WS2018/2019

JOHANNES GUTENBERG UNIVERSITÄT MAINZ

„Genomforschung und Sequenzanalyse

- Einführung in Methoden der Bioinformatik- “

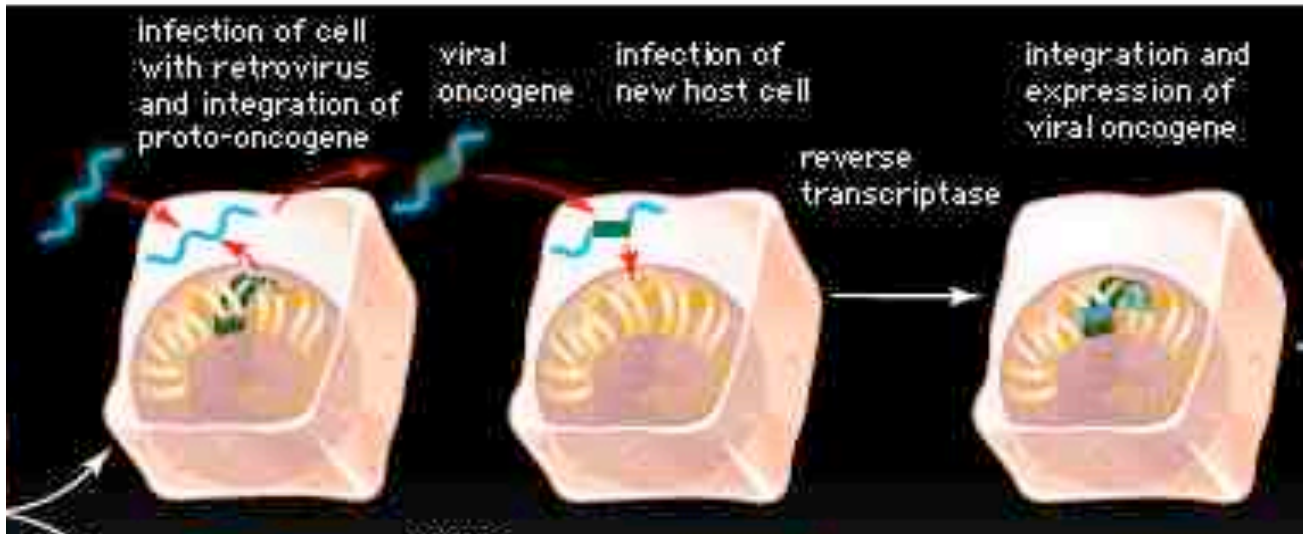
Thomas Hankeln

Alignments & Datenbanksuchen

break-thru…

Doolittle et al. 1983, Waterfield et al. 1983 > DB-Suche

„... das virale Oncogen v-sis ist eine modifizierte Form des zellulären Gens für den platelet-derived growth factor (PDGF)!!“



Spezielle Such-Algorithmen erforderlich…

- „optimale“ Algorithmen wie N-W oder S-W sind viel zu aufwändig für das Durchsuchen großer Datenbanken
- „Heuristische“ Methoden des Sequenzvergleichs ermöglichen schnelle Alignments, jedoch mit geringer Gefahr, eine noch besser passende Sequenz zu übersehen.

Heuristik (altgr. εὑρίσκω *heurísko* „ich finde“; von εὑρίσκειν *heurískein* ‚auffinden‘, ‚entdecken‘) bezeichnet die Kunst, mit begrenztem Wissen (unvollständigen Informationen) und wenig Zeit dennoch zu wahrscheinlichen Aussagen oder praktikablen Lösungen zu kommen.[1]

Allgemeine Strategie heuristischer Methoden

- Suchsequenz in kurze Abschnitte („words“ bzw. „*k*-tuple“) aufbrechen (Wilbur und Lipman, 1983).

- zunächst sehr schnell nach „word hits“ in der DB suchen

- hat man mehrere „word hits“ in einem DB-Eintrag? Dann dort genauer schauen...

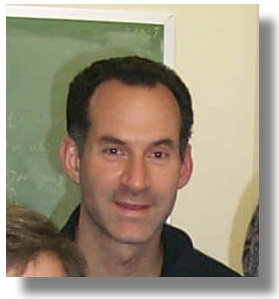
Zwei Programmfamilien für die DB-Suche

- **FASTA (Lipman und Pearson 1983)**
- **BLAST (Altschul et al. 1991, 1997)**

  „Basic Local Alignment Search Tool“

Beide Tools machen lokale Alignments!

