Similarity Searches on Sequence Databases: BLAST, FASTA

Lorenza Bordoli
Swiss Institute of Bioinformatics

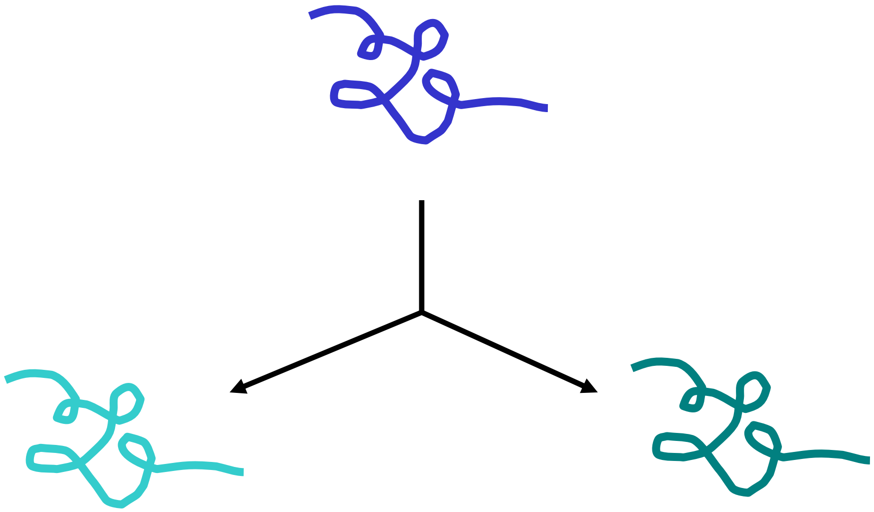
EMBnet Course, Basel, October 2003

Outline

- Importance of Similarity
- Heuristic Sequence Alignment:
  - Principle
  - FASTA algorithm
  - BLAST algorithm
- Assessing the significance of sequence alignment
  - The Extreme Value Distribution (EVD)
  - P-value, E-Value
- BLAST:
  - Protein Sequences
  - DNA Sequences
  - Choosing the right Parameters
- Other members of the BLAST family

Importance of Similarity

Importance of Similarity

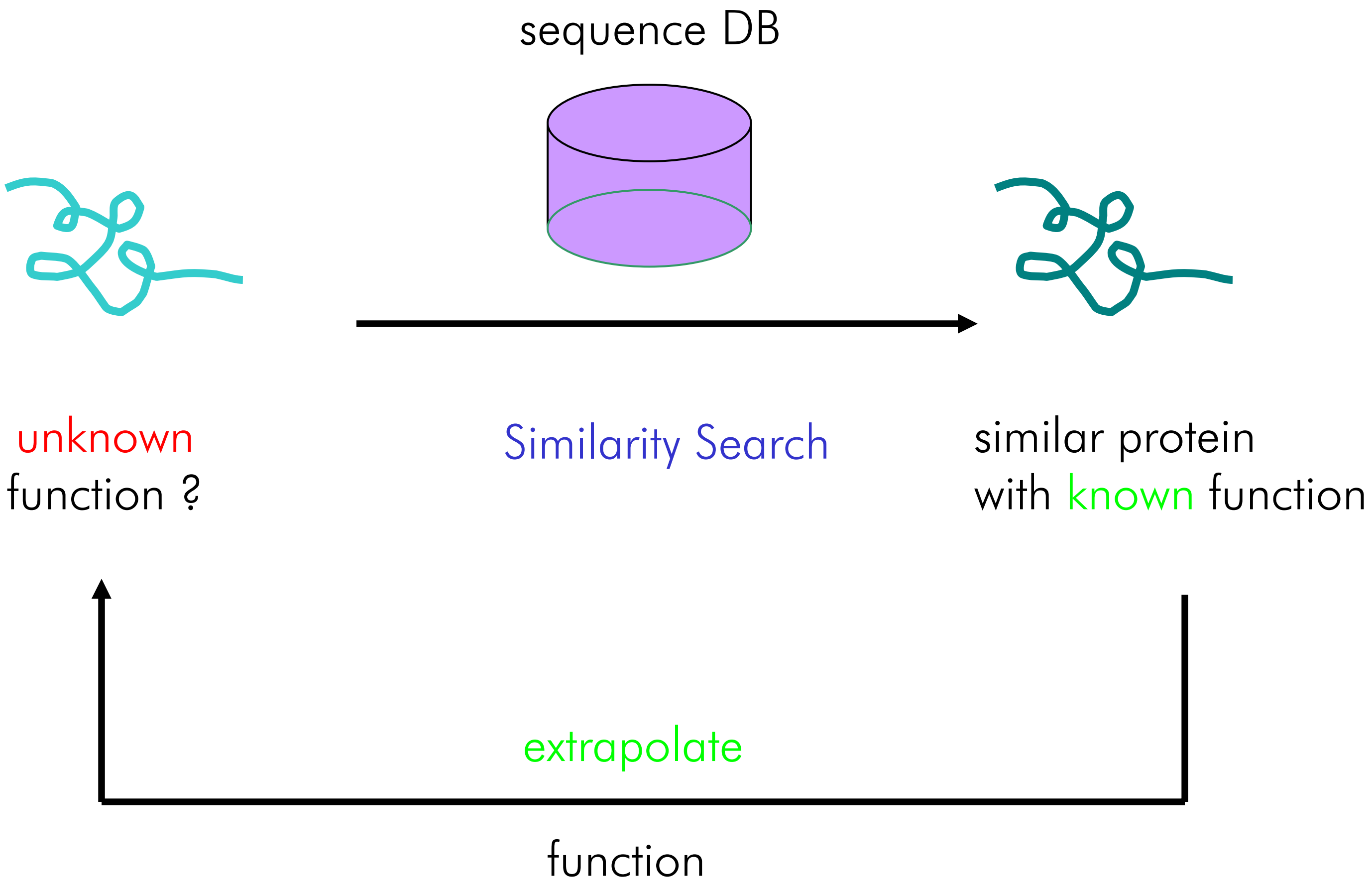


ancestral
protein/gene sequence

similar (homologous)
protein/gene sequences

similar sequences: probably have the same ancestor, share the same structure, and have a similar biological function

Importance of Similarity



sequence DB

unknown
function ?

