1. Ricerca di omologhe in banche dati.

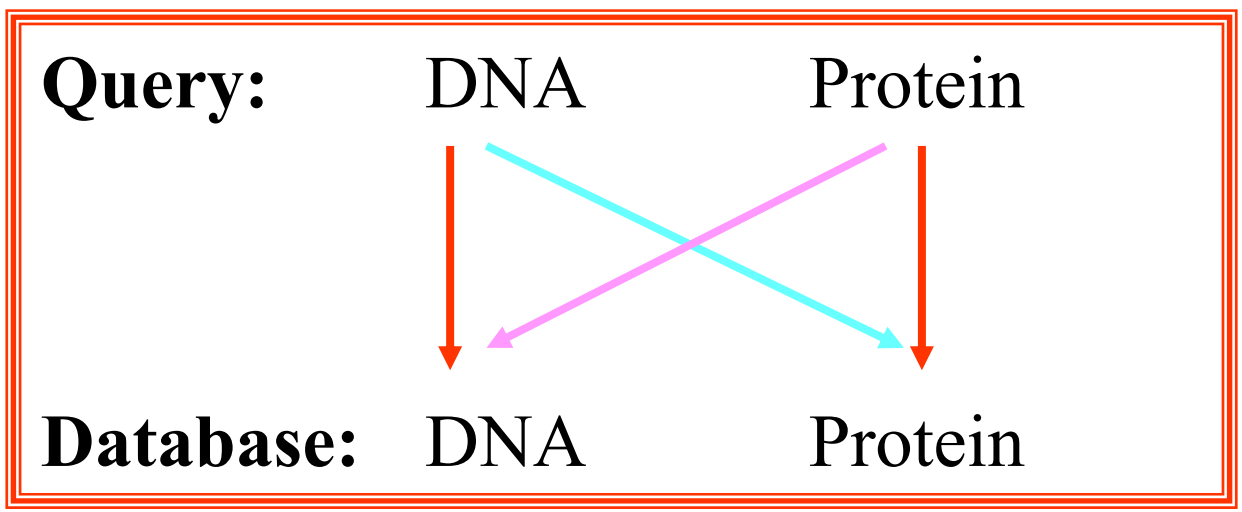
2. Programmi per la ricerca:

   FASTA

   BLAST

Ricerca di omologhe in banche dati

- Proteina *vs.* proteine
- Gene (traduzione in aa) *vs.* proteine
- Gene *vs.* geni
- Proteina *vs.* traduzione in aa di sequenze nucleotidiche (tutti i moduli)



Quando confrontiamo sequenze proteiche cerchiamo la migliore corrispondenza per 20 diversi amminoacidi

Quando confrontiamo sequenze nucleotidiche cerchiamo la migliore corrispondenza per sole 4 basi nucleotidiche

Quando confrontiamo sequenze proteiche cerchiamo la migliore corrispondenza per 20 diversi amminoacidi

Quando confrontiamo sequenze nucleotidiche cerchiamo la migliore corrispondenza per sole 4 basi nucleotidiche

La probabilita' di trovare una buona corrispondenza (allineamento con punteggio alto) per caso è più alta per le sequenze nucleotidiche che per quelle proteiche

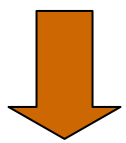
Inoltre, quando confrontiamo sequenze proteiche possiamo tener conto della similarita' tra i diversi amminoacidi

Quando confrontiamo sequenze proteiche cerchiamo la migliore corrispondenza per 20 diversi amminoacidi

Quando confrontiamo sequenze nucleotidiche cerchiamo la migliore corrispondenza per sole 4 basi nucleotidiche

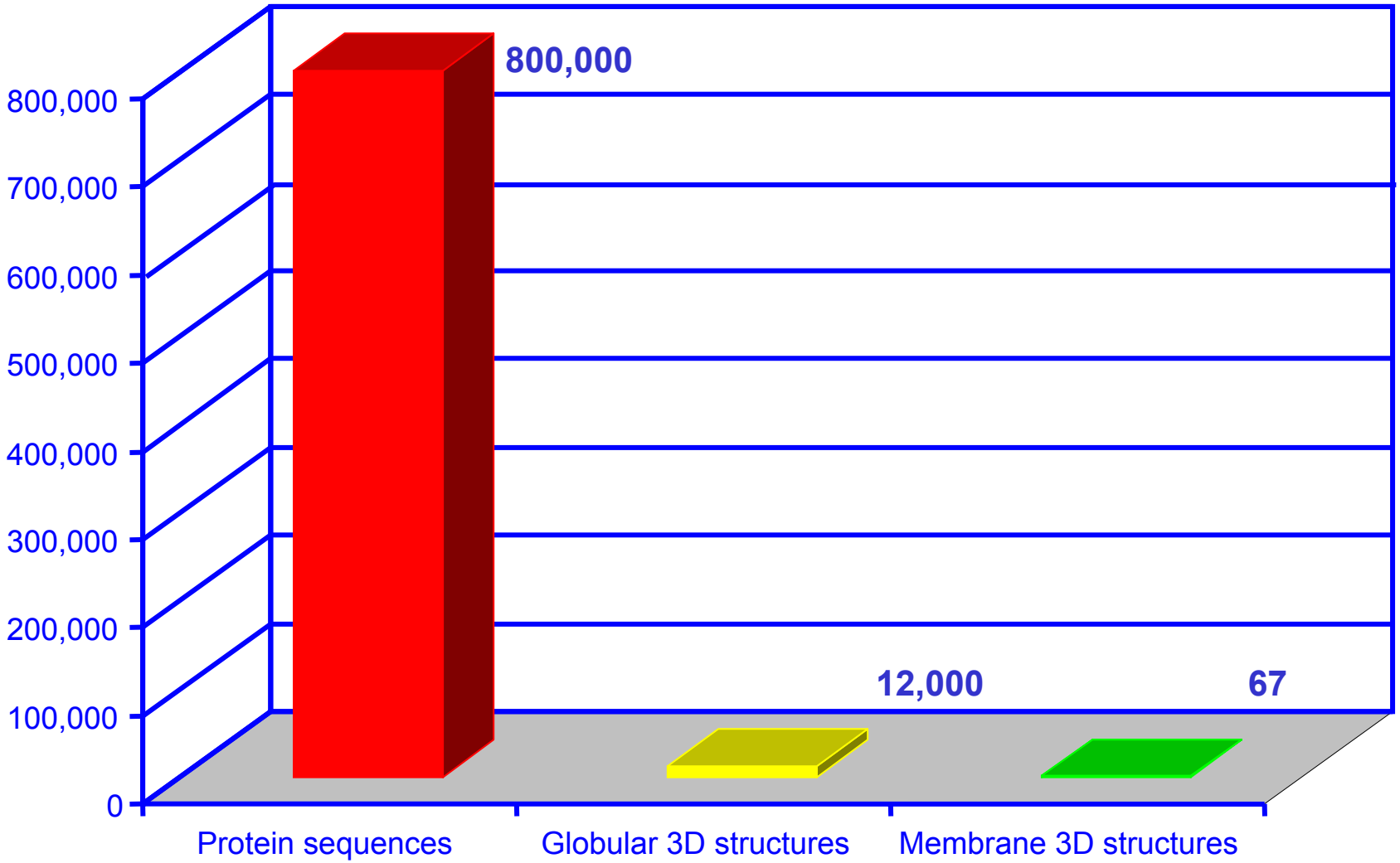
La probabilita' di trovare una buona corrispondenza (allineamento con punteggio alto) per caso è più alta per le sequenze nucleotidiche che per quelle proteiche

Inoltre, quando confrontiamo sequenze proteiche possiamo tener conto della similarita' tra i diversi amminoacidi



Quando è possibile, è preferibile confrontare sequenze proteiche !

Quante sequenze di proteine nelle banche dati ?



Come possiamo “pescare” dai databases di sequenze potenziali omologhe?

Algoritmi esatti (Smith-Waterman)
Lezione precedente

Esatto, garantisce di trovare il/i miglior allineamento/i per una coppia di sequenze.

Per 2 sequenze: A di lunghezza n and B di lunghezza m, Smith-Waterman inpiega n*m passi computazionali.

Cerchiamo omologhe della sequenza query A (n=200 aa)

Cerchiamo nel DB (10^6 sequenze di m=200 aa)

Numero di passi computazionali = 10^6 x 200 x 200 = ~10^{10}

10^3 passi al sec = 10^7 secs = 120 giorni = 4 mesi!

Necessità di algoritmi approssimati

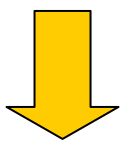
Algoritmi esatti (Smith-Waterman)
Lezione precedente

Esatto, garantisce di trovare il/i miglior allineamento/i per una coppia di sequenze.

Per 2 sequenze: A di lunghezza n and B di lunghezza m, Smith-Waterman inpiega n*m passi computazionali.

.

Come scartiamo gli allineamenti irrilevanti ?

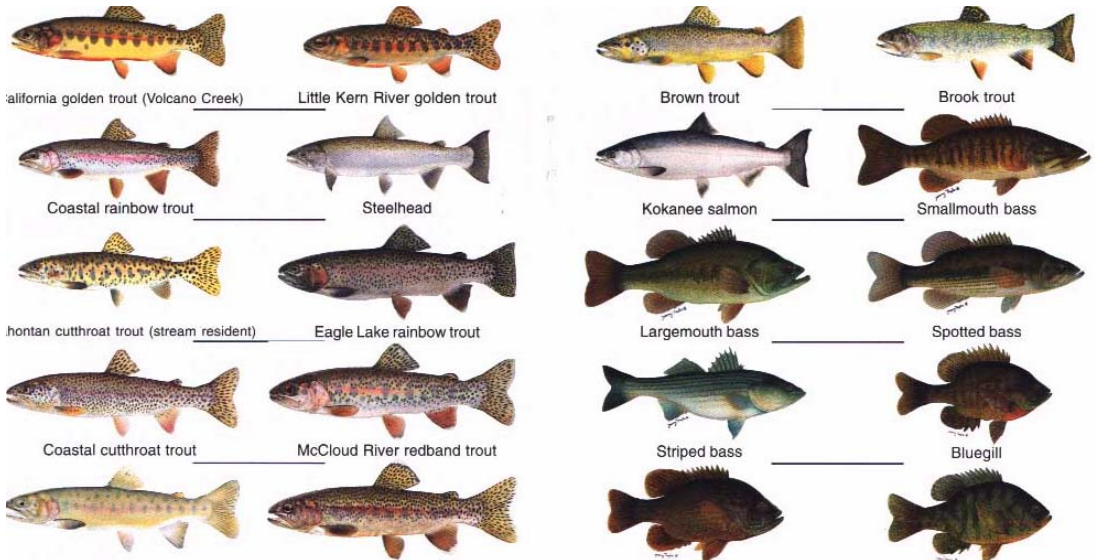


Gli algoritmi euristici (BLAST, FASTA) servono a scartare la gran parte degli allineamenti irrilevanti.