Stephen Altschul

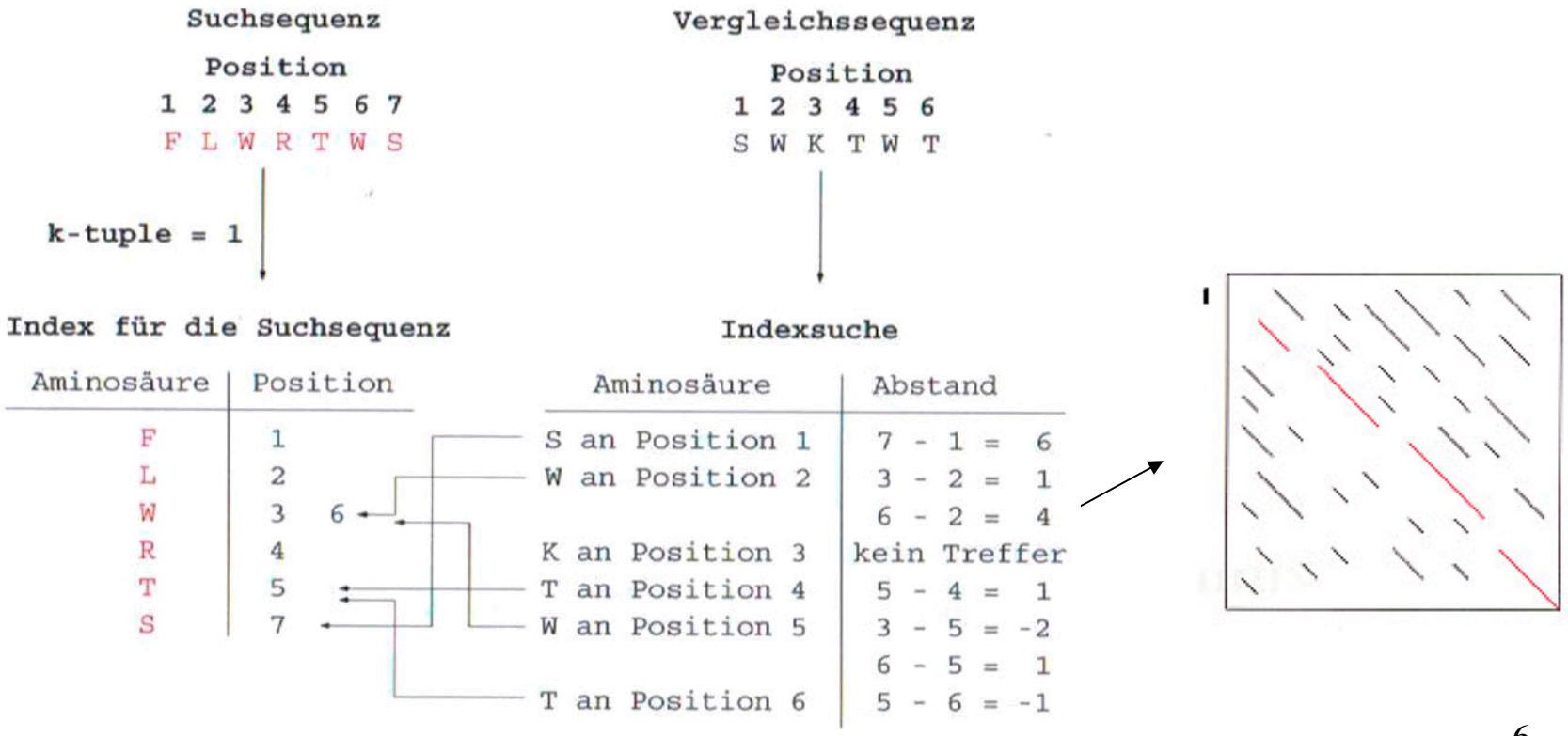


David Lipman

Bill Pearson

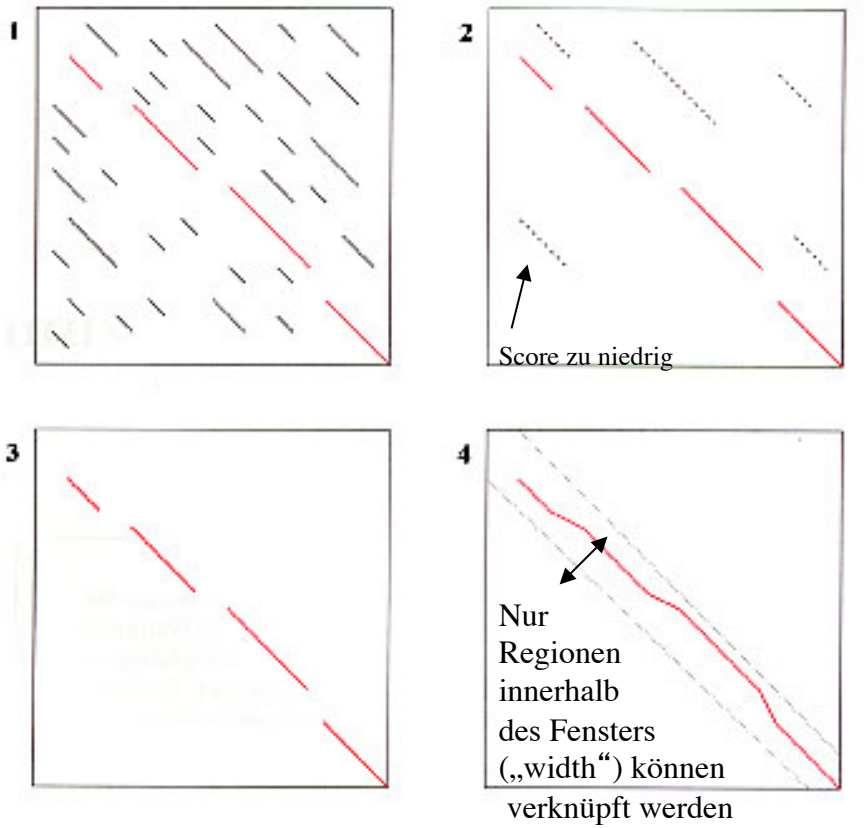
FASTA

1. Erzeugung eines Index (lookup table) der Suchsequenz:
   > Länge der Index-Einträge = k-tuple
   > mit Index wird nach ident. Posit. in Vergleichssequenz gesucht



aus: Hansen, (2004) Bioinformatik, ein Leitfaden...

FASTA



2. Verlängerung der ersten Matches ohne gaps; Berechnung des init1-Scores

3. Verknüpfung unter Einführen von gaps (initn-Score = init1 - joining penalty)

4. Verknüpfung der init1-Regionen mit höchstem Score (> „opt“) nach der sensitiveren S-W-Methode
(nach Normalisierung auf Länge wird Z-Score angegeben)