Similarity Search

similar protein
with known function

extrapolate

function

Importance of Similarity

Rule-of-thumb:
If your sequences are more than 100 amino acids long (or 100 nucleotides long) you can considered them as homologues if 25% of the aa are identical (70% of nucleotide for DNA). Below this value you enter the twilight zone.

Twilight zone = protein sequence similarity between ∼0-20% identity:
is not statistically significant, i.e. could have arisen by chance.

Beware:

- E-value (*Expectation value*)
- length of the segments similar between the two sequences
- The number of insertions/deletions

Heuristic sequence alignment

Heuristic Sequence Alignment

- With the Dynamic Programming algorithm, one obtain an alignment in a time that is proportional to the product of the lengths of the two sequences being compared. Therefore when searching a whole database the computation time grows linearly with the size of the database. With current databases calculating a full Dynamic Programming alignment for each sequence of the database is too slow (unless implemented in a specialized parallel hardware).

- The number of searches that are presently performed on whole genomes creates a need for faster procedures.

⇒ Two methods that are least 50-100 times faster than dynamic programming were developed: FASTA and BLAST

Heuristic Sequence Alignment: Principle

- Dynamic Programming: computational method that provide in mathematical sense the best alignment between two sequences, given a scoring system.

- Heuristic Methods (e.g. BLAST, FASTA) they prune the search space by using fast approximate methods to select the sequences of the database that are likely to be similar to the query and to locate the similarity region inside them

  =>Restricting the alignment process:

  - Only to the selected sequences
  - Only to some portions of the sequences (search as small a fraction as possible of the cells in the dynamic programming matrix)

Heuristic Sequence Alignment: Principle

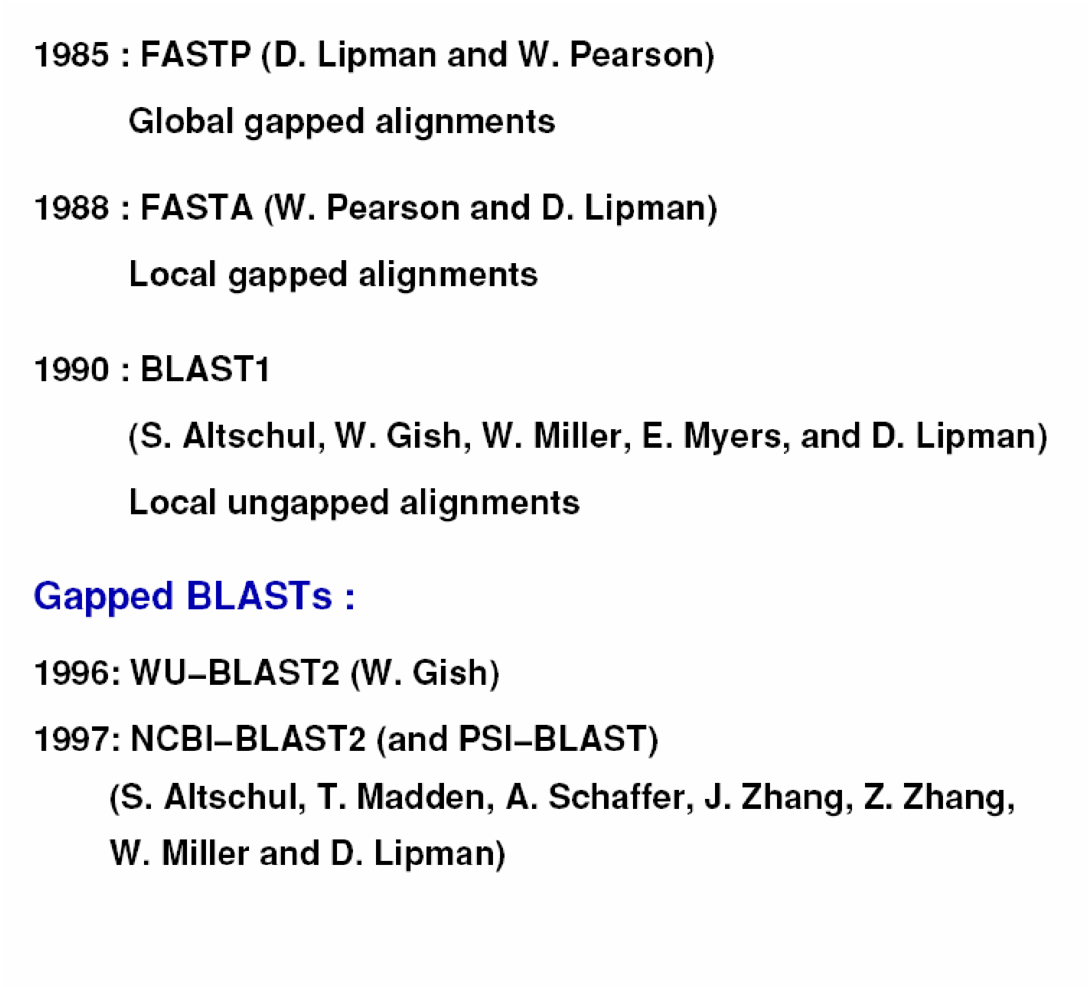
- These methods are heuristic; i.e., an empirical method of computer programming in which rules of thumb are used to find solutions.