Programmi quali **FASTA** e **BLAST**, partendo da una sequenza query:

Programmi quali **FASTA** e **BLAST**, partendo da una sequenza query:

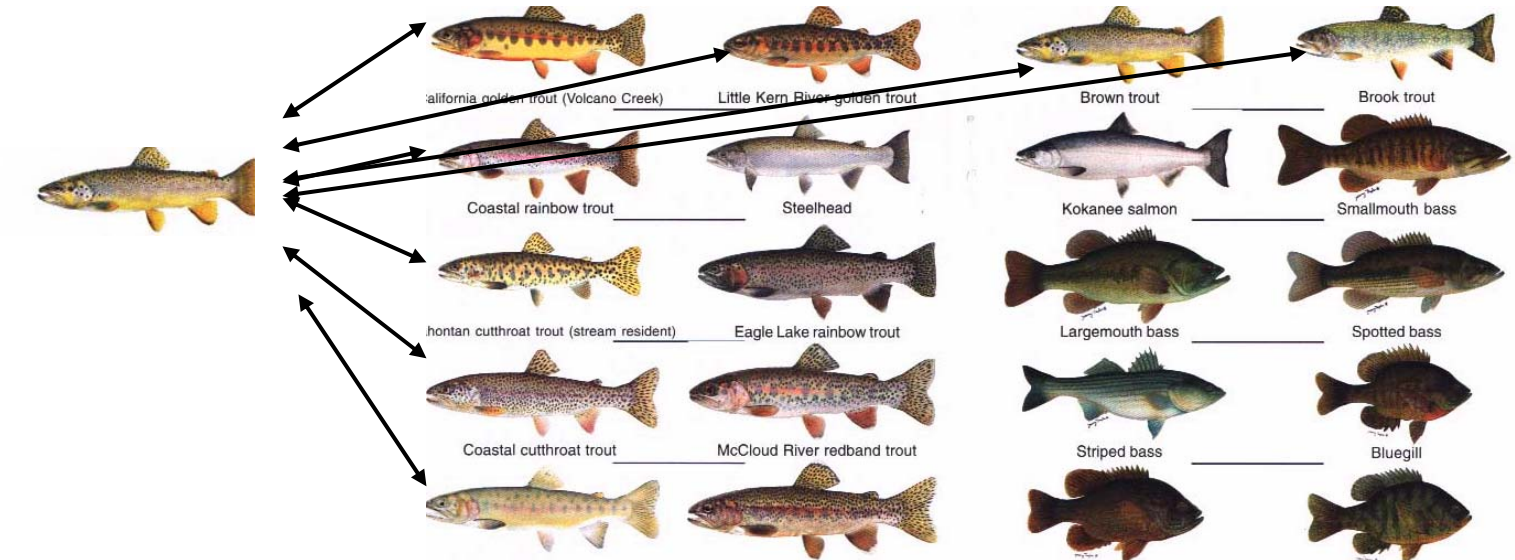
prima "pescano" dalle banche dati un sottoinsieme di sequenze che sono potenziali omologhe



Programmi quali **FASTA** e **BLAST**, partendo da una sequenza query:

prima "pescano" dalle banche dati un sottoinsieme di sequenze che sono potenziali omologhe

poi allineano al meglio ciascuna sequenza di questo sottoinsieme alla sequenza query



FASTA: esempio

query

ACDDEFGSATRMASTRK

Banca dati (DB)

Seq 1: LKDCDDAFSGSTLTLMRASRK

Seq 2: ACKRAEFSGSVTRMLSTRK

Seq 3: ACDDEFGLLLTRYTMASTRK

Step 1 **= Divisione della sequenza in parole di 2 caratteri.**

Parole possibili:

AC, CD, DD, DE, EF, FG, GS, SA, AT, TR, RM, MA, AS, ST, RK

Step 2 = Tabella delle frequenze delle parole

Query: ACDDEFGSATRMASTRK

DB: Seq 1 LKDCDDAFSGSTLTLMRASRK
Seq 2 ACKRAEFSGSVTRMLSTRK
Seq 3 ACDDEFGLLLTRYTMASTRK

Parola	Query	Seq 1	Seq 2	Seq 3	Off1	Off2	Off3
AC	1	-	1	1	-	0	0
CD	2	4	-	2	2	-	0
DD	3	5	-	3	2	-	0
DE	4	-	-	4	-	-	0
EF	5	-	6	5	-	1	0
FG	6	-	-	6	-	-	0
GS	7	10	9	-	3	**2**	-
SA	8	-	-	-	-	-	-
AT	9	-	-	-	-	-	-
TR	10	-	12, 17	11	-	**2**, 7	1
RM	11	-	13	-	-	**2**	-
MA	12	-	-	15	-	-	3
AS	13	18	-	16	5	-	3
ST	14	11	16	17	-3	**2**	3
RK	16	20	18	19	4	**2**	3

Step 3 = Calcolo dello *score* di similarita' *Init1* (basato sulla tabella dello step 2)

Query: `ACDDEFGSATRMASTRK`

DB: Seq 1 `LKDCDDAFSGSTLTLMRASRK`
Seq 2 `ACKRAEFSGSVTRMLSTRK`
Seq 3 `ACDDEFGLLLTRYTMASTRK`

Query		A	C	D	D	E	F	G	S	A	T	R	M	A	S	T	R	K	
Pos		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	
								X	X		X	X			X	X	X	X	
Seq2	A	C	K	R	A	E	F	S	G	S	V	T	R	M	L	S	T	R	K	T
Pos	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Off								2			2				2	2	2		

Init1(seq2) basato su questo allineamento approssimato usando una matrice tipo PAM250

Step 3 = Calcolo dello *score* di similarita' *Init1* (basato sulla tabella dello step 2)

Query: `ACDDEFGSATRMASTRK`

DB: Seq 1 `LKDCDDAFSGSTLTLMRASRK`
Seq 2 `ACKRAEFSGSVTRMLSTRK`
Seq 3 `ACDDEFGLLLTRYTMASTRK`

Query	A	C	D	D	E	F	G	S	A	T	R	M	A	S	T	R	K			
Pos	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17			
		X	X	X	X	X	X													
Seq2	A	C	D	D	E	F	G	L	L	L	T	R	Y	T	M	A	S	T	R	K
Pos	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Off		0	0	0	0	0	0													

Init1(seq3) basato su questo allineamento approssimato usando una matrice tipo PAM250

Step 3 = Calcolo dello *score* di similarita' *Init1* (basato sulla tabella dello step 2)

Query: `ACDDEFGSATRMASTRK`

DB: Seq 1 `LKDCDDAFSGSTLTLMRASRK`
Seq 2 `ACKRAEFSGSVTRMLSTRK`
Seq 3 `ACDDEFGLLLTRYTMASTRK`

Query	A	C	D	D	E	F	G	S	A	T	R	M	A	S	T	R	K
Pos	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
												X	X	X	X	X	X
Seq2	A	C	D	D	E	F	G	L	L	L	T	R	Y	T	M	A	S	T	R	K
Pos	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Off												3	3	3	3	3	3			

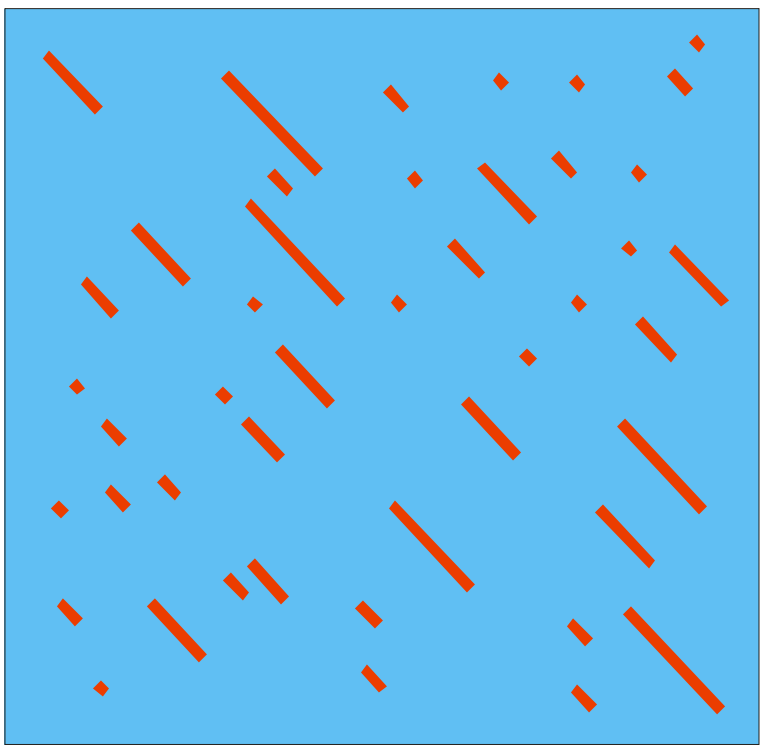
Init1(seq3) basato su allineamento sia con off 0 che off 1
(nel caso di proteine lunghe si prendono i migliori 10 offs)

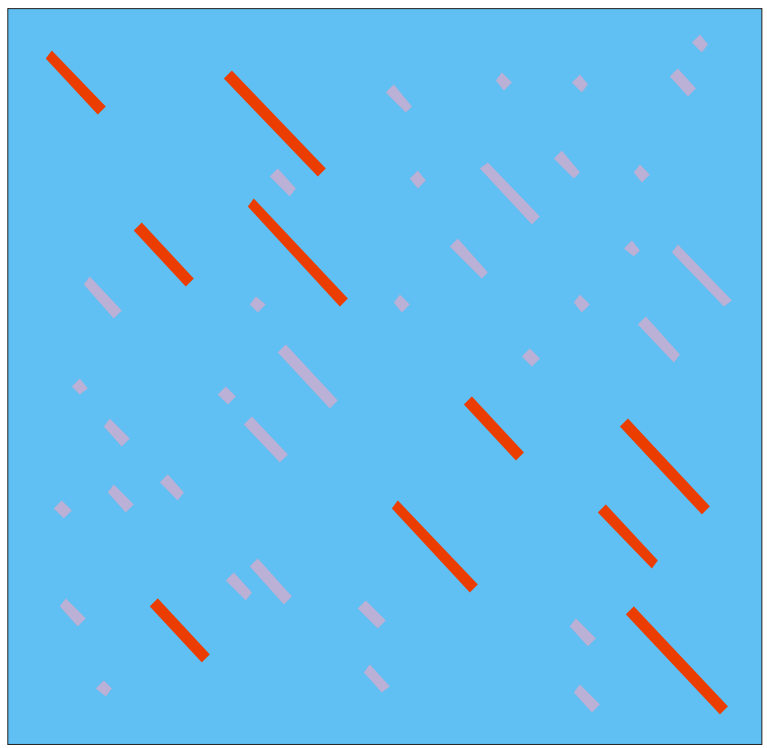
Sulla base degli score di similarita' *Init1* selezionare un sottoinsieme (grande) di sequenze per l'analisi successiva

Query: ACDDEFGSATRMASTRK

DB: Seq 1 LKDCDDAFSGSTLTLMRASRK
Seq 2 ACKRAEFSGSVTRMLSTRK
Seq 3 ACDDEFGLLLTRYTMASTRK

Per calcolare Init1 si scelgono solo le 10 regioni con il maggiore allineamento





Sulla base degli score di similarita' *Init1* selezionare un sottoinsieme (grande) di sequenze per l'analisi successiva

Query: `ACDDEFGSATRMASTRK`

DB: Seq 1 `LKDCDDAFSGSTLTLMRASRK`
Seq 2 `ACKRAEFSGSVTRMLSTRK`
Seq 3 `ACDDEFGLLLTRYTMASTRK`

In questo modo abbiamo selezionato omologhe senza permettere inserzioni o delezioni.