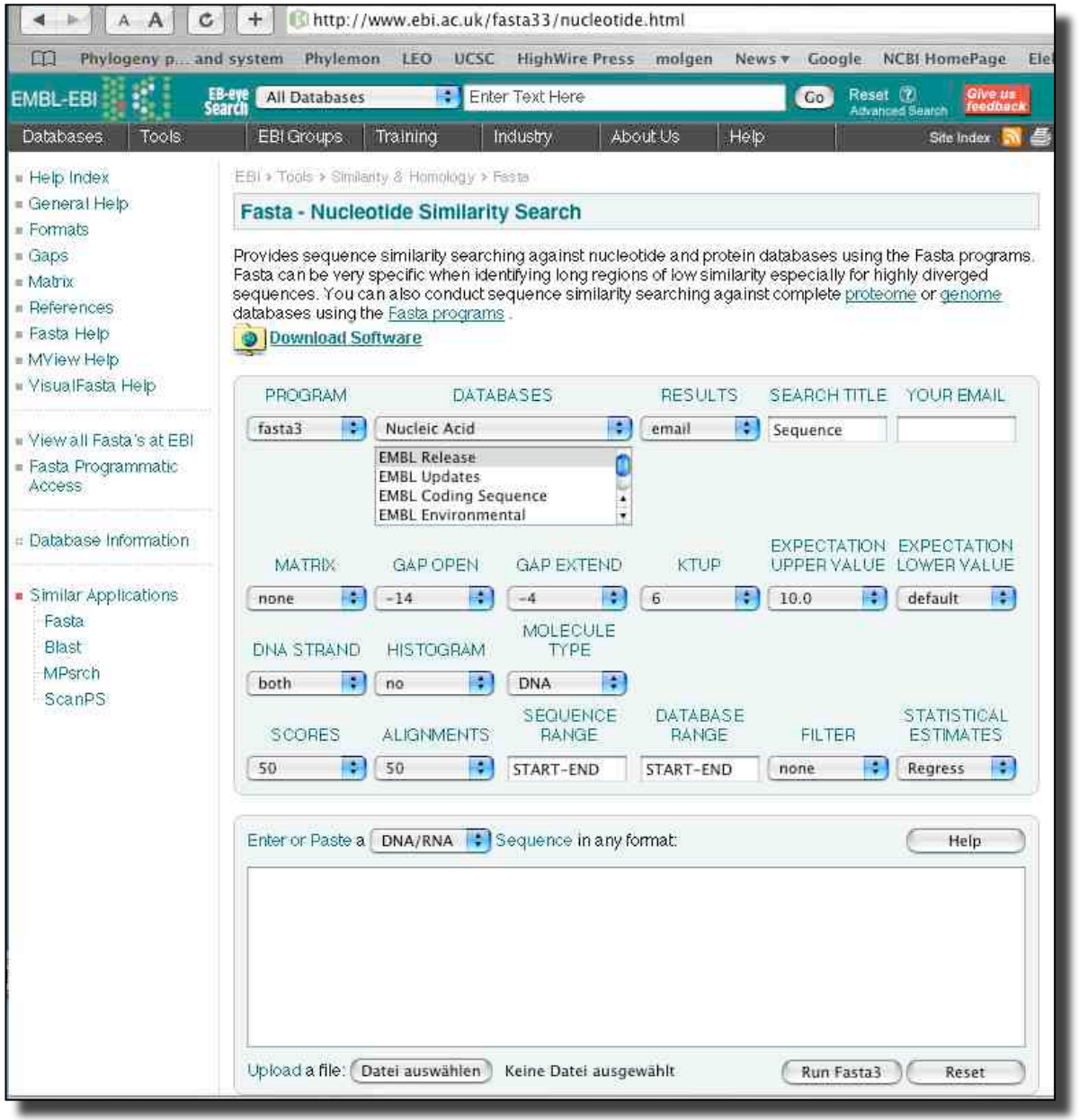
FASTA

- sensitiv, aber vergleichsweise etwas langsam
- *default:* k-tup (DNA) = 6, k-tup (Protein) = 2
- höheren k-tup
  > mehr speed, weniger noise, weniger Sensitivität

  niedrigeren k-tup > höhere Sensitivität für entfernte Matches

- größter Nachteil: nur ein einziges optimales lokales Alignment wird gezeigt

  > Nachbearbeitung (z.B. mit LALIGN) erforderlich

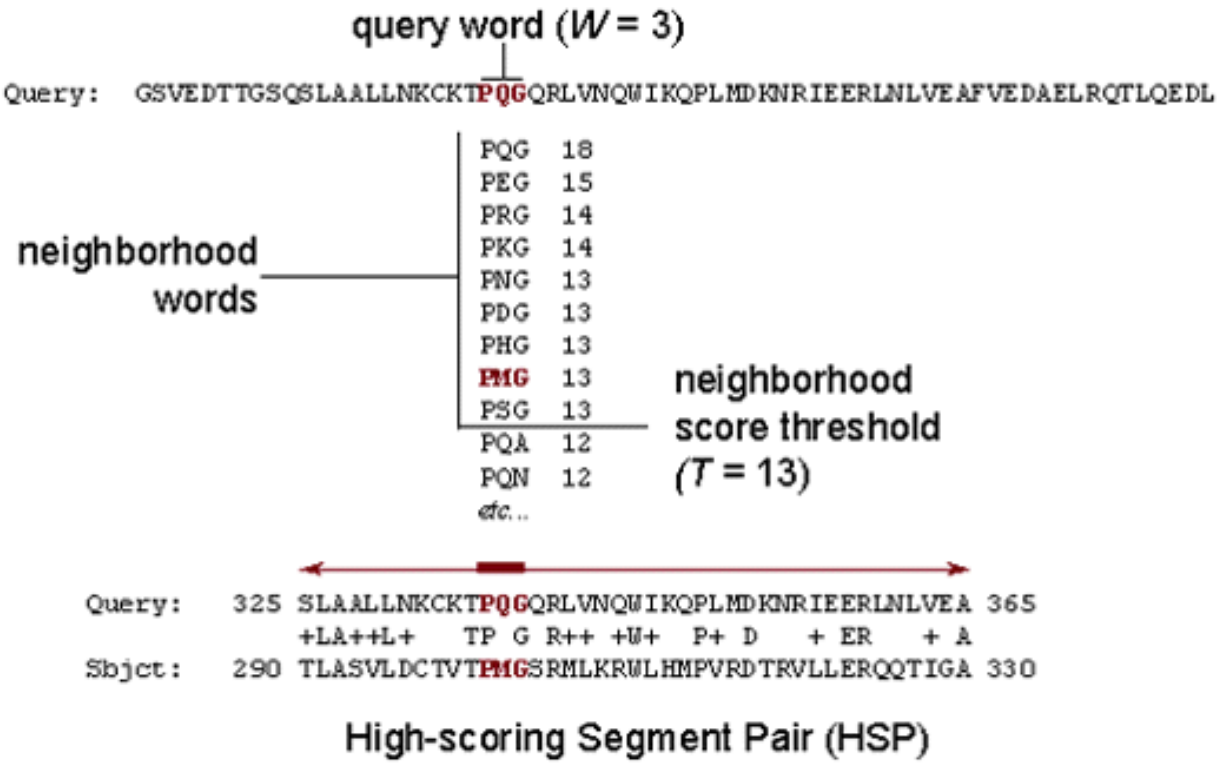


Verschiedene Typen von FASTA-Suchen stehen auf EBI-Seite zur Verfügung...