- They almost always works to find related sequences in a database search but does not have the underlying guarantee of an optimal solution like the dynamic programming algorithm.

- Advantage: This methods that are least 50-100 times faster than dynamic programming therefore better suited to search DBs.

FASTA & BLAST

FASTA & BLAST: story



1985 : FASTP (D. Lipman and W. Pearson)

Global gapped alignments

1988 : FASTA (W. Pearson and D. Lipman)

Local gapped alignments

1990 : BLAST1

(S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman)

Local ungapped alignments

Gapped BLASTs :

1996: WU–BLAST2 (W. Gish)

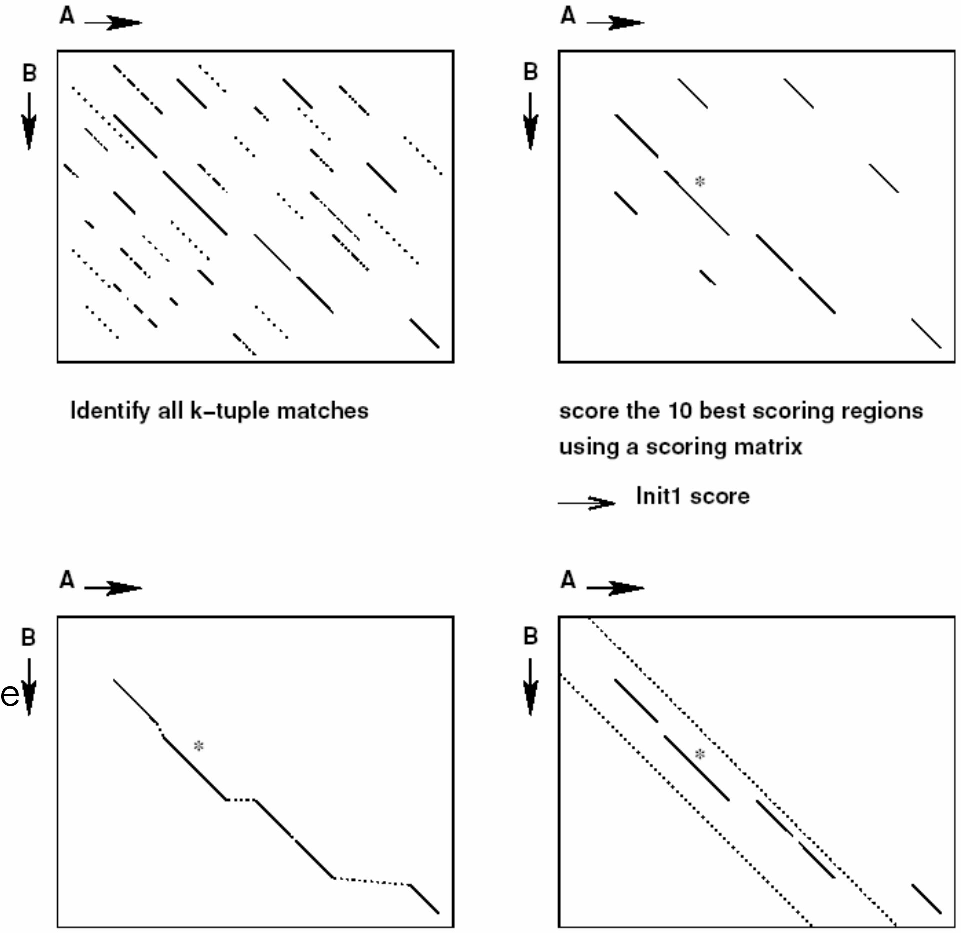
1997: NCBI–BLAST2 (and PSI–BLAST)
(S. Altschul, T. Madden, A. Schaffer, J. Zhang, Z. Zhang,
W. Miller and D. Lipman)

FASTA

FASTA: algorithm (4 steps)

Localize the 10 best regions of similarity between the two seq.
Each identity between two “word” is represented by a dot

Find the best combination of the diagonals-> compute a score.
Only those sequences with a score higher than a threshold will go to the fourth step



Identify all k–tuple matches

score the 10 best scoring regions using a scoring matrix

→ Init1 score

Apply joining procedure

→ Initn score

Apply limited DP

→ Opt score

Each diagonal: ungapped alignment

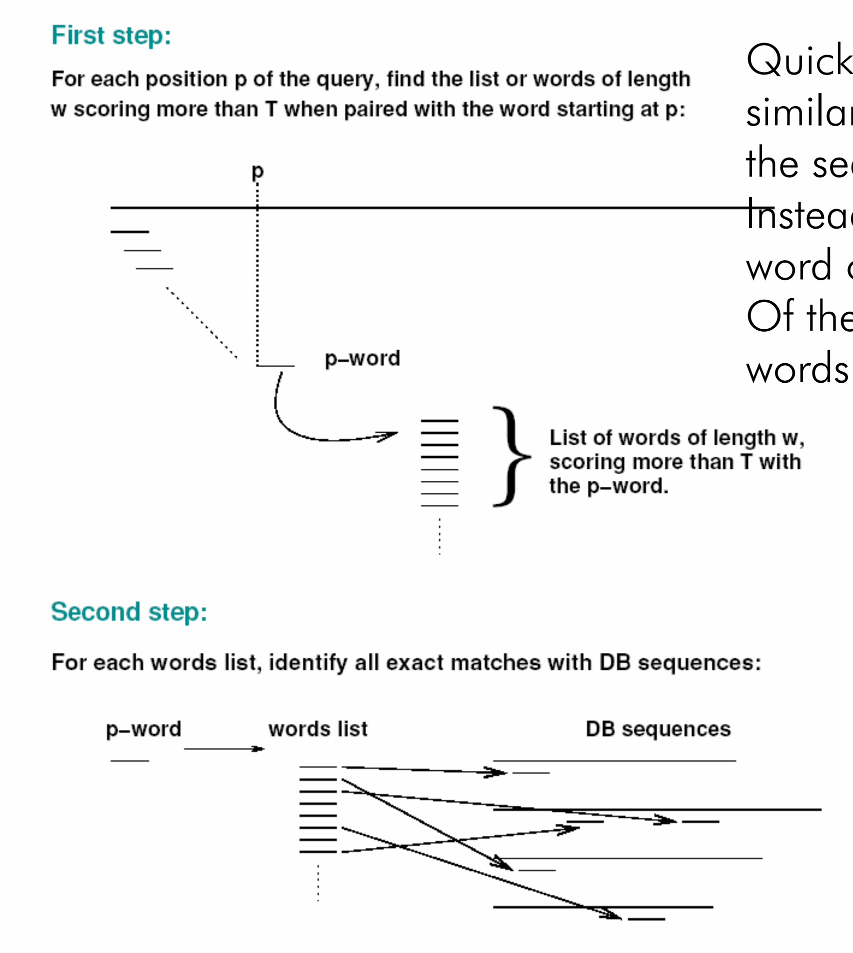
The smaller the k, The sensitive the method but slower