Introduzione dello score di similarità InitN.

Query: **ACDDEFGSATRMASTRK**

Tabella delle frequenze delle parole

Parola	Query	Seq 1	Seq 2	Seq 3	Off1	Off2	Off3
AC	1	-	1	1	-	**0**	0
CD	2	4	-	2	2	-	0
DD	3	5	-	3	2	-	0
DE	4	-	-	4	-	-	0
EF	5	-	6	5	-	**1**	0
FG	6	-	-	6	-	-	0
GS	7	10	9	-	3	**2**	-
SA	8	-	-	-	-	-	-
AT	9	-	-	-	-	-	-
TR	10	-	12, 17	11	-	**2**, 7	1
RM	11	-	13	-	-	**2**	-
MA	12	-	-	15	-	-	3
AS	13	18	-	16	5	-	3
ST	14	11	16	17	-3	**2**	3
RK	16	20	18	19	4	**2**	3

Step 4 = Calcolo dello *score* di similarita' *InitN* (basato sull'allineamento dello step 3 e sulla tabella dello step 2)

Query: ACDDEFGSATRMASTRK

DB: Seq 1 LKDCDDAFSGSTLTLMRASRK
Seq 2 ACKRAEFSGSVTRMLSTRK
Seq 3 ACDDEFGLLLTRYTMASTRK

Query	A	C	-	D	D	E	F	-	G	S	A	T	R	M	A	S	T	R	K	
Pos	1	2		3	4	5	6		7	8	9	10	11	12	13	14	15	16	17	
	X	X				X	X		X	X		X	X			X	X	X	X	
Seq2	A	C	K	R	A	E	F	S	G	S	V	T	R	M	L	S	T	R	K	T
Pos	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Off	0					1			2			2				2		2		

InitN(seq2) = ***Init1***(seq2) + score(nuovi matchs) – ***K***(gap)

Sulla base degli score di similarita' *InitN* selezionare un sottoinsieme ridotto di sequenze per l'analisi successiva

Query: ACDDEFGSATRMASTRK

DB: Seq 1 LKDCDDAFSGSTLTLMRASRK
Seq 2 ACKRAEFSGSVTRMLSTRK
Seq 3 ACDDEFGLLLTRYTMASTRK

Step 5 **= Allineamento vero e proprio delle sequenze con il miglior *InitN alla query* e calcolo del coefficiente finale *opt* (punteggio per il nuovo allineamento, completo)**

NOTA. **La scelta del sottoinsieme di sequenze nel DB con cui si effettua l'allineamento ottimale è basata sui punteggi approssimati Init1 e InitN**

Algoritmo di FASTA

1. Suddivide le sequenze in parole (2 aa)

2. Trova le parole nelle sequenze del DB e calcola l'offset

3. Calcola la similarita' delle dieci regioni con maggiori parole identiche per ciascuna sequenza del DB (init1).
   Prima esclusione di sequenze

4. Calcola la similarita' delle dieci regioni con maggiori parole identiche includendo gaps (initN)
   Seconda esclusione di sequenze

5. Allinea accuratamente le sequenze rimaste (opt)
   init1 $\rightarrow$ initN $\rightarrow$ opt

Quanto e’ buono un allineamento?

Quanto e' buono un allineamento?

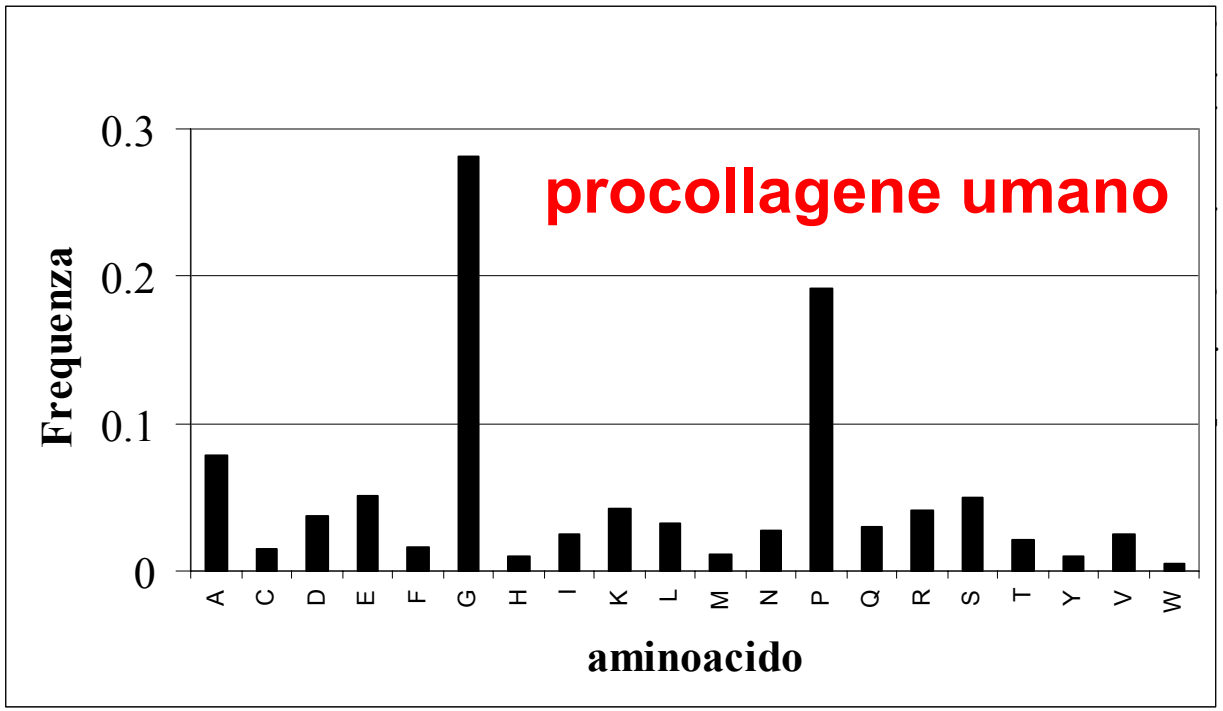
=

Quanto e' migliore di uno casuale?

- Sequenze che danno un allineamento casuale:
  - Sequenze non omologhe
  - Sequenze rimescolate ("shuffled")
  - Sequenze generate casualmente
  - Sequenze a bassa complessità

bassa complessità

MMSFVQKGSW LLLALLHPTI ILAQQEAVEG GCSHLGQSYA DRDVWKPEPC QICVCDSGSV LCDDIICDDQ
ELDCPNPEIP FGECCAVCPQ PPTAPTRPPN GQGPQGPKGD PGPPGIPGRN GDPGIPGQPG SPGSPGPPGI
CESCPTGPQN YSPQYDSYDV KSGVAVGGLA GYPGPAGPPG PPGPPGTSGH PGSPGSPGYQ GPPGEPGQAG
PSGPPGPPGA IGPSGPAGKD GESGRPGRPG ERGLPGPPGI KGPAGIPGFP GMKGHRGFDG RNGEKGETGA
PGLKGENGLP GENGAPGPMG PRGAPGERGR PGLPGAAGAR GNDGARGSDG QPGPPGPPGT AGFPGSPGAK
GEVGPAGSPG SNGAPGQRGE PGPQGHAGAQ GPPGPPGING SPGGKGEMGP AGIPGAPGLM GARGPPGPAG
ANGAPGLRGG AGEPGKNGAK GEPGPRGERG EAGIPGVPGA KGEDGKDGSP GEPGANGLPG AAGERGAPGF
RGPAGPNGIP GEKGPAGERG APGPAGPRGA AGEPGRDGVP GGPGMRGMPG SPGGPGSDGK PGPPGSQGES
GRPGPPGPSG PRGQPGVMGF PGPKGNDGAP GKNGERGGPG GPGPQGPPGK NGETGPQGPP GPTGPGGDKG
DTGPPGPQGL QGLPGTGGPP GENGKPGEPG PKGDAGAPGA PGGKGDAGAP GERGPPGLAG APGLRGGAGP
PGPEGGKGAA GPPGPPGAAG
GPAGQPGDKG EGGAPGLPGI
VAGPPGGSGP AGPPGPQGVK
TGAPGSPGVS GPKGDAGQPG
GKPGANGLSG ERGPPGPQGL
PPGPVGPAGK SGDRGESGPA
PGPAGQQGAI GSPGPAGPRG
GAPGPCCGGV GAAAIAGIGG
NCRDLKFCHP ELKSGEYWVD
SMDGGFQFSY GNPELPEDVL
KAEGNSKFTY TVLEDGCTKH



Si tratta di sequenze omologhe?

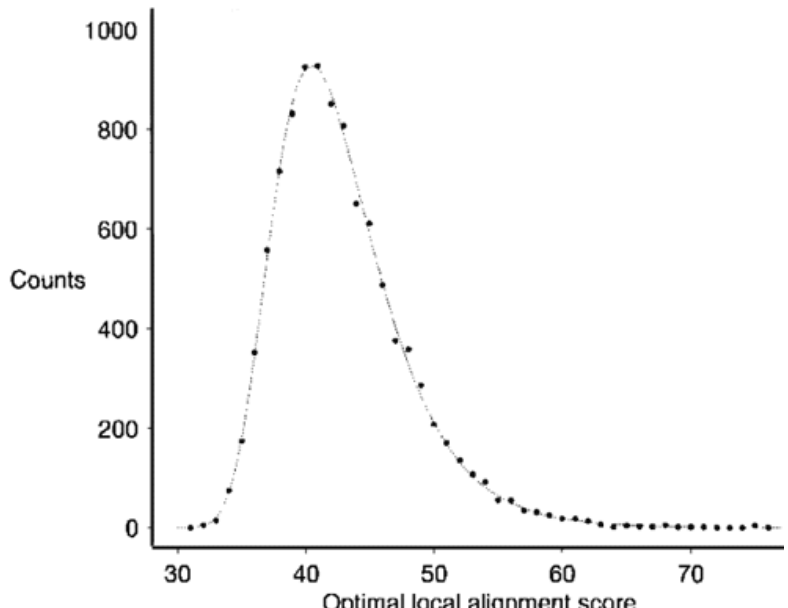
Valutazione della significatività dell'allineamento

a) Generazione di un ampio numero di sequenze casuali con la stessa composizione della seq query (sequenze "shuffled")

b) Ripetizione della ricerca di similarita' su sottoinsiemi casuali dei DB utilizzando come query ciascuna delle seq. casuali

c) Calcolo degli *opt* corrispondenti, del loro valore medio $M_{casuale}$ e della corrispondente deviazione standard $\sigma_{casuale}$