BLAST

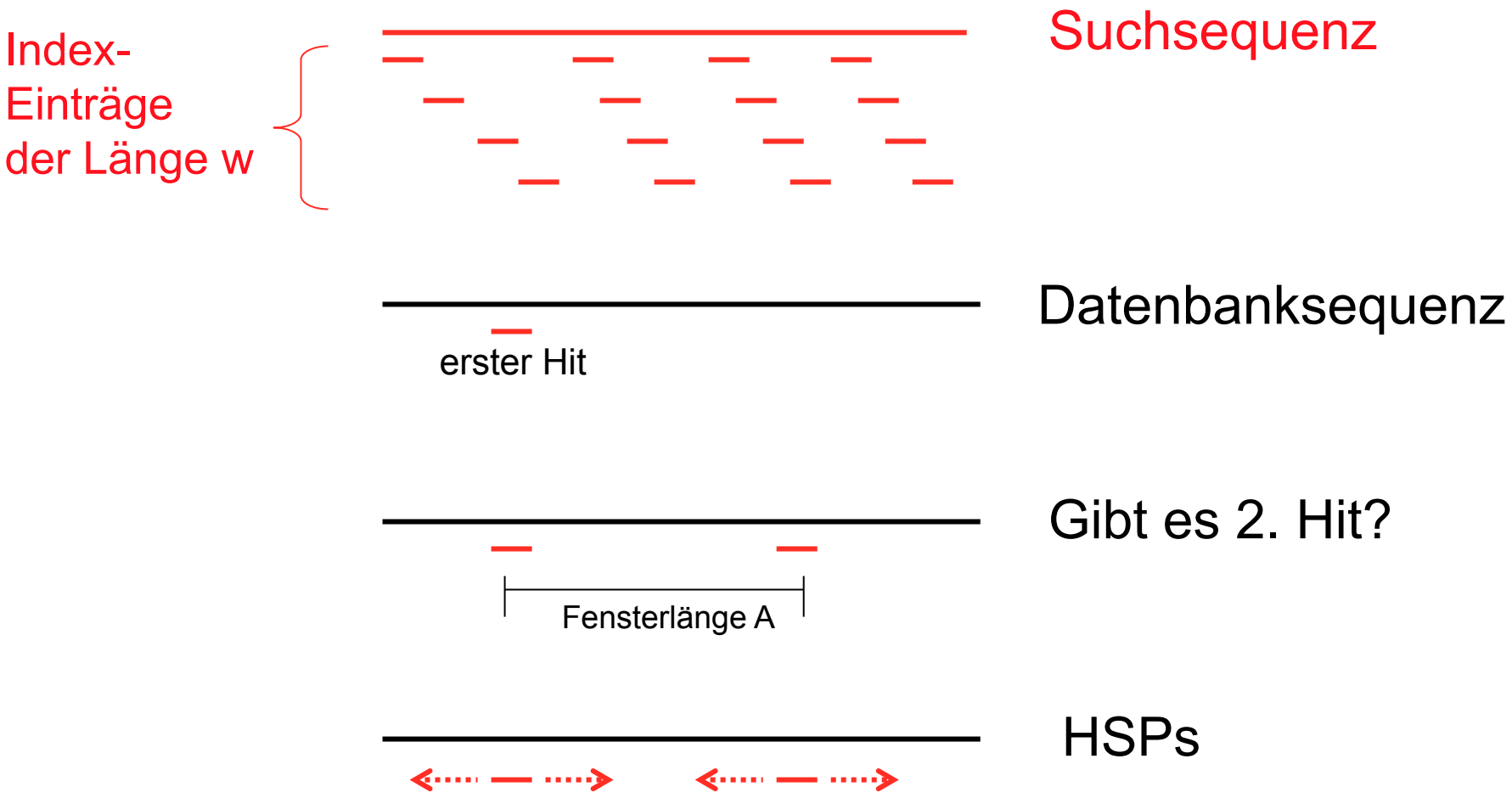
Altschul et al. 1990, 1997
> 74000 Zitate

- schneller als FASTA!
- liefert mehrere lokale Alignments
- berücksichtigt Ähnlichkeiten!



(word size W = 11 bei DNA)

BLAST



zwei lokale Alignments,
Verknüpfung über Lücken falls möglich erlaubt

BLAST

1. Suchsequenz wird in ‚words' der Länge w „zerbrochen"
2. mit Index dieser ‚words' wird DB durchsucht
3. ein „word hit" liegt vor, wenn das ‚word' exakt oder in ähnlicher Form* (threshhold-Score >T) erkannt wird

   > word size kann hoch bleiben (speed) ohne Sensitivitätsverlust
   > erhöhe T : weniger ‚background words', schneller
   > erniedrige T : entfernte Verwandschaften zu finden

4. ausgehend von ‚word hit' wird lokales optimales alignment verlängert, bis Score S durch mismatches stark abfällt (= HSP, high-scoring segment pair)

   > dabei können kleine Lücken toleriert werden

*das kann FASTA nicht!

BLAST bewertet die Signifikanz eines Alignments !!

Score E-Wert

3.	dbj|BAA29916| (AP000003) 170aa long hypothetical protein [P...	107	6e-23
4.	sp|Q57951|Y531_METJA HYPOTHETICAL PROTEIN MJ0531 >gi|212801...	91	4e-18
5.	gi|2622094 (AE000872) conserved protein [Methanobacterium t...	85	4e-16
6.	gi|2621993 (AE000865) conserved protein [Methanobacterium t...	81	4e-15
7.	gi|2621194 (AE000803) conserved protein [Methanobacterium t...	80	7e-15

$$E = k\ m\ n\ e^{-\lambda S}$$

k, Konstante
λ, Konstante für Normalisierung des HSP-scores
m, Nukleotidanzahl in Suchsequenz
n, Nukleotidanzahl in Datenbank
S = score des HSP-matches

Der E (Expect)-Wert gibt die Zahl der Treffer an, die in einer Datenbank der verwendeten Größe zufällig erwartet werden können.

(Je kleiner der Wert, desto höher die Signifikanz des betrachteten Treffers)

https://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html

Wann habe ich einen guten Treffer?

Faustregel:

- DNA: < e -6, >60% Sequenzidentität
- Protein: < e -3, >25% Identität

Grenzfall „Neuroglobin“:

```
                                                                  Score     E
Sequences producing significant alignments:                       (bits)  Value
dbj|AU036042.1|AU036042  AU036042 Sugano mouse brain mncb Mu...    41   0.003
gb|BE648697.1|BE648697  UI-M-BG1-aid-e-09-0-UI.r1 NIH_BMAP_M...    37   0.045
gb|AW548186.1|AW548186  L0032E08-3 Mouse E12.5 Female Mesone...    32   0.89
gb|AW546198.1|AW546198  L0005A02-3 Mouse E12.5 Female Mesone...    32   0.89
gb|AW548428.1|AW548428  L0036F07-3 Mouse E12.5 Female Mesone...    32   1.1
emb|AL362383.1|AL362383  AL362383 ICRFp 522 and 523 Mus musc...    32   1.3


dbj|AU036042.1|AU036042 AU036042 Sugano mouse brain mncb Mus musculus cDNA clone MNCb-7114.
Length = 740  Score = 40.8 bits (126), Expect = 0.003
Identities = 33/154 (21%), Positives = 63/154 (40%), Gaps = 5/154 (3%)
Frame = +3

Query: 1  MNSDEVQLIKKTWEIPVATPTDSGAAILTQFFNRFPSNLEKFPFRDVPL---EELSGNAR 57
          M   E +LI+++W +  +P + G  +  + F   PS L F +       E+   +
Sbjct:156 MERPESELIRQSWRVVSRSPLEHGTVLFARLFALEPSLLPLFQYNGRQFSSPEDCLSSPE 335
```