DP applied around The best scoring diagonal.

BLAST

BLAST1: Algorithm

With w=2 :
(20x20=400
Possible words,
w=3, 8000
Possible words,…)



First step:

For each position p of the query, find the list or words of length w scoring more than T when paired with the word starting at p:

p
p–word
List of words of length w, scoring more than T with the p–word.

Second step:

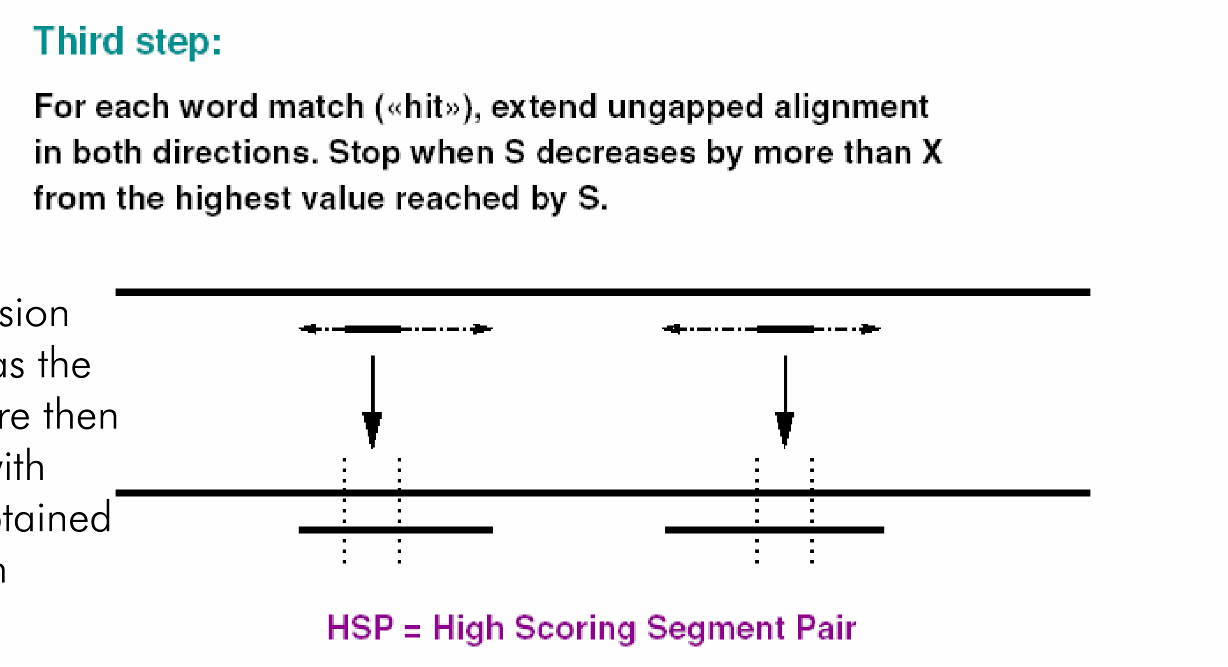
For each words list, identify all exact matches with DB sequences:

p–word
words list
DB sequences

Quickly locate ungapped
similarity regions between
the sequences.
Instead of comparing each
word of the query with each word
Of the DB: create a list of “similar”
words.

BLAST1: Algorithm

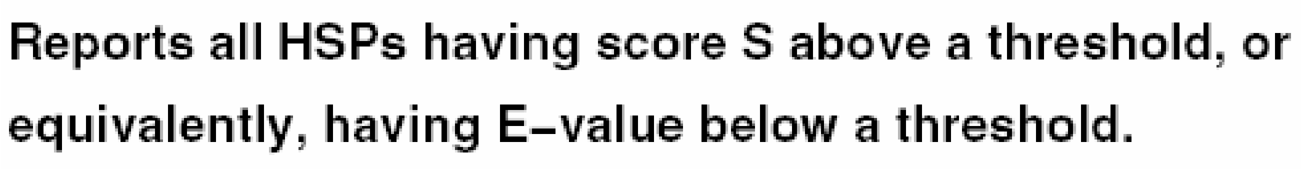
Each match is then extended. The extension is stopped as soon as the score decreases more then X when compared with the highest value obtained During the extension process



Third step:

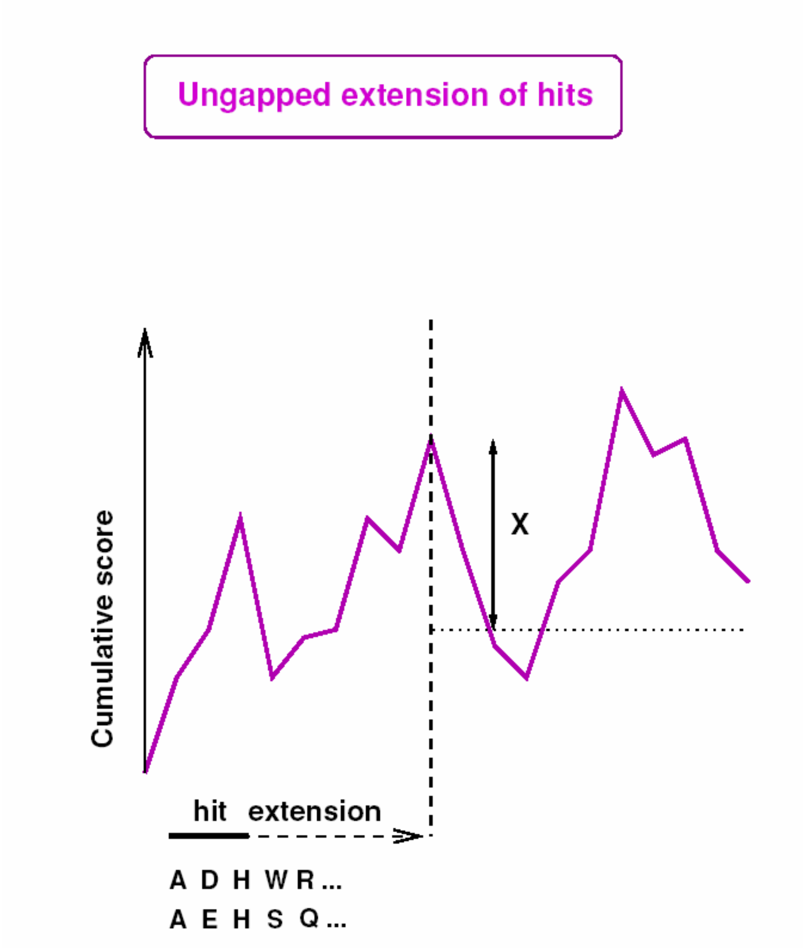
For each word match («hit»), extend ungapped alignment in both directions. Stop when S decreases by more than X from the highest value reached by S.

HSP = High Scoring Segment Pair



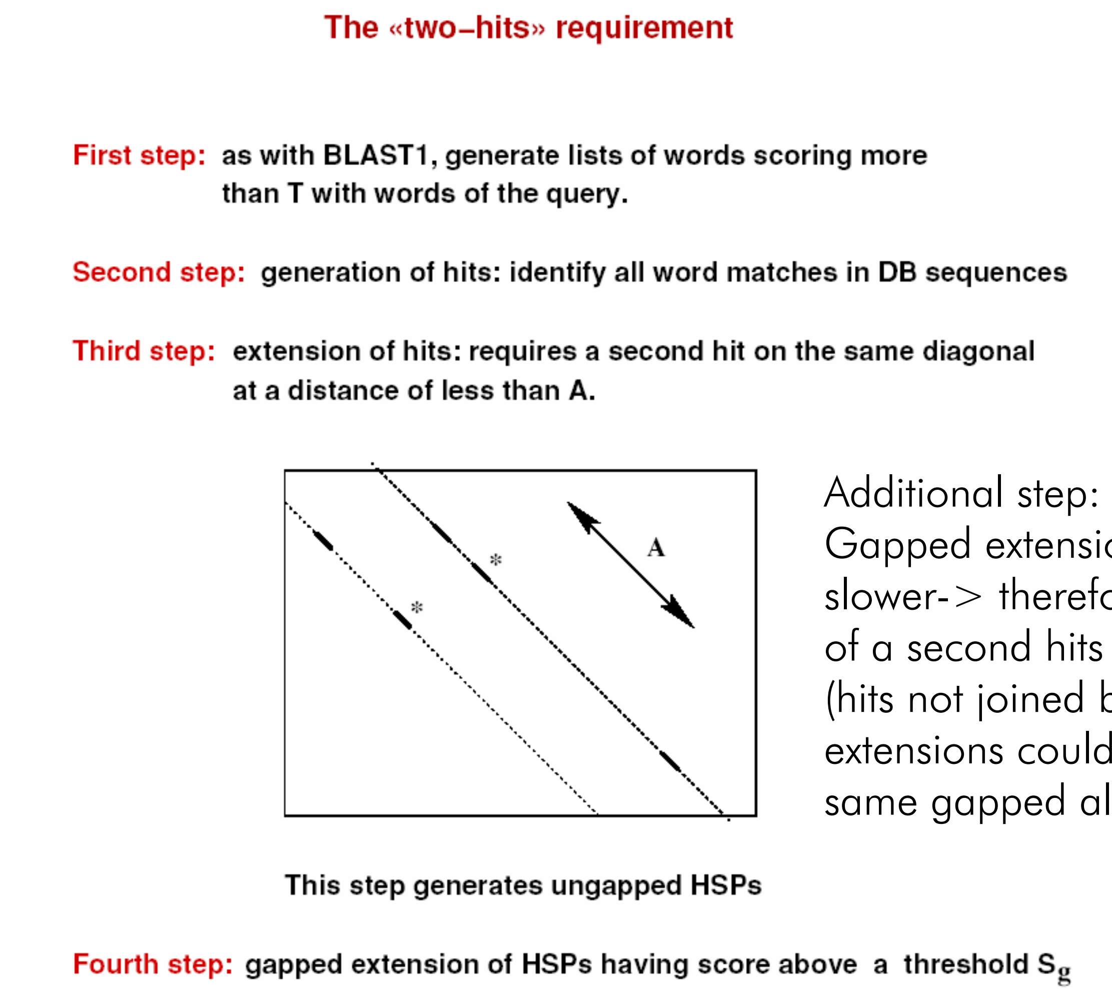
Reports all HSPs having score S above a threshold, or equivalently, having E–value below a threshold.

