und nur das aligniert werden?

Woher kommen die Namen *needle* und *water*?

Von den Entdeckern der Verfahren (Algorithmen), mit denen optimale

- globale [Needleman SB, Wunsch CD (1970) J. Mol. Biol. 48; 443-453]
- lokale [Smith TF, Waterman MS (1981) J. Mol. Biol. 147(1); 195-197]

Alignments effizient berechnet werden können

Wann global, wann lokal?

Ein globales Alignment macht dann und nur dann Sinn, wenn beide Sequenzen einander über ihre gesamte Länge hinweg ähnlich (und damit auch in etwa gleich lang) sind.

Ein lokales Alignment sucht in jeder Sequenz nach dem am besten zur anderen Sequenz passenden Teilstück und aligniert nur diese Teilstücke.

Dies macht vor allem dann Sinn, wenn die Sequenzen nicht über ihre ganze Länge homolog (evolutionär verwandt) sind, sondern nur in Teilen.

# Gapkosten

Das 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, das während der Evolution in einer der Sequenzen Teile gelöscht wurden bzw. hinzugekommen sind.

Solche Ereignisse sind bei sehr nah verwandten Sequenzen recht selten, so dass das Einfügen eines Gaps in ein Alignment mit einem niedrigen (stark negativen) Score gewertet wird.

Es ist plausibel, dass in so einem Fall (wenn es überhaupt passiert) nicht genau eine Aminosäure gelöscht oder hinzugefügt wird, sondern evtl. gleich mehrere. Daher sollte man einen zusammenhängenden Gap der Länge 3 anders behandeln als drei einzelne Gaps der Länge 1.

Verwendet man geringere Gap-extend Kosten als Gap-open Kosten, spricht man von **affinen Gap-Kosten**.

Sind die Kosten für open und extend gleich, spricht man von **linearen Gap-Kosten**. Es sind noch kompliziertere Modelle denkbar.

# Suche in Sequenz-Datenbanken

Eine häufig auftretende Situation ist die, dass man irgendwo her ein „Stück“ Sequenz (DNA, RNA, oder Protein) erhält, und etwas darüber herausfinden möchte.

Man sucht daher nach ähnlichen Sequenzen, über die schon etwas bekannt ist. Es wäre aber lästig, wenn man jede bekannte Sequenz in das EMBOSS-Formular eingeben müsste, bis man endlich eine relativ ähnliche findet!

Daher gibt es Datenbanksuchprogramme, die automatisch

- eine Sequenz (die Anfrage oder Query)
- mit jeder Sequenz einer Sequenzdatenbank vergleichen und nur die hinreichend ähnlichen Datenbank-Sequenzen, ggf. mit Alignments, ausgeben.

Das bekannteste solche Programm (eigentlich eine Programmsammlung) heißt **BLAST** (**B**asic **L**ocal **A**lignment **S**earch **T**ool).

Es arbeitet auch auf großen Datenbanken sehr schnell und sucht nach lokalen Ähnlichkeiten zwischen Query und Datenbank-Sequenzen. Es ist am NCBI unter <http://www.ncbi.nlm.nih.gov/blast/Blast.cgi> verfügbar.

# BLAST-Webseite am NCBI

BLAST besteht in Wirklichkeit aus vielen einzelnen speziellen Werkzeugen. Diese können auf der Übersichtsseite ausgewählt werden.

The screenshot shows the top navigation bar of the NCBI BLAST website. It includes the BLAST logo and the text 'Basic Local Alignment Search Tool'. Below the navigation bar are tabs for 'Home', 'Recent Results', 'Saved Strategies', and 'Help'. The main content area features a heading 'BLAST Assembled Genomes' and a list of species genomes to search, including Human, Mouse, Rat, Arabidopsis thaliana, Oryza sativa, Bos taurus, Danio rerio, Drosophila melanogaster, Gallus gallus, Pan troglodytes, Microbes, and Apis mellifera. Below this is the 'Basic BLAST' section, which lists various BLAST programs such as nucleotide blast, protein blast, blastx, tblastn, and tblastx, each with a brief description of its search criteria and algorithms.

Protein oder DNA-  
Sequenz gegen  
ganze Genome

Auswahl nach Sequenztyp  
der Query und der Datenbank

<http://www.ncbi.nlm.nih.gov/blast/Blast.cgi>

# BLAST-Varianten

Programm	Querytyp	DB-Typ	Bemerkung
blastn	Nukleotid	Nukleotid	standard, auch für entfernte Sequenzen
megablast	Nukleotid	Nukleotid	nur für sehr ähnliche Seq., sehr schnell
disc. megabl.	Nukleotid	Nukleotid	auch für etwas weiter entfernte Sequenzen
blastp	Protein	Protein	standard
psi-blast	Protein	Protein	langsamer, findet entfernte Verwandte
phi-blast	Protein	Protein	Einschränkung auf bestimmte Motive
blastx	Nukleotid (T)	Protein	(*)
tblastn	Protein	Nukleotid (T)	(*)
tblastx	Nukleotid (T)	Nukleotid (T)	(*)

## 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

Die mit (\*) markierten Programme vergleichen auf Protein-Ebene.