BLAST :
Endecke die Möglichkeiten...

blastn DNA-Sequenz ÷ DNA-DB

> für nahe Verwandschaft; beide Stränge verglichen

blastp As-Sequenz ÷ Protein-DB

> für entfernte Verwandschaft (default: BLOSUM62)

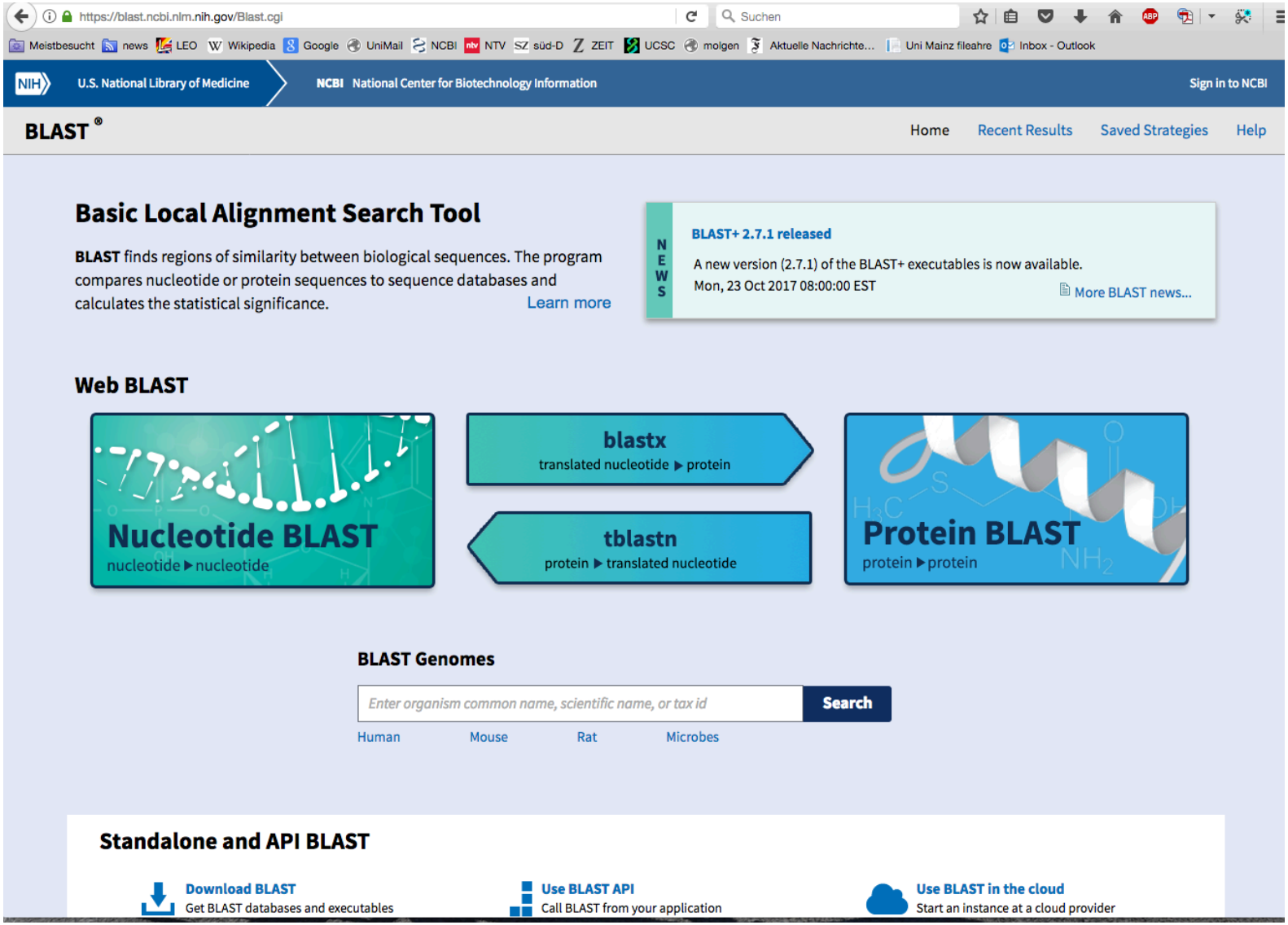
Ich habe die DNA-Sequenz aus einer exotischen Spezies neu entschlüsselt.

Ich will wissen, ob diese DNA-Sequenz ein bekanntes Protein kodiert, und welches Protein aus welcher anderen Spezies am Ähnlichsten ist...

Was muss BLAST idealerweise können, um das zu beantworten?