BLAST1: Algorithm



Each match is then extended. The extension is stopped as soon as the score decreases more then X when compared with the highest value obtained During the extension process

BLAST2: (NCBI)



The «two–hits» requirement

First step: as with BLAST1, generate lists of words scoring more than T with words of the query.

Second step: generation of hits: identify all word matches in DB sequences

Third step: extension of hits: requires a second hit on the same diagonal at a distance of less than A.

This step generates ungapped HSPs

Fourth step: gapped extension of HSPs having score above a threshold S_g

Additional step:
Gapped extension of the hits slower-> therefore: requirement of a second hits on the diagonal.
(hits not joined by ungapped extensions could be part of the same gapped alignmnet)

Assessing the significance of sequence alignment

Assessing the significance of sequence alignment

- Scoring System:
  - 1. Scoring (Substitution) matrix: In proteins some mismatches are more acceptable than others. Substitution matrices give a score for each substitution of one amino-acid by another (e.g. PAM, BLOSUM)
  - 2. Gap Penalties: simulate as closely as possible the evolutionary mechanisms involved in gap occurrence. Gap opening penalty: Counted each time a gap is opened in an alignment and Gap extension penalty: Counted for each extension of a gap in an alignment.
- Based on a given scoring system: you can calculate the raw score of the alignment
  - Raw score= sum of the amino acid substitution scores and gap penalties

Assessing the significance of sequence alignment

Caveats:

1. We need a normalised score to compare different alignments, based on different scoring systems, e.g. different substitution matrices.

2. It is possible that a good long alignment gets a better raw score than a very good short alignment

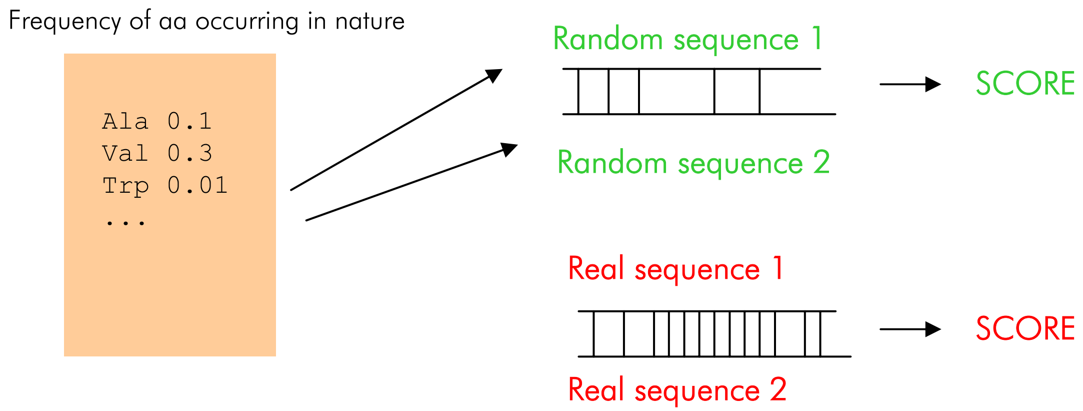
=> a method to asses the statistical significance of the alignment is needed (is an alignment biological relevant?) : E-value

Assessing the significance of sequence alignment

- How?

  ⇒ Evaluate the probability that a score between random or unrelated sequences will reach the score found between two real sequences of interest:

  If that probability is very low, the alignment score between the real sequences is significant.



Frequency of aa occurring in nature

```
Ala 0.1
Val 0.3
Trp 0.01
...
```

Random sequence 1

Random sequence 2

SCORE

Real sequence 1

Real sequence 2

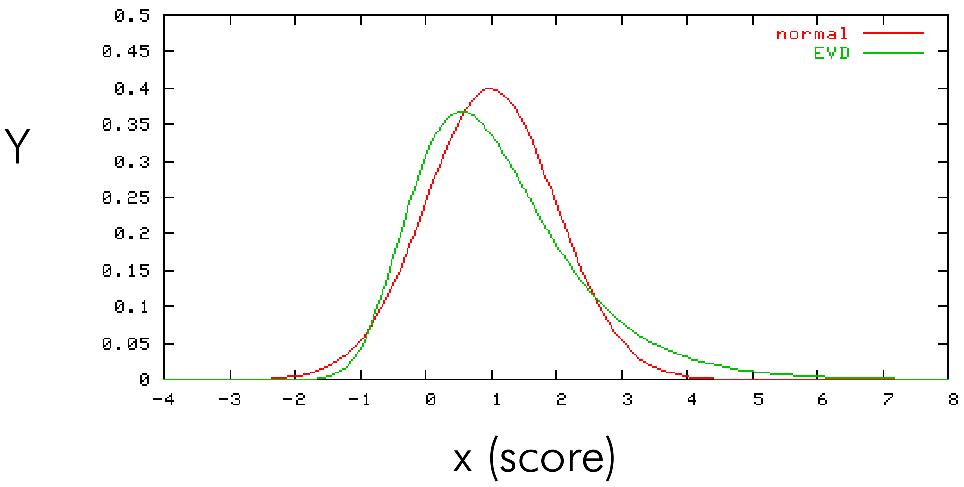
SCORE

If SCORE > SCORE => the alignment between the real sequences is significant

The Extreme Value Distribution (EVD)