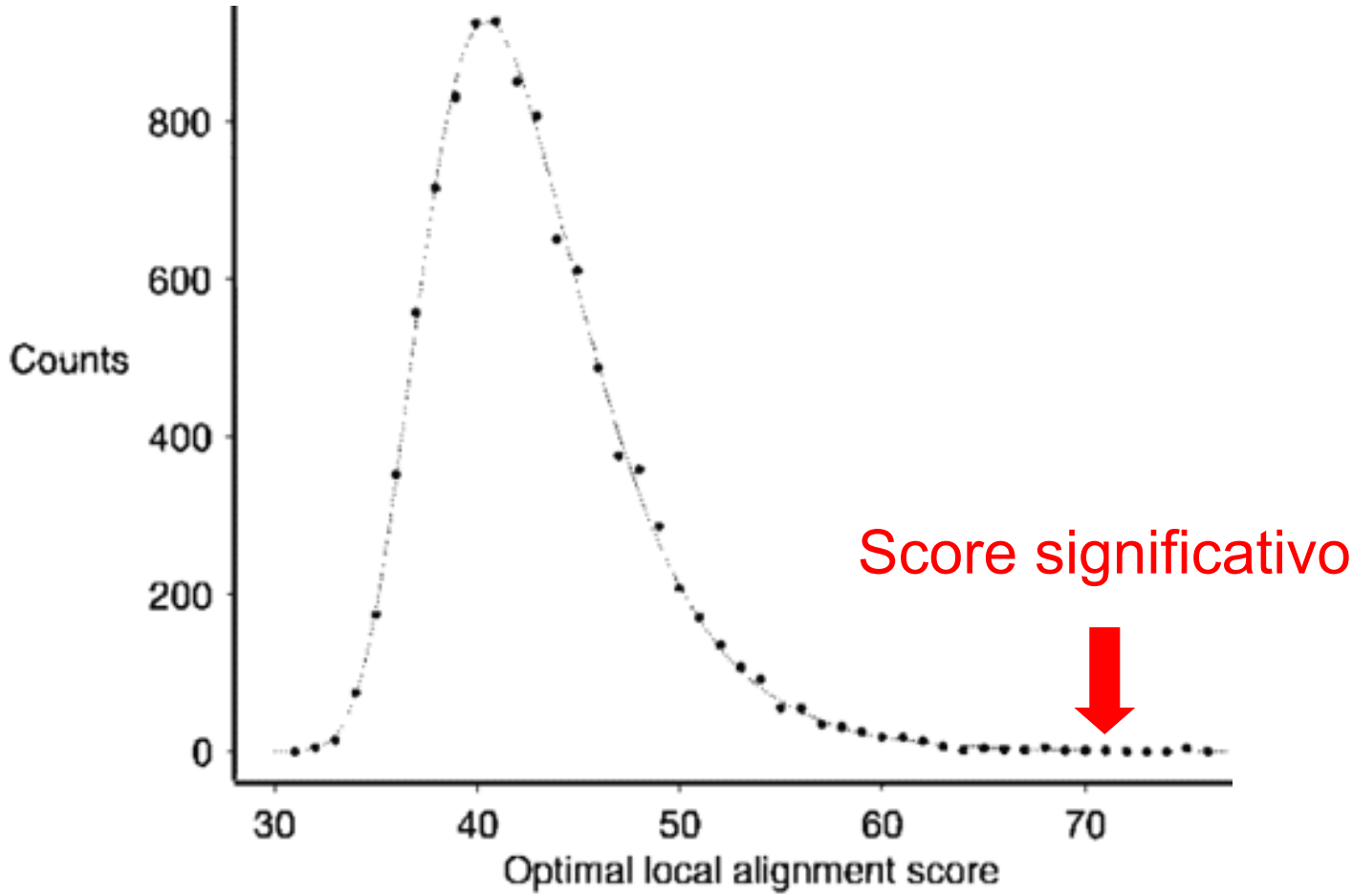
Distribuzione dei punteggi casuali



Si tratta di sequenze omologhe?

Valutazione della significatività dell'allineamento

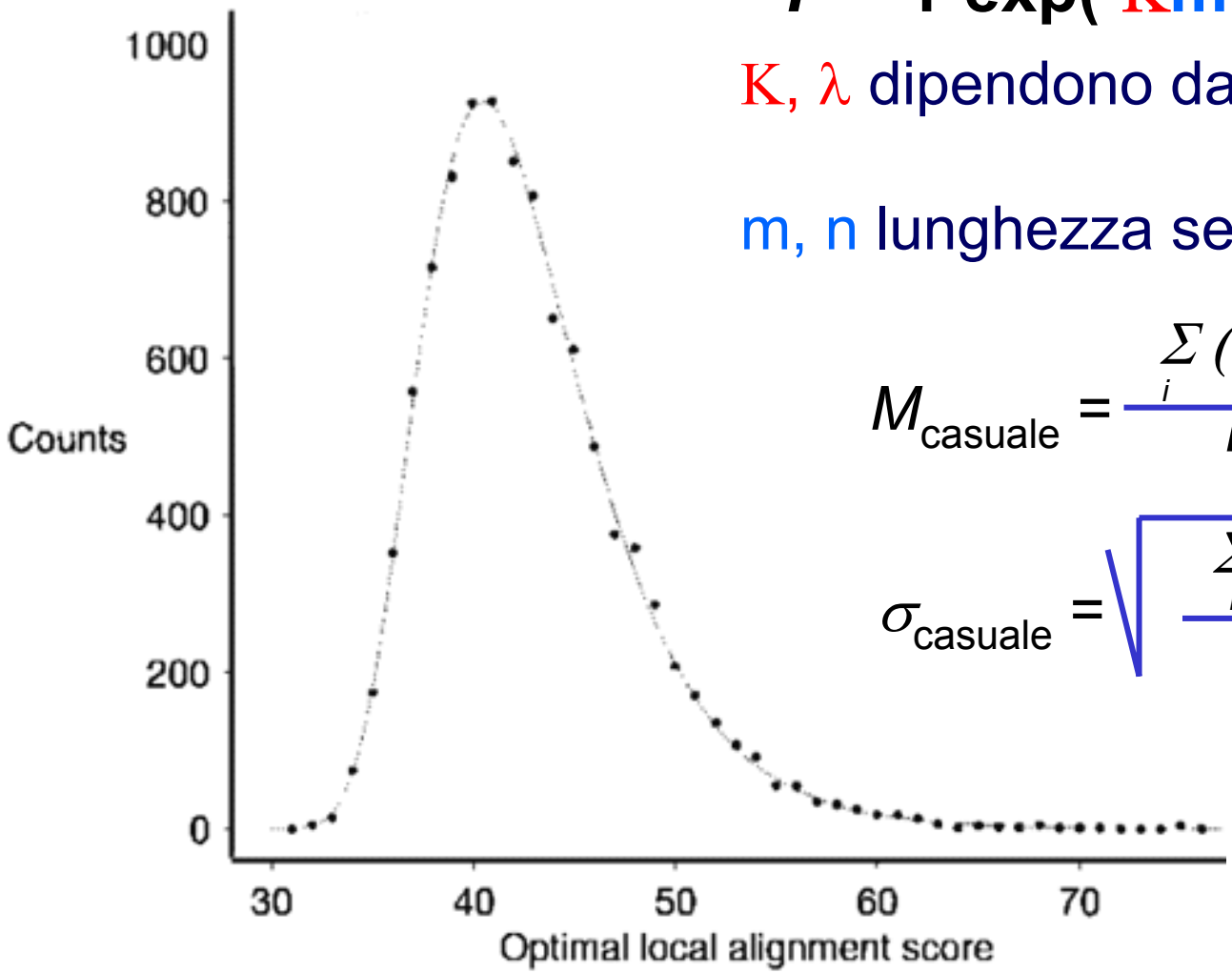
Due sequenze possono essere considerate omologhe se il punteggio per il loro allineamento ottimale (opt) cade fuori dalla distribuzione dei punteggi ottenuti per caso



Si tratta di sequenze omologhe?

Valutazione della significatività dell'allineamento

Score di sequenze casuali: distribuzione di Poisson



$$P = 1-\exp(-Kmn \exp(-\lambda S))$$

K, λ dipendono da matrice e DB

m, n lunghezza sequenze comparate

$$M_{\text{casuale}} = \frac{\sum_i (opt_i)}{n}$$

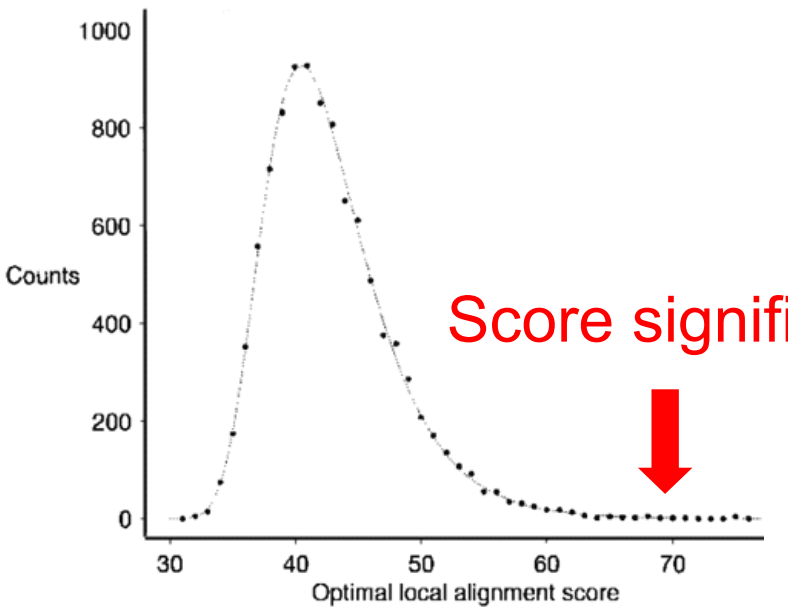
$$\sigma_{\text{casuale}} = \sqrt{\frac{\sum_i (opt_i - M_{\text{casuale}})}{n}}$$

Si tratta di sequenze omologhe?
Valutazione della significatività dell'allineamento

Calcolo dello **Z-score** e dell' **E-value** per l'allineamento della sequenza query con le sue possibili omologhe

Z-score = numero di deviazioni standard che separano lo *punteggio (opt)* della *query* della media dei punteggi casuali

$$\text{Z-score} = (opt_{\text{query}} - M_{\text{casuale}}) / \sigma_{\text{casuale}}$$



Score significativo

Z-score ≥ 4

➔ opt_{query} fuori dalla distribuzione casuale.

Si tratta di sequenze omologhe?
Valutazione della significatività dell'allineamento

E-value = *expectation value*: numero atteso di sequenze che danno per caso il punteggio S (o opt)
Indica quanto e' probabile che si trovi il punteggio S per caso in una distribuzione di Poisson con valore medio $M_{casuale}$

$$\textit{E-value} = Kmn \exp(-\lambda S) = mn\, 2^{-S'}$$

Bit score : $S' = (\lambda S - \ln K) / \ln 2$

(Ricordarsi che $2^{1/\ln 2} = e$)

E' possibile confrontare direttamente i bit score per ricerche con diverse matrici e diverse banche dati

Si tratta di sequenze omologhe?

Valutazione della significatività dell'allineamento

Opt	grande
Z-score	> 4
E-value	> 0.01
Bit-score	grande

- Help Index
- General Help
- Formats
- Gaps
- Matrix
- References
- Fasta Help
- MView Help
- VisualFasta Help

- View all Fasta's at EBI
- Fasta Programmatic Access

- Database Information
  - UniProt
  - UniParc

Fasta Protein Database Query

Provides sequence similarity searching against nucleotide and protein databases using the Fasta programs. Fasta can be very specific when identifying long regions of low similarity especially for highly diverged sequences. You can also conduct sequence similarity searching against complete proteome or genome databases using the Fasta programs.