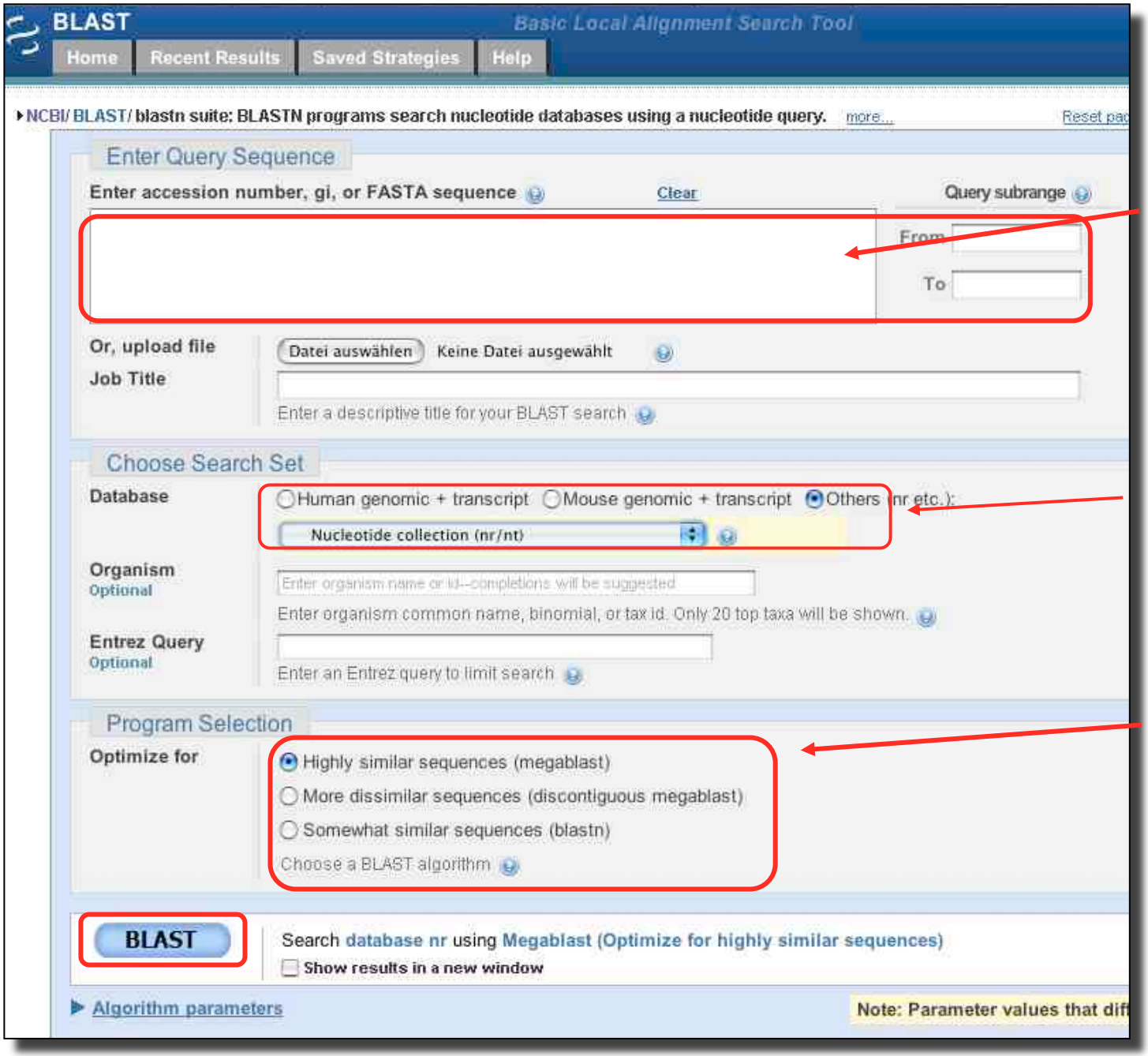
BLAST : Endecke die Möglichkeiten...

blastx	DNA-Seq > in 6 Leserahmen translatiert ÷ Protein-DB > findet mögliche Proteine in einer nicht-charakterisierten („anonymen“) DNA-Sequenz (z.B. EST)!
tblastn	As-Seq gegen DNA-DB (6-frame translatiert!) > findet nicht-annotierte Genregionen in DNA-DB-Sequenzen
tblastx	6-frame-Translation einer DNA-Seq ÷ 6-frame-Translation einer DNA-DB > Analyse von ESTs auf Proteinebene zur Detektion entfernter Verwandschaft > kann nicht mit nr-DB benutzt werden (zu aufwändig)



Viele spezialisierte BLAST-Optionen weiter unten auf der Seite...

BLAST-Suche (1)



Copy/paste

DB wählen!

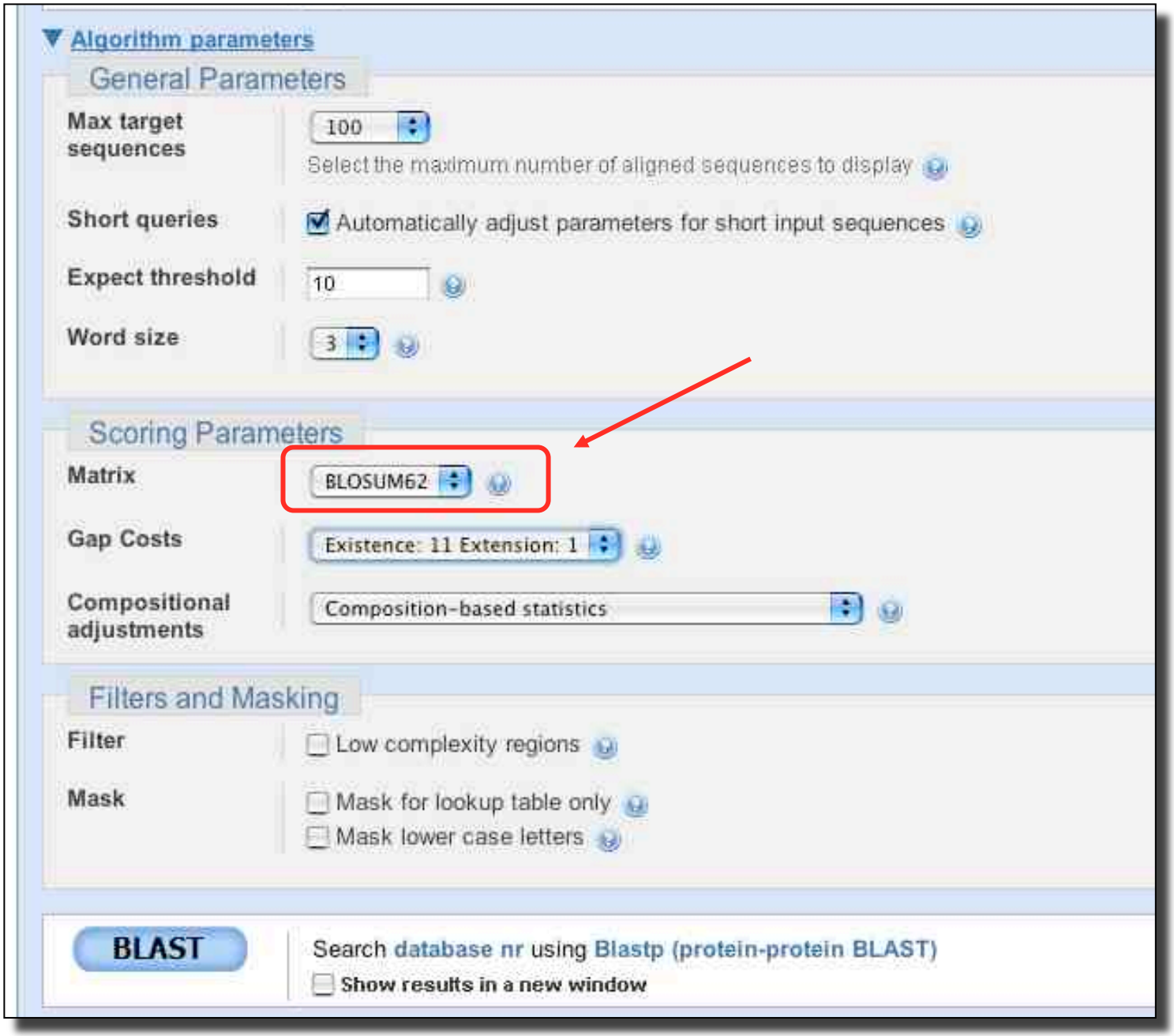
Algorithmus wählen

BLAST-Algorithmen auf Nt-Ebene

- Megablast: längere Word size, daher schneller
  für gut passende matches, aber weniger
  sensitiv als BlastN:
  für Suchen mit >80 % Identität