Mit (T) markierte Nukleotid-Sequenzen werden zuvor mittels genetischem Code in Proteinsequenzen übersetzt.

# Protein-BLAST (blastp)

► [NCBI/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 [Non-redundant protein sequences \(nr\)](#)

Organism [Optional](#)   
Enter organism name or id--completions will be suggested  
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?

Hier können weitere Parameter eingestellt werden (nächste Folie).

# Parameter von Protein-BLAST (blastp)

The screenshot shows the 'Algorithm parameters' section of the NCBI BLAST web interface. It is divided into three sub-sections: 'General Parameters', 'Scoring Parameters', and 'Filters and Masking'.  
- **General Parameters:** 'Max target sequences' is set to 100. 'Short queries' has a checked box for 'Automatically adjust parameters for short input sequences'. 'Expect threshold' is 10, marked with (\*\*\*) and a help icon. 'Word size' is 3, marked with (\*).  
- **Scoring Parameters:** 'Matrix' is BLOSUM62. 'Gap Costs' is 'Existence: 11 Extension: 1'. 'Compositional adjustments' is 'Conditional compositional score matrix adjustment', marked with (\*\*\*) and a help icon.  
- **Filters and Masking:** 'Filter' has a checked box for 'Low complexity regions'. 'Mask' has unchecked boxes for 'Mask for lookup table only' and 'Mask lower case letters'.  
At the bottom, there is a 'BLAST' button and a checkbox for 'Show results in a new window'.

Wie viele Treffer werden maximal angezeigt?

Für kurze Queries Einstellungen anpassen?

Wortlänge: Bevor eine Sequenz als Treffer gelten kann, muss eine Übereinstimmung dieser Länge mit der Query festgestellt werden.

Welche Scorematrix und Gapkosten sollen verwendet werden?

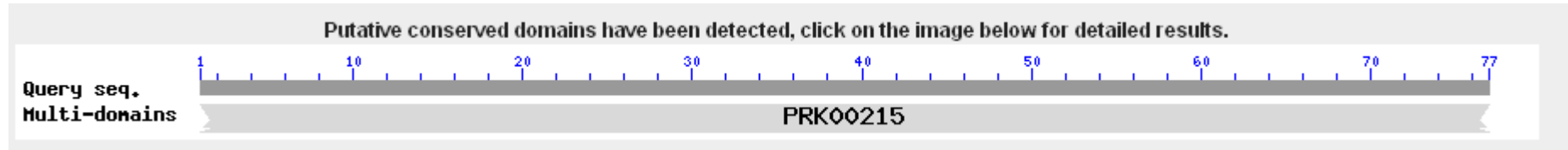
Soll die Sequenz vor der Suche noch gefiltert werden, z.B. bekannte Repeats?

Zu (\*): Eine grössere Wortlänge macht BLAST schneller, führt aber dazu, dass entfernt verwandte Sequenzen u.U. nicht mehr gefunden werden.

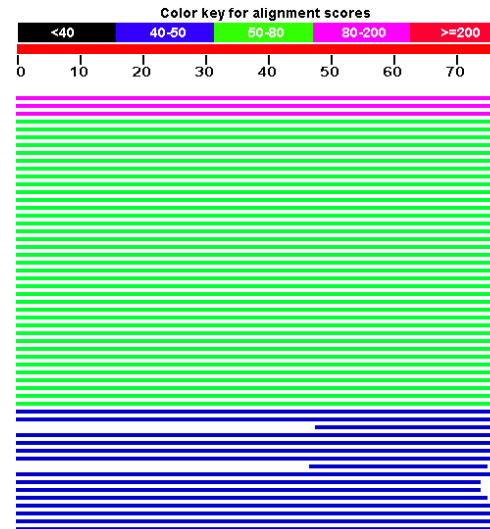
Mit (\*\*\*) markierte Parameter erfordern ein wenig mehr Theorie! Dazu gleich mehr.

# Ergebnisliste von BLAST

Bereits während der Suche wird eine Graphik mit in der Query enthaltenen möglichen „conserved domains“ angezeigt.  
Das sind wiederkehrende strukturelle und funktionelle Einheiten, aus denen Proteine modular aufgebaut sind.



Es folgt eine graphische Übersicht über die besten Treffer in der Datenbank.  
Die Ausdehnung des Balkens zeigt an, wo das ähnlichste Segment mit der Datenbank-Sequenz in der Query liegt (lokales Alignment!).  
Die Farbe zeigt den Grad der Ähnlichkeit an (hier gemessen in Prozent Identität).



# Ergebnisliste von BLAST

Es folgt die 5-spaltige Ergebnisliste mit gefundenen Treffern in der Datenbank.  
Die Spalten enthalten folgende Informationen:

1. Link + Kombination von Schlüsseln der gefundenen Sequenz in den Datenbanken
2. Name (evtl. Kurzbeschreibung) des gefundenen Datenbank-Eintrags
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)

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 [Nocardia 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>

# Ergebnisliste von BLAST

Unter der Liste mit einzeiligen Beschreibungen folgen detaillierte Informationen zu den einzelnen Treffern, insbesondere auch die Alignments, sowie Eigenschaften der Alignments (identities, positives, gaps).