The Extreme Value Distribution

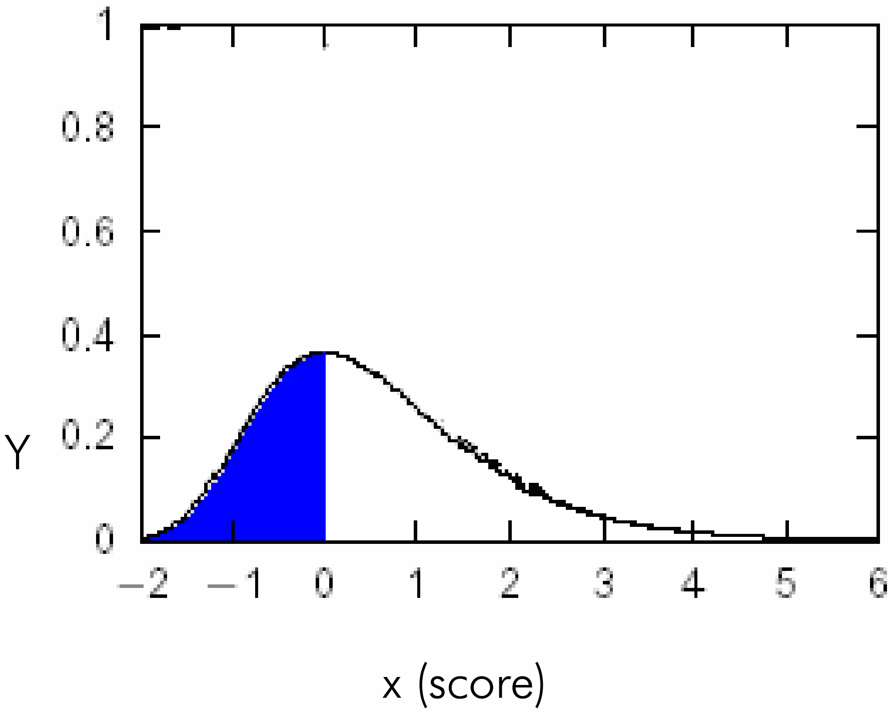
- Karlin and Altschul observed that in the framework of local alignments without gaps: the distribution of random sequence alignment scores follow an EVD.



$$Y = \lambda \exp[-\lambda(x-\mu) - e^{-\lambda(x-\mu)}]$$

μ, λ : parameters depend on the length and composition of the sequences and on the scoring system