- *discontiguous*
  Megablast: „unterbrochene“ word hits erlaubt;
  ignoriert mismatches der 3. Kodonposition
  in kodierenden Sequenzen;
  sensitiver als BlastN für entfernte Suchen

- BlastN: „gut für den Rest...“



BLAST-Suche (2)

evtl. die Parameter verändern

BLAST-Suche (3) ..das Ergebnis

BLASTN 2.2.4 [Aug-26-2002]

Reference:
Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database search
programs", Nucleic Acids Res. 25:3389-3402.

RID: 1038246134-013402-11992

Query=
(626 letters)

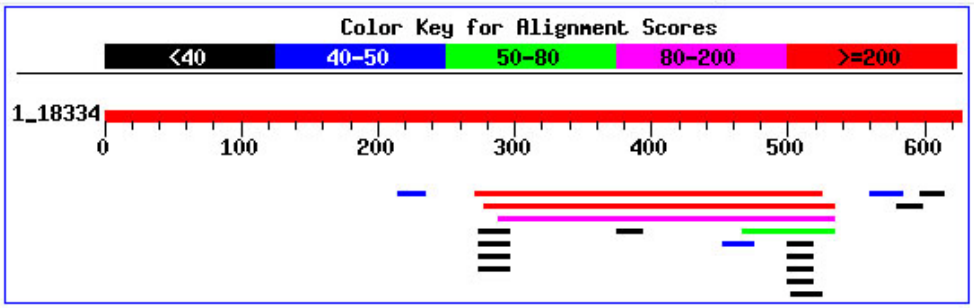
Database: GenBank non-mouse and non-human EST entries
6,214,058 sequences; 3,097,472,311 total letters

If you have any problems or questions with the results of this search
please refer to the **BLAST FAQs**

Taxonomy reports

Distribution of 19 Blast Hits on the Query Sequence

Mouse-over to show defline and scores. Click to show alignments

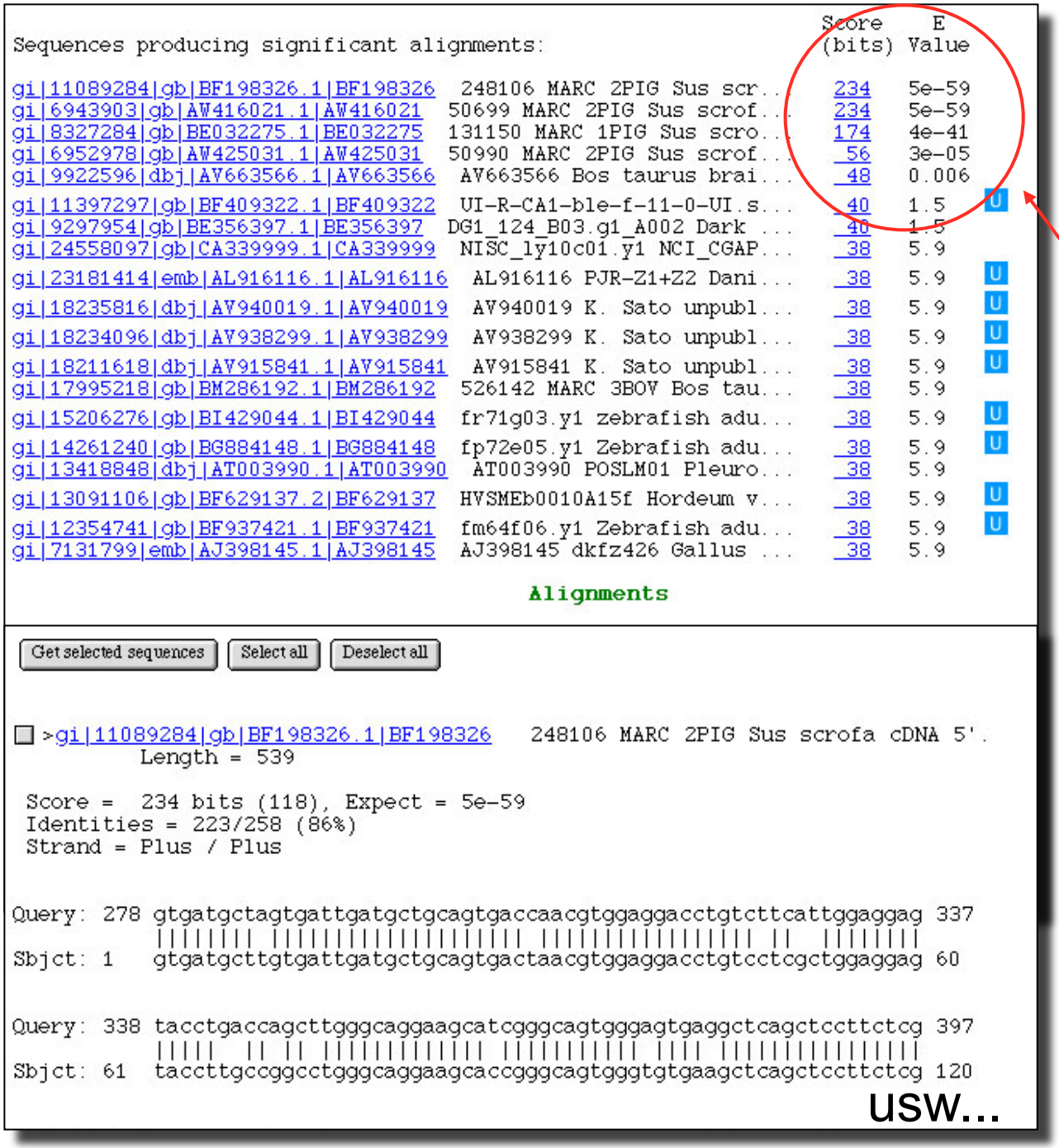


Zur Erinnerung...