Bei der Bewertung der Alignments werden verschiedene Zahlen angegeben:

- Score in bits (80.9)
- Score in Rohform (in Klammern, 198)
- E-Wert (Expect-Wert,  $3 \cdot 10^{-14}$ )
- Mit welcher Methode 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---SVAISPDPAETSPTS FVPIVGSIAAG 57
RDPNKPRAVDVR LP+ + K GPK + SP P S TSF+P+VG IAAG
Sbjct 103 RDPNKPRAVDVRHLPE---TDMRTKAGPKAKARPTAGAS PQPELASSTS FIPVVGKIAAG 159

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

# BLAST-Statistik: Bewertung von Alignments

Bei der Bewertung der Alignments angegeben:

- Score in bits (80.9)
- Score in Rohform (in Klammern, 198)
- E-Wert (Expect-Wert,  $3 \cdot 10^{-14}$ )
- Berechnungsmethode des E-Werts

```
>[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 RDPMKPRAVDVRLPDPPIPSKPKGRKPGPKKS---SVAISPDPAETSPTSFPVPIVGSIAAG 57
RDPMKPRAVDVRLP+ + K GPK + SP P S TSP+P+VG IAAG
Sbjct 103 RDPMKPRAVDVRLP+ + K GPK + SP P S TSP+P+VG IAAG 159

Query 58 NPILAEENVDCYFPFPSEIV 77
+PILAE+M++ Y+P P++IV
Sbjct 160 SPILAEQNIEEYYPLPADIV 179
```

Der Score in Rohform (198) ist die Summe der Scores über die Spalten des Alignments.

Intuitiv: Längere Queries haben eine höhere Chance, ein gutes lokales(!) Alignment mit Datenbanksequenzen zu bekommen. Je länger Sequenzen sind, um so länger werden auch lokale Ähnlichkeiten, die nur auf Zufall basieren.

Ausserdem ist unklar, auf welcher Skala der Roh-Score angegeben wird. Beispielsweise könnte man alle Scores mit 10 multiplizieren, ohne dass sich grundsätzlich etwas ändert.

Also gibt man einen längennormalisierten, genormten Score-Wert (Bit-Score) an.

Man weiss aber noch nicht, ab wann ein Bit-Score wirklich (sehr) hoch ist. Daher wird ein Bit-Score in eine besser interpretierbare Zahl umgerechnet: den E-Wert.

# BLAST-Statistik: E-Werte

Bei der Bewertung der Alignments angegeben:

- Score in bits (80.9)
- Score in Rohform (in Klammern, 198)
- E-Wert (Expect-Wert,  $3 \cdot 10^{-14}$ )
- Berechnungsmethode des E-Werts

```
>[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   RDPMKPRAVDVRLPDPPIPSKPKGRKPKKS---SVAISPDPAETSPTS FVPIVGSIAAG 57
          RDPMKPRAVDVRLP+ + K GPK + SP P S TSP+P+VG IAAAG
Sbjct 103 RDPMKPRAVDVRLP+ + K GPK + SP P S TSP+P+VG IAAAG 159

Query 58  NPILAEENVDCYFPFPSEIV 77
          +PILAE+M++ Y+P P++IV
Sbjct 160  SPILAEQNIEEYYPADIV 179
```

Der E-Wert gibt an:

Wie viele mit diesem oder höheren Bit-Score würde man ausschließlich aufgrund zufälliger Sequenzähnlichkeiten erwarten?

Ist dieser Wert sehr klein (also z.B.  $< 10^{-10}$ ), ist es sehr unwahrscheinlich, dass die beobachtete Ähnlichkeit nur auf einem zufällig ähnlichen Stück Sequenz beruht, und man darf hoffen, dass eine evolutionäre Verwandtschaft vorliegt.

Ist der E-Wert hoch (nahe bei oder größer als 1), lässt sich dieser Treffer auch durch puren Zufall erklären und sollte daher nicht notwendigerweise als biologisch relevant gelten.

# BLAST-Statistik: E-Werte und Parameter

Wir klären jetzt die Rolle der verbleibenden (\*\*\*) Parametereinstellungen.

Es gibt mehrere Methoden, „zufällige Ähnlichkeit“ zu formalisieren. Diese kann man bei den Parametern unter „Compositional adjustments“ variieren. Die Voreinstellung ist vernünftig, wir gehen nicht weiter darauf ein!

Durch eine Angabe einer Grenze für den E-Wert („Expect threshold“) lassen sich die Treffer, die BLAST ausgibt, begrenzen. Bei der Voreinstellung 10 werden nur Treffer mit einem E-Wert  $< 10$  ausgegeben.

Will man die Suche schneller und strikter gestalten (und riskieren, dass weiter entfernte verwandte Sequenzen nicht mehr gefunden werden), kann man hier z.B. auch  $1E-6$  o.ä. verwenden.

The screenshot shows the 'Algorithm parameters' section of a BLAST search interface, divided into three sub-sections:

- 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

At the bottom, there is a 'BLAST' button and a checkbox for 'Show results in a new window'.