Download Software

YOUR EMAIL	SEARCH TITLE	RESULTS	PROGRAM	DATABASES
	Sequence	interactive	fasta3 fastx3 fasty3	Protein UniProt UniRef100 UniRef90
GAP PENALTIES	**SCORES & ALIGNMENTS**	**KTUP/ HISTOGRAM**	**DNA STRAND**	**MATRIX**
OPEN -10 RESIDUE -2	SCORES 50 ALIGN 50	KTUP 2 HIST no	none	BLOSUM50
EXPECTATION UPPER VALUE	**EXPECTATION LOWER VALUE**	**SEQUENCE RANGE**	**DATABASE RANGE**	**MOLECULE TYPE**
10.0	default	START-END	START-END	Protein

Enter or Paste a PROTEIN Sequence in any format: Help

```
>gi|4504111|ref|NP_002077.1| growth factor
receptor-bound protein 2 isoform 1 [Homo
sapiens]
MEAIAKYDFKATADDELSFKRGDILKVLNEECDQNWYKAELNGKDGF
IPKNYIEMKPHPWFFGKIPRAKA
EEMLSKQRHDGAFLIRESESAPGDFSLSVKFGNDVQHFKVLRDGAGK
YFLWVVKFNSLNELVDYHRSTSV
```

- Help
- General Help
- Formats
- Gaps
- Matrix
- References
- Fasta Help
- MView Help
- VisualFasta Help

- Database Information
  - UniProt
  - UniParc

Fasta Summary Table

SUBMISSION PARAMETERS			
Title	Sequence	**Database**	uniprot
Sequence length	217	**Sequence type**	p
Program	fasta	**Version**	3.4t25 Sept 2, 2005
Expectation upper value	10.0	**Matrix**	BL50
Sequence range	1-	**Number of scores**	50
Number of alignments	50	**Word size**	2
Open gap penalty	-10	**Gap extension penalty**	-2
Histogram	false		

Show Annotation | Fasta Result | MView | VisualFasta | XML | SUBMIT ANOTHER JOB

Show Alignments | Clear all | Check all | Invert selection | Reset

Alignment	DB:ID	Source	Length	Identity%	Similar%	Overlap	E()
1 ☑	UNIPROT:Q2PG25_MACFA	Hypothetical protein.	217	100.000	100.000	217	2.1e-97
2 ☑	UNIPROT:Q5BKA7_RAT	Growth factor receptor bound	217	100.000	100.000	217	2.1e-97
3 ☑	UNIPROT:GRB2_HUMAN	Growth factor receptor-bound	217	100.000	100.000	217	2.1e-97
4 ☑	UNIPROT:Q3T0F9_BOVIN	Hypothetical protein MGC12	217	100.000	100.000	217	2.1e-97
5 ☑	UNIPROT:GRB2_PONPY	Growth factor receptor-bound	217	100.000	100.000	217	2.1e-97
6 ☑	UNIPROT:GRB2_RAT	Growth factor receptor-bound p	217	100.000	100.000	217	2.1e-97
7 ☑	UNIPROT:GRB2_MOUSE	Growth factor receptor-bound	217	99.539	100.000	217	4e-97

```
>>UNIPROT:DRK_DROME Q08012 Protein E(sev)2B (Protein enh  (211 aa)
 initn: 923 init1: 364 opt: 942  Z-score: 1173.2  bits: 224.1 E(): 4.5e-57
Smith-Waterman score: 942;  65.566% identity (85.377% similar) in 212 aa overlap (1-212:1-208)

               10        20        30        40        50        60
Sequen MEAIAKYDFKATADDELSFKRGDILKVLNEECDQNWYKAELNGKDGFIPKNYIEMKPHPW
       :::::::::::::::::: .::.:: : :::::::::::::::::::: : :
UNIPRO MEAIAKHDFSATADDELSFRKTQILKILNMEDDSNWYRAELDGKEGLIPSNYIEMKNHDW
               10        20        30        40        50        60

               70        80        90       100       110       120
Sequen FFGKIPRAKAEEMLSKQRHDGAFLIRESESAPGDFSLSVKFGNDVQHFKVLRDGAGKYFL
       .:::: :: :::::: .::::::: ::::::::::::  . :::::::: .:::: 
UNIPRO YYGRITRADAEKLLSN-KHEGAFLIRISESSPGDFSLSVKCPDGVQHFKVLRDAQSKFFL
               70         80        90       100       110

              130       140       150       160       170       180
Sequen WVVKFNSLNELVDYHRSTSVSRNQQIFLRDIEQVPQQPTYVQALFDFDPQEDGELGFRRG
       :::::::::::::::::::::::::: ::::  ::::   ::::::: ::::::: ::::
UNIPRO WVVKFNSLNELVEYHRTASVSRSQDVKLRDM--IPEE-MLVQALYDFVPQESGELDFRRG
     120       130       140       150           160       170

              190       200       210
Sequen DFIHVMDNSDPNWWKGACHGQTGMFPRNYVTPVNRNV
       : : : :: :::::   .. :::: .::::
UNIPRO DVITVTDRSDENWWNGEIGNRKGIFPATYVTPYHS
        180       190       200       210
```

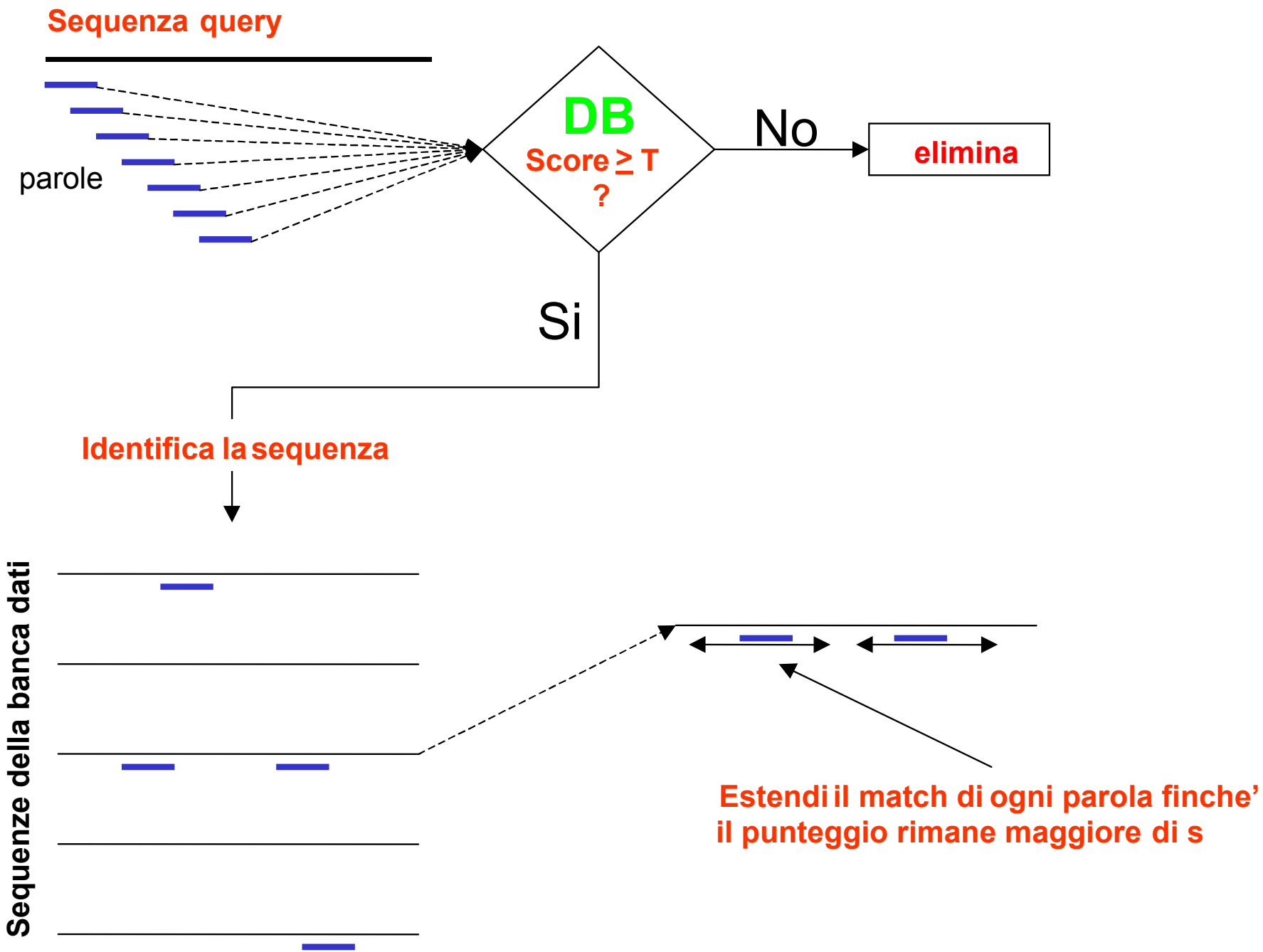
```
>>UNIPROT:Q7PV64_ANOGA Q7PV64 ENSANGP00000020137.         (211 aa)
 initn: 915 init1: 367 opt: 935  Z-score: 1164.5  bits: 222.5 E(): 1.4e-56
Smith-Waterman score: 935;  64.623% identity (85.377% similar) in 212 aa overlap (1-212:1-208)

               10        20        30        40        50        60
Sequen MEAIAKYDFKATADDELSFKRGDILKVLNEECDQNWYKAELNGKDGFIPKNYIEMKPHPW
       :::::::::::::::::::::::::: : :::::::::::::::::::: : :
UNIPRO MEAVAKHDFNATADDELSFRKSQVLKILNMEDDMNWYRAELDGKEGLIPSNYIEMKNHDW
               10        20        30        40        50        60

               70        80        90       100       110       120
Sequen FFGKIPRAKAEEMLSKQRHDGAFLIRESESAPGDFSLSVKFGNDVQHFKVLRDGAGKYFL
       .:::: :: :::::: .::::::: ::::::::::::  .. :::::::: .:::: 
UNIPRO YYGRITRADAEKLLSN-KHEGAFLIRISESSPGDFSLSVKCSDGVQHFKVLRDAQGKFFL
               70         80        90       100       110
```

BLAST
(Basic Local Alignment Search Tool)

1. Suddivide la sequenza query in parole (*default*, 3 aa)
2. Seleziona parole molto simili mediante matrici di similarità e quindi per ogni tripletta della query genera una famiglia di triplette
3. Mediante matrici di similarità confronta ogni parola con regioni di uguali dimensioni delle entries del DB e ne calcola lo *score*
4. Se lo *score* e' $\geq$ di un valore soglia ***T*** al di sotto del quale la similarita' e' considerata troppo bassa, estende la regione allineata cercando regioni di alta similarità (score sopra un secondo valore di soglia ***S***), fermandosi quando lo score non può più essere migliorato


Sequenza query
parole
DB
Score ≥ T
?
No
elimina
Si
Identifica la sequenza
Sequenze della banca dati
Estendi il match di ogni parola finche'
il punteggio rimane maggiore di s

BLAST Step 1

- Data una **parola di lunghezza *w*** (di solito 3 per le proteine) e una matrice di scoring (es. BLOSUM62):