Suchsequenz („Query“)

Matches mit unterschiedlicher Qualität

BLAST-Suche (4) ..das Ergebnis



...das erste Alignment
(Query = Suchsequenz)

Anschauen:
- **Score >50**
- **E << 1**

PSI-BLAST

Position-specific iterated BLAST

- speziell für die Suche sehr entfernt verwandter Proteine, die durch BLASTP nicht gefunden werden

Multiples Alignment

1..10	11	12	13	14	15	..	342	Position	
..	E	L	V	K	A	G	..	**Suchsequenz mit L = 342**	
..	G	L	V	E	P	G	..	Vergleichssequenzen	
..	V	S	V	A	L	G	..	aus der Datenbank	
..	L	P	V	T	P	G	..		
..	V	L	V	A	P	G	..	Konsensussequenz	= „Profil“

Positionsspezifische Matrix

		A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
1																					
.	.																				
.	.																				
10	V	3	-2	3	4	0	4	-1	3	-1	4	4	1	1	1	-2	1	2	6	-6	-2
11	L	2	-2	-2	-1	3	0	-1	3	-1	6	5	-1	3	0	-1	3	1	4	1	-1
12	V	2	2	-2	-2	2	2	3	11	-2	8	6	-2	1	-2	-2	0	2	15	-9	-1
13	A	6	-2	5	6	-5	4	1	0	5	-2	0	3	3	3	1	3	6	0	-6	-4
14	P	6	-1	0	1	-2	2	0	1	0	2	2	0	8	2	0	2	2	3	-5	-4
15	G	7	1	7	5	-6	15	-1	-3	0	-4	-3	4	3	2	-3	6	4	2	-11	-7
.	.																				
.	.																				
342																					

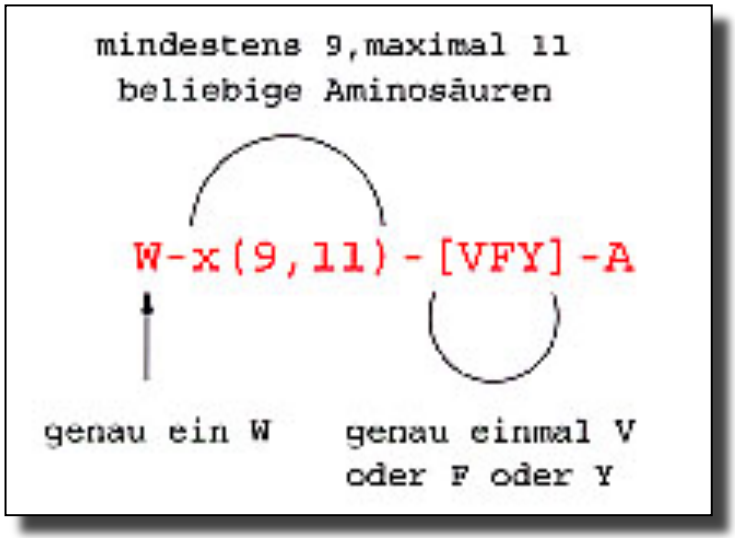
1. Erste Suche = einfacher BLAST
2. Matches untereinander schreiben,
   > Konsensussequenz errechnen („Profil“)
3. „Positions-spezifische“ Substitutions-Matrix errechen
4. BLAST mit dem „Profil“ und der PSSM mehrfach wiederholen

➢ Gezieltere Suche nach verwandten Proteinen wegen Positionsinformation

PHI-BLAST

Pattern-hit initiated BLAST

- sucht Sequenzmuster („Signatur“), das typisch für Proteindomäne ist
- Suche über „qualitatives“ Sequenzmotif (PSI-Blast über quantitatives Motiv)

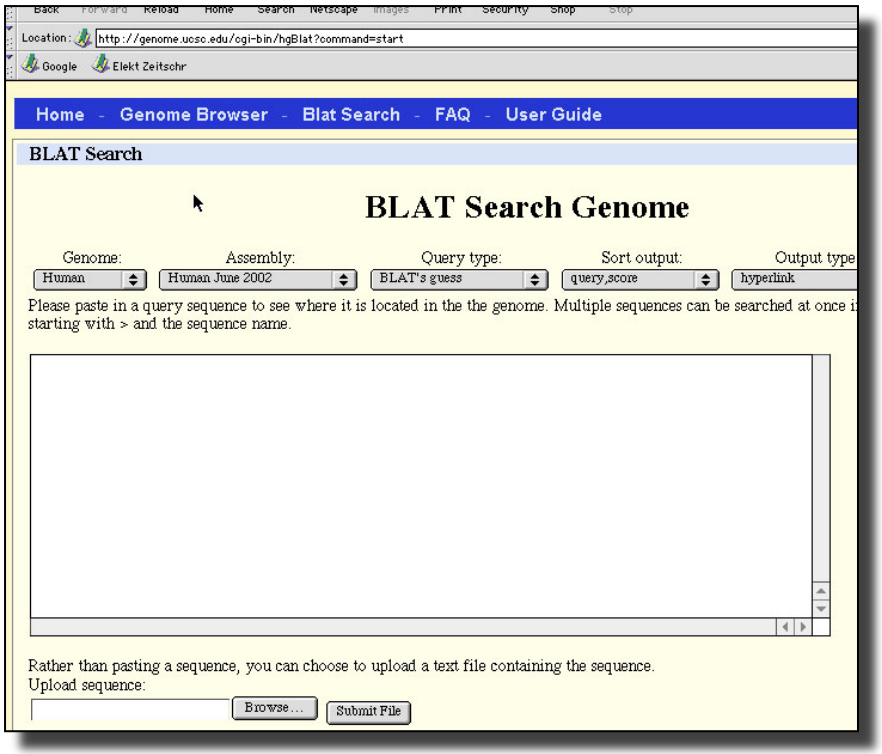


➢Muster zusammen mit Suchsequenz gegen DB laufen lassen

➢Treffer = Proteine mit Ähnlichkeit zur Suchsequenz und das Motiv enthaltend

Ultraschnelle DB-Suche über BLAT

„BLAST-like alignment tool“



http://genome.ucsc.edu/cgi-bin/hgBlat?command=start

- DNA-BLAT findet 40 Bp (>95% id) bzw. perfekte matches von >33Bp
- Protein-BLAT findet 20 aa (<80%id)
- Index (DNA) enthält alle nicht-überlappenden 11-mere des Genoms (1 Gb RAM)!!!
- Index wird gebraucht um passende Regionen im Genom schnell zu identifizieren, die dann für genaueren Vergleich „hochgeladen“ werden

WWW-Seiten mit „tools“