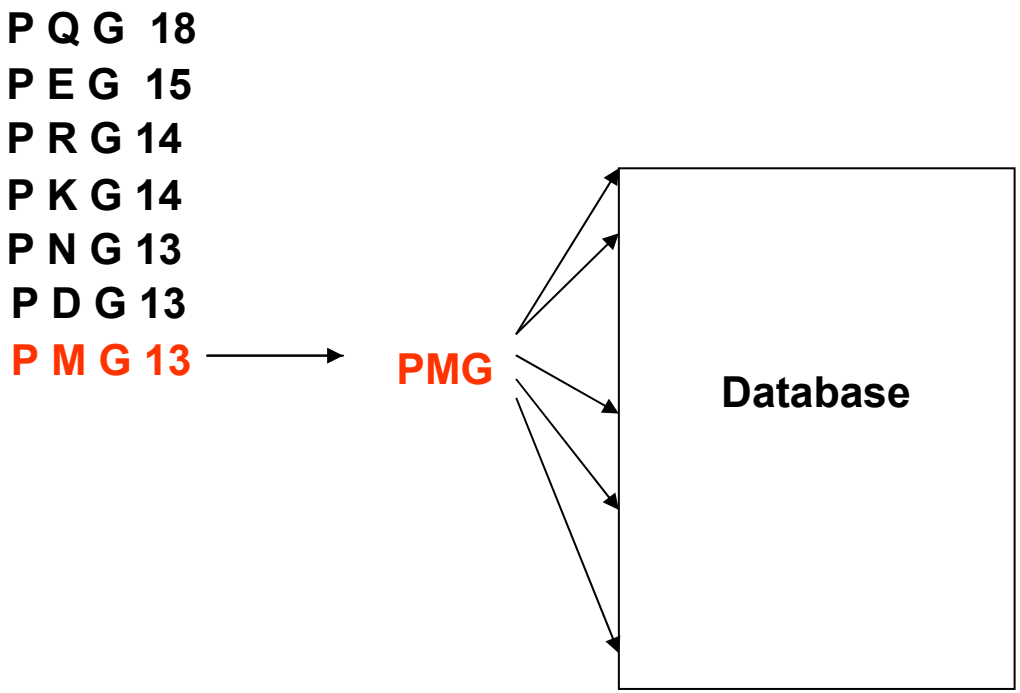
  Crea una lista di tutte le parole (*w*-lettere) cha danno uno **score >*T*** se confrontate con le parole di lunghezza *w*- della query.

Sequenza query **L N K C K T P Q G Q R L V N Q**

	P Q G 18	**Parola della query**
	P E G 15	**Parole simili**
	P R G 14	
	P K G 14	
	P N G 13	
	P D G 13	
	P M G 13	
Sotto Soglia (T=13)	**P Q A 12**	
	P Q N 12	
	etc.	

BLAST Step 2

- Identifica tutte le **posizioni nel database** dove si trova una parola sufficientemente simile (hit list).



BLAST Step 3

- Il programma prova a **estendere** l'allineamento in entrambe le direzioni aggiungendo coppie di residui. I residui sono aggiunti fino a che lo score non è più migliorabile. Considera solo gli allineamenti con score sopra il valore di soglia (S).

High Scoring Segment Pairs
(Tratto di allineamento ad alto punteggio)

BLAST e **FASTA** differiscono per il modo in cui “pescano” le potenziali omologhe nel DB (similarità/identità).

BLAST e **FASTA** differiscono per il modo in cui "pescano" le potenziali omologhe nel DB (similarità/identità).

Altra differenza fondamentale tra **BLAST** e **FASTA** sta nelle modalita' di calcolo della distribuzione dei punteggi casuali:

FASTA la calcola ogni volta che gli viene sottomessa una nuova query per la ricerca su un certo DB

BLAST usa distribuzioni precalcolate su ciascun DB per insiemi di sequenze casuali di composizione standard

$$E\text{-}value = Kmn\, e^{-\lambda S}$$

"**n**" in **FASTA** = lunghezza della sequenza allineata

"**n**" in **BLAST** = lunghezza totale delle sequenze nel DB usato ➔ a prescindere dalla *query*

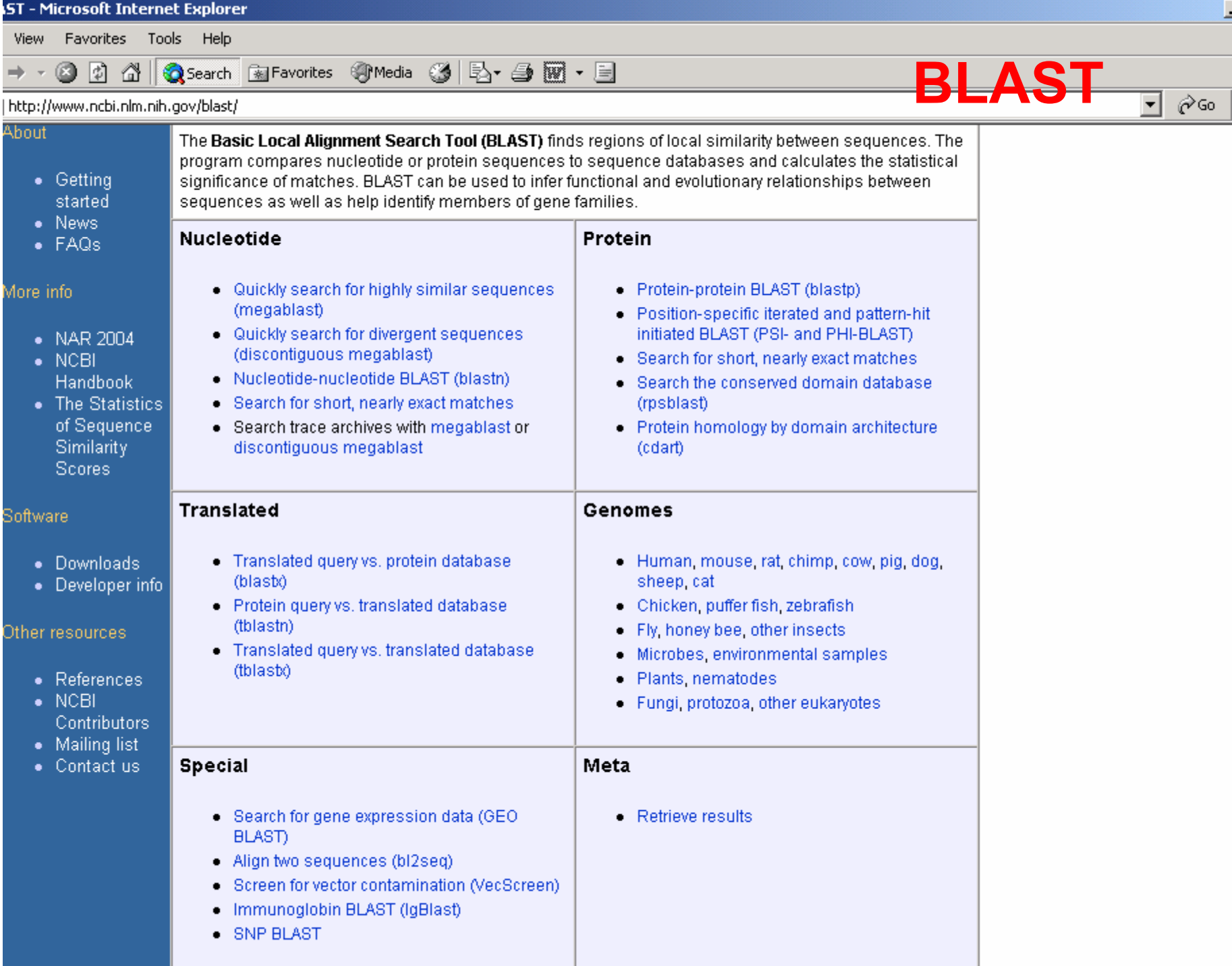
BLAST e **FASTA** differiscono per il modo in cui “pescano” le potenziali omologhe nel DB (similarità/identità).

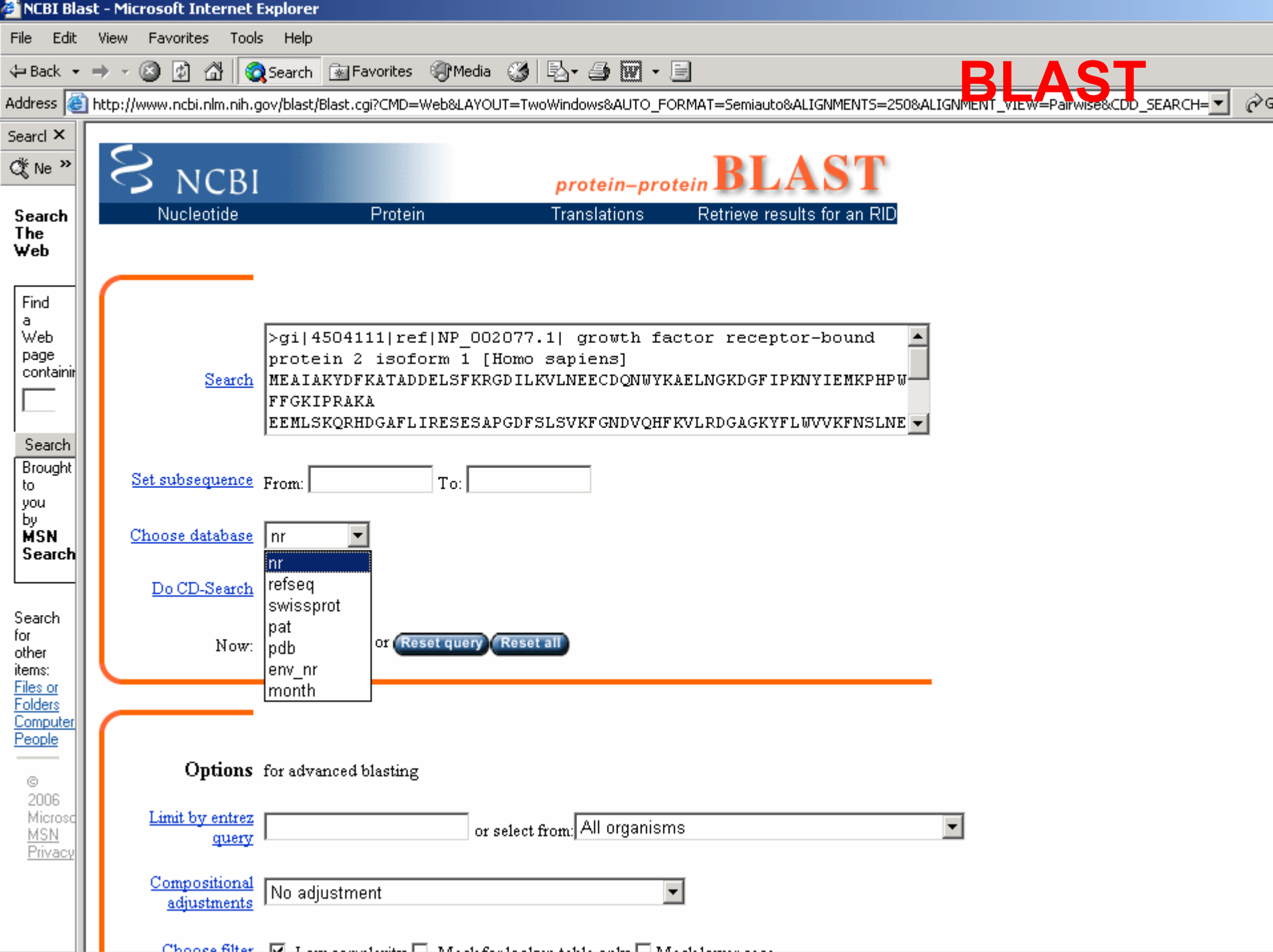
Inoltre una differenza fondamentale tra **BLAST** e **FASTA** sta nelle modalita’ di calcolo della distribuzione dei punteggi casuali:

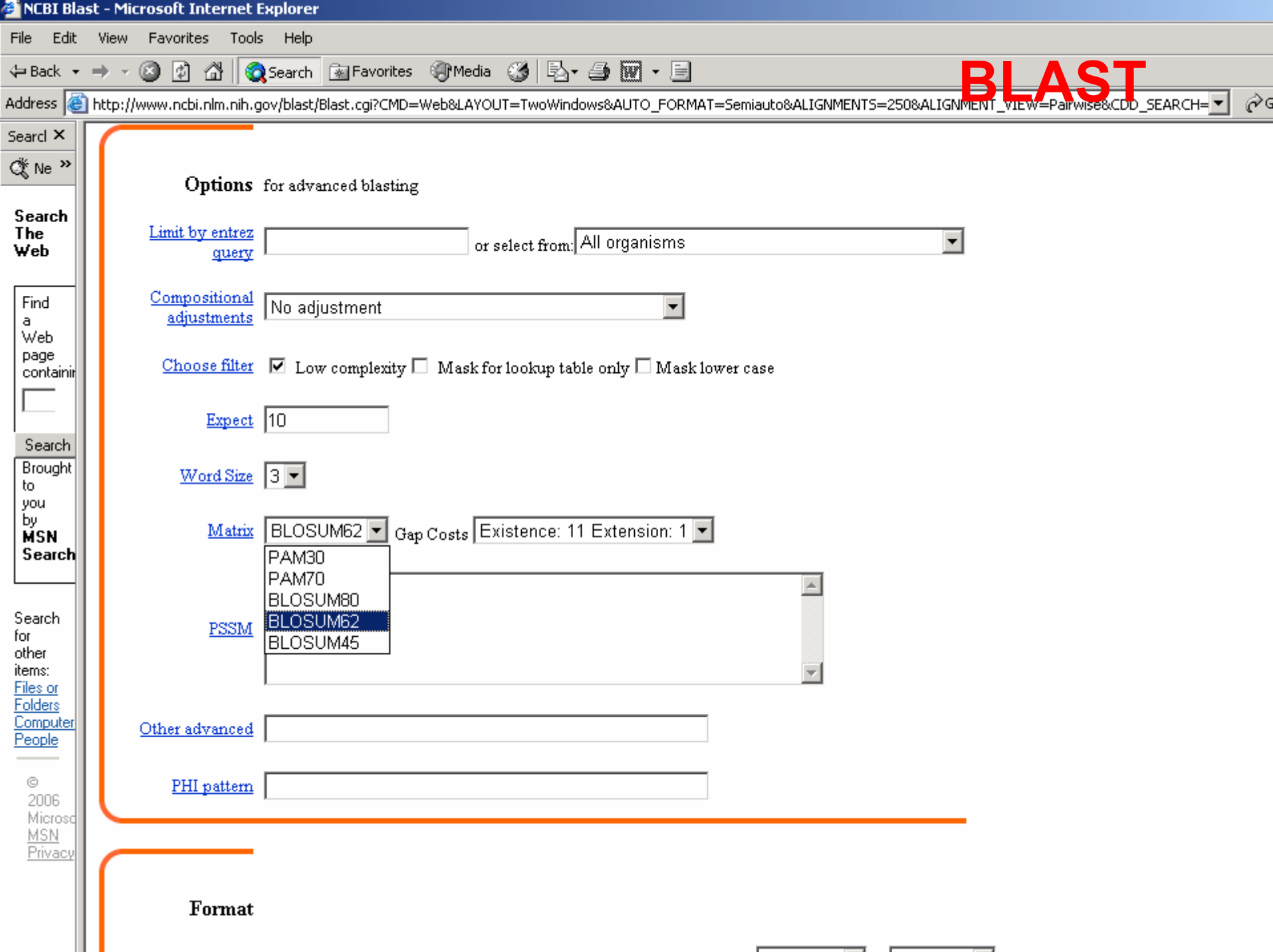
FASTA la calcola ogni volta che gli viene sottomessa una nuova query per la ricerca su un certo DB

BLAST usa distribuzioni precalcolate su ciascun DB per insiemi di sequenze casuali di composizione standard

per questo BLAST “maschera” le regioni di sequenza della query a bassa complessita’


ST - Microsoft Internet Explorer
View Favorites Tools Help
Search Favorites Media
BLAST
http://www.ncbi.nlm.nih.gov/blast/
Go
About
Getting started
News
FAQs
More info
NAR 2004
NCBI Handbook
The Statistics of Sequence Similarity Scores
Software
Downloads
Developer info
Other resources
References
NCBI Contributors
Mailing list
Contact us
The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.
Nucleotide
Quickly search for highly similar sequences (megablast)
Quickly search for divergent sequences (discontiguous megablast)
Nucleotide-nucleotide BLAST (blastn)
Search for short, nearly exact matches
Search trace archives with megablast or discontiguous megablast
Protein
Protein-protein BLAST (blastp)
Position-specific iterated and pattern-hit initiated BLAST (PSI- and PHI-BLAST)
Search for short, nearly exact matches
Search the conserved domain database (rpsblast)
Protein homology by domain architecture (cdart)
Translated
Translated query vs. protein database (blastx)
Protein query vs. translated database (tblastn)
Translated query vs. translated database (tblastx)
Genomes
Human, mouse, rat, chimp, cow, pig, dog, sheep, cat
Chicken, puffer fish, zebrafish
Fly, honey bee, other insects
Microbes, environmental samples
Plants, nematodes
Fungi, protozoa, other eukaryotes
Special
Search for gene expression data (GEO BLAST)
Align two sequences (bl2seq)
Screen for vector contamination (VecScreen)
Immunoglobin BLAST (IgBlast)
SNP BLAST
Meta
Retrieve results


NCBI Blast - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
BLAST
Address http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?CMD=Web&LAYOUT=TwoWindows&AUTO_FORMAT=Semiauto&ALIGNMENTS=250&ALIGNMENT_VIEW=Pairwise&CDD_SEARCH=
Search
Search The Web
Find a Web page containi
Search
Brought to you by MSN Search
Search for other items: Files or Folders Computer People
© 2006 Microsc MSN Privacy
NCBI
protein–protein BLAST
Nucleotide Protein Translations Retrieve results for an RID
Search
>gi|4504111|ref|NP_002077.1| growth factor receptor-bound protein 2 isoform 1 [Homo sapiens]
MEAIAKYDFKATADDELSFKRGDILKVLNEECDQNWYKAELNGKDGFIPKNYIEMKPHPW
FFGKIPRAKA
EEMLSKQRHDGAFLIRESESAPGDFSLSVKFGNDVQHFKVLRDGAGKYFLWVVKFNSLNE
Set subsequence From: To:
Choose database nr
nr
refseq
swissprot
pat
pdb
env_nr
month
Do CD-Search
Now: or Reset query Reset all
Options for advanced blasting
Limit by entrez query or select from: All organisms
Compositional adjustments No adjustment
Choose filter


NCBI Blast - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
BLAST
Address http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?CMD=Web&LAYOUT=TwoWindows&AUTO_FORMAT=Semiauto&ALIGNMENTS=250&ALIGNMENT_VIEW=Pairwise&CDD_SEARCH=
Search The Web
Find a Web page containi
Search
Brought to you by MSN Search
Search for other items: Files or Folders Computer People
© 2006 Microsc MSN Privacy
Options for advanced blasting
Limit by entrez query
or select from: All organisms
Compositional adjustments
No adjustment
Choose filter
Low complexity
Mask for lookup table only
Mask lower case
Expect
10
Word Size
3
Matrix
BLOSUM62
Gap Costs
Existence: 11 Extension: 1
PAM30
PAM70
BLOSUM80
BLOSUM62
BLOSUM45
PSSM
Other advanced
PHI pattern
Format

Edit View Favorites Tools Help

Back Search Favorites Media

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

Query = gi|4504111|ref|NP_002077.1| growth factor receptor-bound protein 2 isoform 1 [Homo sapiens] (217 letters)

Putative conserved domains have been detected, click on the image below for detailed results.

The request ID is 1142860123-17210-152380642238.BLASTQ4

Format! or Reset all

The results are estimated to be ready in 9 seconds but may be done sooner.

Please press "FORMAT!" when you wish to check your results. You may change the formatting options for your result via the form below and press "FORMAT!" again. You may also request results of a different search by entering any other valid request ID to see other recent jobs.

Format

Show ☑ Graphical Overview ☑ Linkout ☑ Sequence Retrieval ☑ NCBI-gi Alignment in HTML format

☐ CDS feature

Masking Character Lower Case Masking Color Grey

Number of: Descriptions 500 Alignments 250

Alignment view Pairwise

I tratti di sequenza "mascherati" perche' a bassa complessita' sono riportati in lettere minuscole colorate in grigio

Internet

Edit View Favorites Tools Help

Search Favorites Media

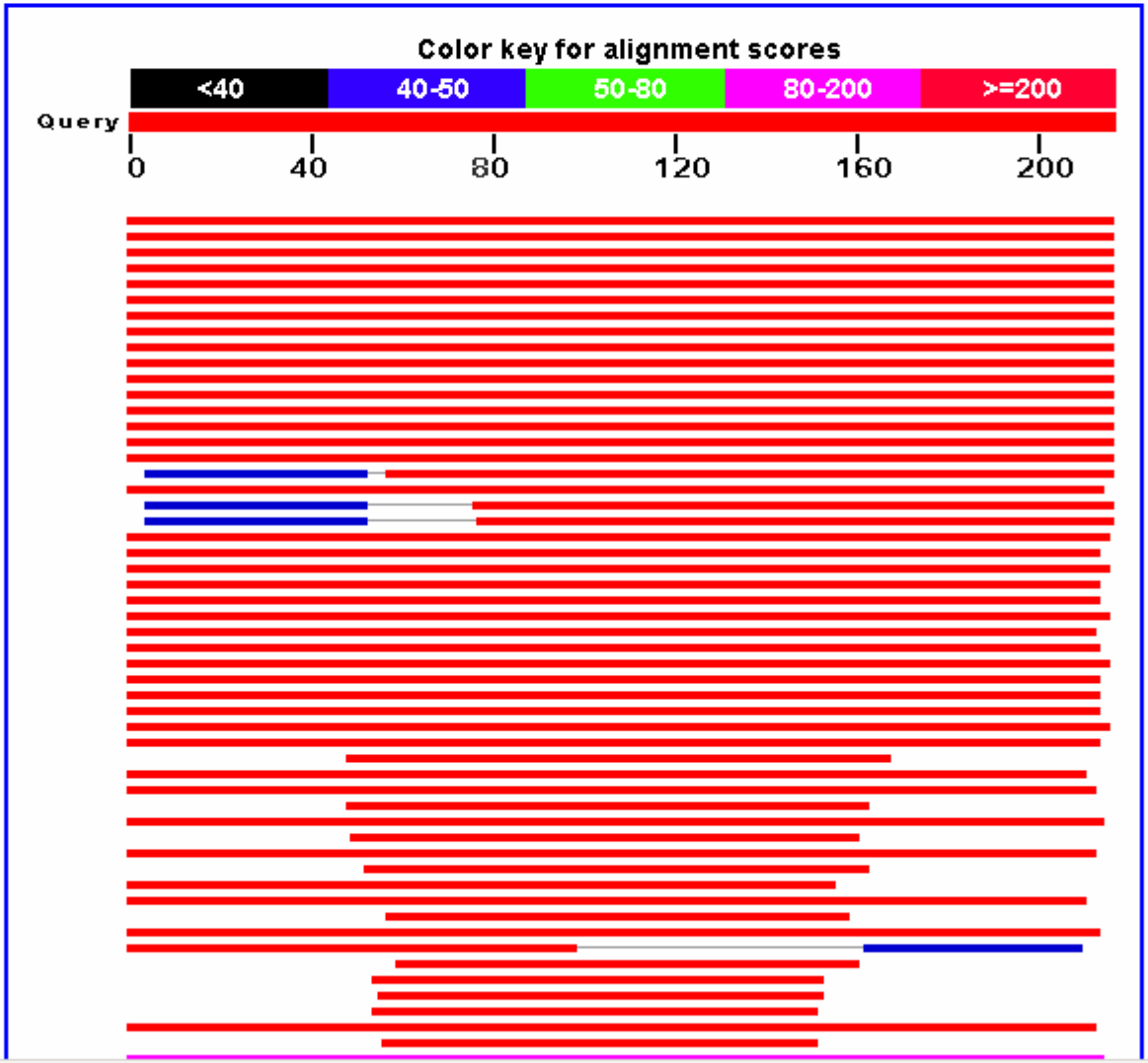
http://www.ncbi.nlm.nih.gov/blast/Blast.cgi

Go

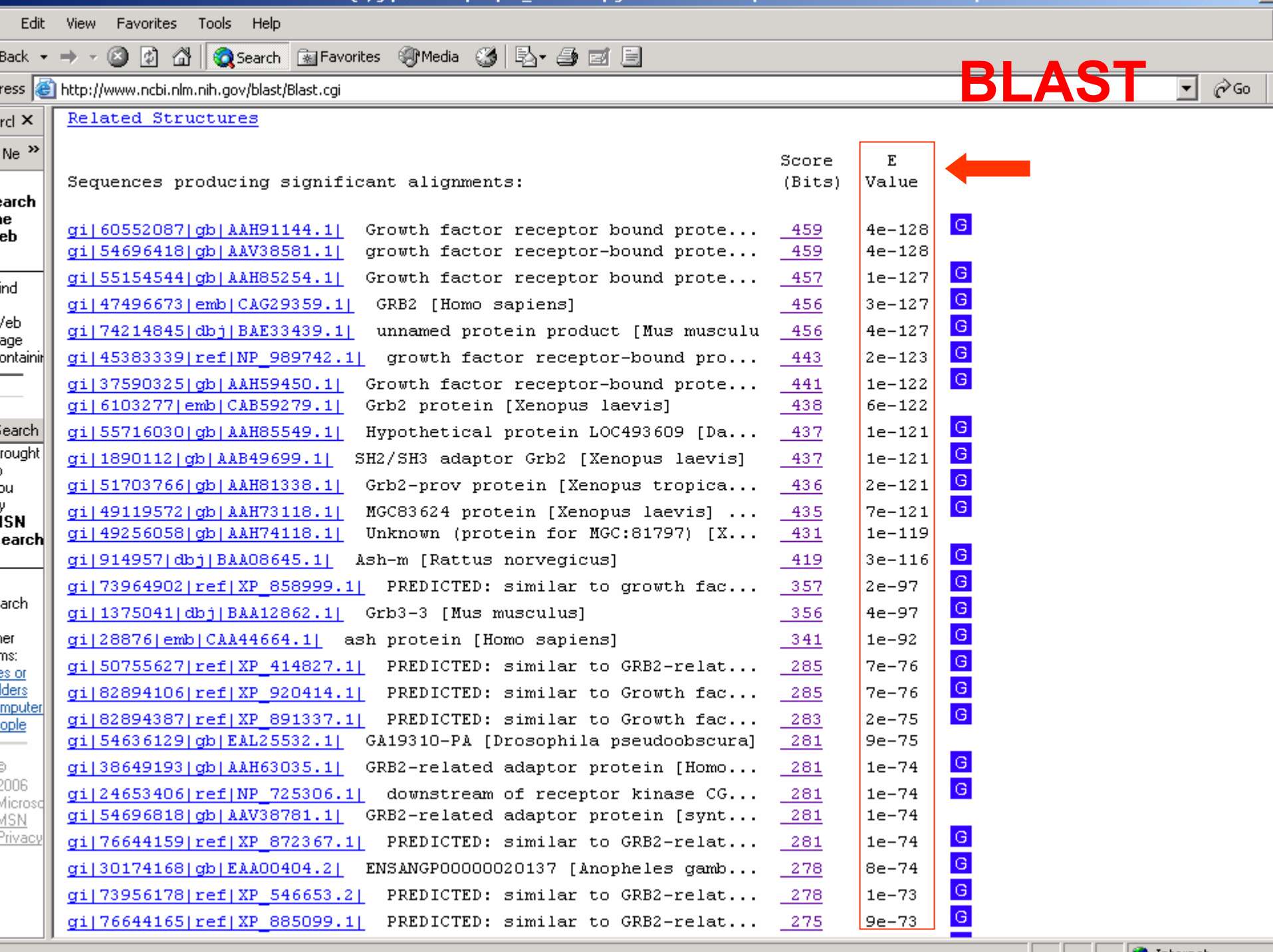
BLAST

Distribution of 985 Blast Hits on the Query Sequence

Mouse-over to show defline and scores, click to show alignments



Internet


Edit View Favorites Tools Help
Back Search Favorites Media
http://www.ncbi.nlm.nih.gov/blast/Blast.cgi
Go
BLAST
Related Structures
Sequences producing significant alignments:
Score (Bits)
E Value
gi|60552087|gb|AAH91144.1| Growth factor receptor bound prote... 459 4e-128
gi|54696418|gb|AAV38581.1| growth factor receptor-bound prote... 459 4e-128
gi|55154544|gb|AAH85254.1| Growth factor receptor bound prote... 457 1e-127
gi|47496673|emb|CAG29359.1| GRB2 [Homo sapiens] 456 3e-127
gi|74214845|dbj|BAE33439.1| unnamed protein product [Mus musculu 456 4e-127
gi|45383339|ref|NP_989742.1| growth factor receptor-bound pro... 443 2e-123
gi|37590325|gb|AAH59450.1| Growth factor receptor-bound prote... 441 1e-122
gi|6103277|emb|CAB59279.1| Grb2 protein [Xenopus laevis] 438 6e-122
gi|55716030|gb|AAH85549.1| Hypothetical protein LOC493609 [Da... 437 1e-121
gi|1890112|gb|AAB49699.1| SH2/SH3 adaptor Grb2 [Xenopus laevis] 437 1e-121
gi|51703766|gb|AAH81338.1| Grb2-prov protein [Xenopus tropica... 436 2e-121
gi|49119572|gb|AAH73118.1| MGC83624 protein [Xenopus laevis] ... 435 7e-121
gi|49256058|gb|AAH74118.1| Unknown (protein for MGC:81797) [X... 431 1e-119
gi|914957|dbj|BAA08645.1| Ash-m [Rattus norvegicus] 419 3e-116
gi|73964902|ref|XP_858999.1| PREDICTED: similar to growth fac... 357 2e-97
gi|1375041|dbj|BAA12862.1| Grb3-3 [Mus musculus] 356 4e-97
gi|28876|emb|CAA44664.1| ash protein [Homo sapiens] 341 1e-92
gi|50755627|ref|XP_414827.1| PREDICTED: similar to GRB2-relat... 285 7e-76
gi|82894106|ref|XP_920414.1| PREDICTED: similar to Growth fac... 285 7e-76
gi|82894387|ref|XP_891337.1| PREDICTED: similar to Growth fac... 283 2e-75
gi|54636129|gb|EAL25532.1| GA19310-PA [Drosophila pseudoobscura] 281 9e-75
gi|38649193|gb|AAH63035.1| GRB2-related adaptor protein [Homo... 281 1e-74
gi|24653406|ref|NP_725306.1| downstream of receptor kinase CG... 281 1e-74
gi|54696818|gb|AAV38781.1| GRB2-related adaptor protein [synt... 281 1e-74
gi|76644159|ref|XP_872367.1| PREDICTED: similar to GRB2-relat... 281 1e-74
gi|30174168|gb|EAA00404.2| ENSANGP00000020137 [Anopheles gamb... 278 8e-74
gi|73956178|ref|XP_546653.2| PREDICTED: similar to GRB2-relat... 278 1e-73
gi|76644165|ref|XP_885099.1| PREDICTED: similar to GRB2-relat... 275 9e-73

Affine Phospho Peptide

gi|20150610|pdb|1JYQ|A S Chain A, Xray Structure Of Grb2 Sh2 Domain Complexed With A Highly
Affine Phospho Peptide
Length=96

Score = 189 bits (481), Expect = 5e-47
Identities = 92/92 (100%), Positives = 92/92 (100%), Gaps = 0/92 (0%)

```
Query  60   WFFGKIPRAKAEEMLSKQRHDGAFLIRESESAPGDFSLSVKFGNDVQHFKVLRDGAGKYF  119
            WFFGKIPRAKAEEMLSKQRHDGAFLIRESESAPGDFSLSVKFGNDVQHFKVLRDGAGKYF
Sbjct  5    WFFGKIPRAKAEEMLSKQRHDGAFLIRESESAPGDFSLSVKFGNDVQHFKVLRDGAGKYF  64

Query  120  LWVVKFNSLNELVDYHRSTSVSRNQQIFLRDI  151
            LWVVKFNSLNELVDYHRSTSVSRNQQIFLRDI
Sbjct  65   LWVVKFNSLNELVDYHRSTSVSRNQQIFLRDI  96
```

> gi|72081594|ref|XP_788430.1| G PREDICTED: similar to CG6033-PA, isoform A [Strongylocentrotus purpuratus]
Length=177

Score = 178 bits (451), Expect = 1e-43
Identities = 103/214 (48%), Positives = 128/214 (59%), Gaps = 43/214 (20%)

```
Query  2    EAIAKYDFKATADDELSFKRGDILKVLNEECDQNWYKAELNGKDGFIPKNYIEMKPHPWF  61
            EA AK+DF    + ELSFK+  ILKV                +DG
Sbjct  3    EATAKHDFNGQEESELSFKKNSILKVT---------------RDG---------------  32

Query  62   FGKIPRAKAEEMLSKQRHDGAFLIRESESAPGDFSLSVKFGNDVQHFKVLRDGAGKYFLW  121
                    AEE+L K   DGAFLIRESE  PGD+SLSVKF + VQHFKVLRDGAGKYFLW
Sbjct  33   --------AEELL-KNDGDGAFLIRESEGTPGDYSLSVKFVDGVQHFKVLRDGAGKYFLW  83

Query  122  VVKFNSLNELVDYHRSTSVSRNQQIFLRDIEQVPQQPTYVQALFDFDPQEDGELGFRRGD  181
            VVKFNSLN+LV+YHR++SVSR+Q I+L+D  +  +    V AL+DF   E+GEL F++GD
Sbjct  84   VVKFNSLNQLVEYHRTSSVSRSQTIYLKD--RKSESIHLVLALYDFTAGEEGELSFKKGD  141

Query  182  FIHVMDNSDPNWWKG--ACHGQTGMFPRNYVTPV  213
```

1. Ricerca di omologhe in banche dati.

2. Programmi per la ricerca:

   FASTA

   BLAST