The Extreme Value Distribution

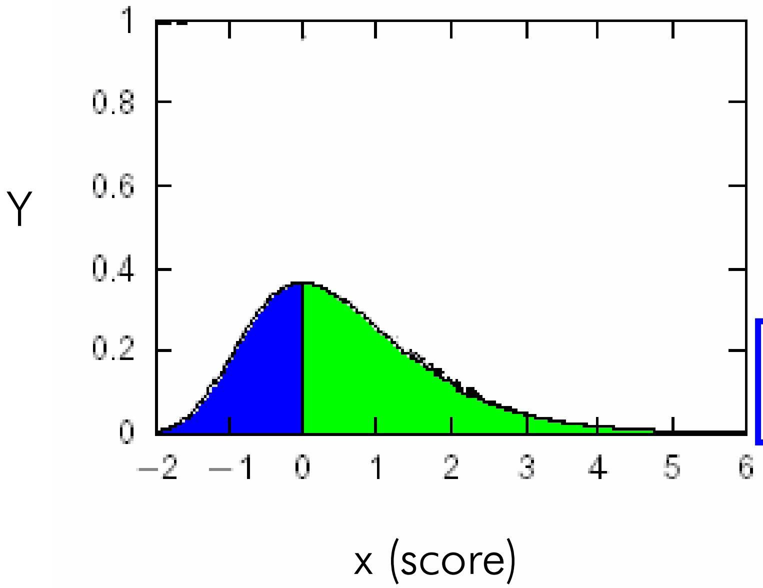


$$Y = \lambda \exp[-\lambda(x-\mu) - e^{-\lambda(x-\mu)}]$$

$$\Big\downarrow \int$$

$$\boxed{P(S < x) = \exp[-e^{-\lambda(x-\mu)}]}$$

The Extreme Value Distribution



$$Y = \lambda \exp[-\lambda(x-\mu) - e^{-\lambda(x-\mu)}]$$

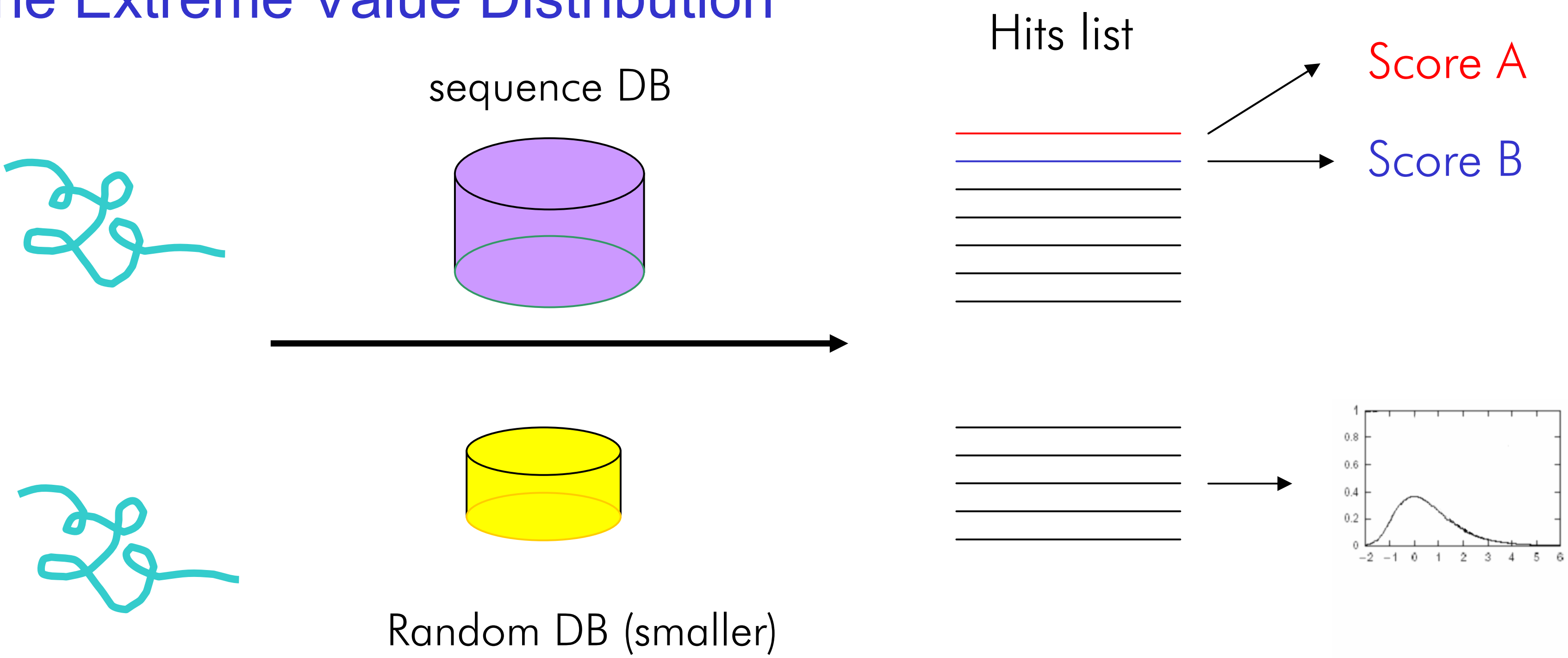
$$\Big\downarrow \int$$

$$\boxed{P(S < x) = \exp[-e^{-\lambda(x-\mu)}]}$$

$$\boxed{P(S \geq x) = 1 - \exp[-e^{-\lambda(x-\mu)}]}$$

P-value = the probability of obtaining a score equal or greater than *x* by chance

The Extreme Value Distribution



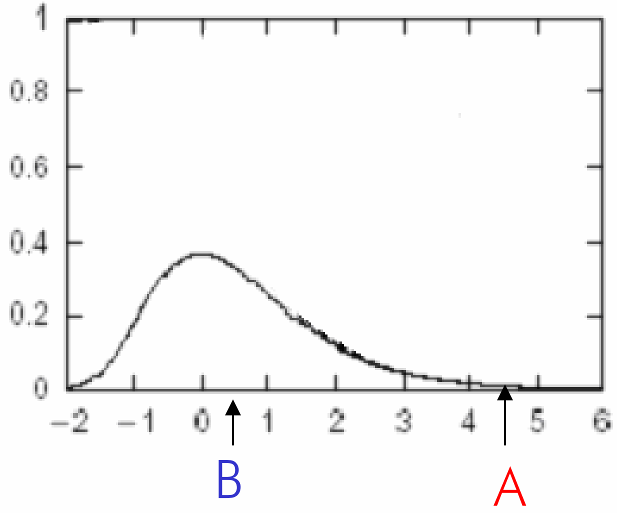
sequence DB

Hits list

Score A

Score B

Random DB (smaller)



Score A: is significant

Score B: is NOT significant

Assessing the significance of sequence alignment

In a database of size N: P x N = E

- P-value:

Probability that an alignment with this score occurs by chance in a database of size N.

The closer the P-value is towards 0, the better the alignment

- E-value:

Number of matches with this score one can expect to find by chance in a database of size N.

The closer the E-value is towards 0, the better the alignment

Assessing the significance of sequence alignment

- Local alignment without gaps:
  - Theoretical work: Karlin-Altschul statistics: -> Extreme Value Distribution

- Local alignments with gaps:
  - Empirical studies: -> Extreme Value Distribution

EVD: More formalisms (1)

$$P(S \geq x) = 1 - \exp[-e^{-\lambda(x-\mu)}] \qquad (1)$$

μ, λ : parameters depend on the length and composition of the sequences and on the scoring system: μ is the mode (highest point) of the distribution and λ is the decay parameter
-They can me estimated by making many alignments of random or shuffled sequences.
- For alignments without gaps they can be calculated from the scoring matrix and then :

$$P(S \geq x) = 1 - \exp[-Kmne^{-\lambda x}] \qquad (2)$$

K: is a constant that depend on the scoring matrix values and the frequencies of the different residues in the sequences.
m,n : sequence lengths

EVD: More formalisms (2)

- To facilitate calculations, the score S may be normalized to produce a score S′. The effect of normalization is to change the score distribution with a $\mu=0$ and a $\lambda=1$. S′ can be calculated from equation (2):

$$S' = \lambda S - \ln Kmn$$

- And then replacing S by S′ in (1) :

$$P(S' \geq x) = 1 - \exp[-e^{-x}]$$

Assessing the significance of sequence alignment

- BLAST2:
  - Artificial random sequences
- FASTA:
  - Uses results from the search: real unrelated sequences

BLASŤ

*<u>B</u>*asic *<u>L</u>*ocal *<u>A</u>*lignment *<u>S</u>*earch *<u>T</u>*ool

BLASTing protein sequences

BLASTing protein sequences

blastp = Compares a protein sequence with a protein database

If you want to find something about the function of your protein, use blastp to compare your protein with other proteins contained in the databases

tblastn = Compares a protein sequence with a nucleotide database

If you want to discover new genes encoding proteins, use tblastn to compare your protein with DNA sequences translated into their six possible reading frames

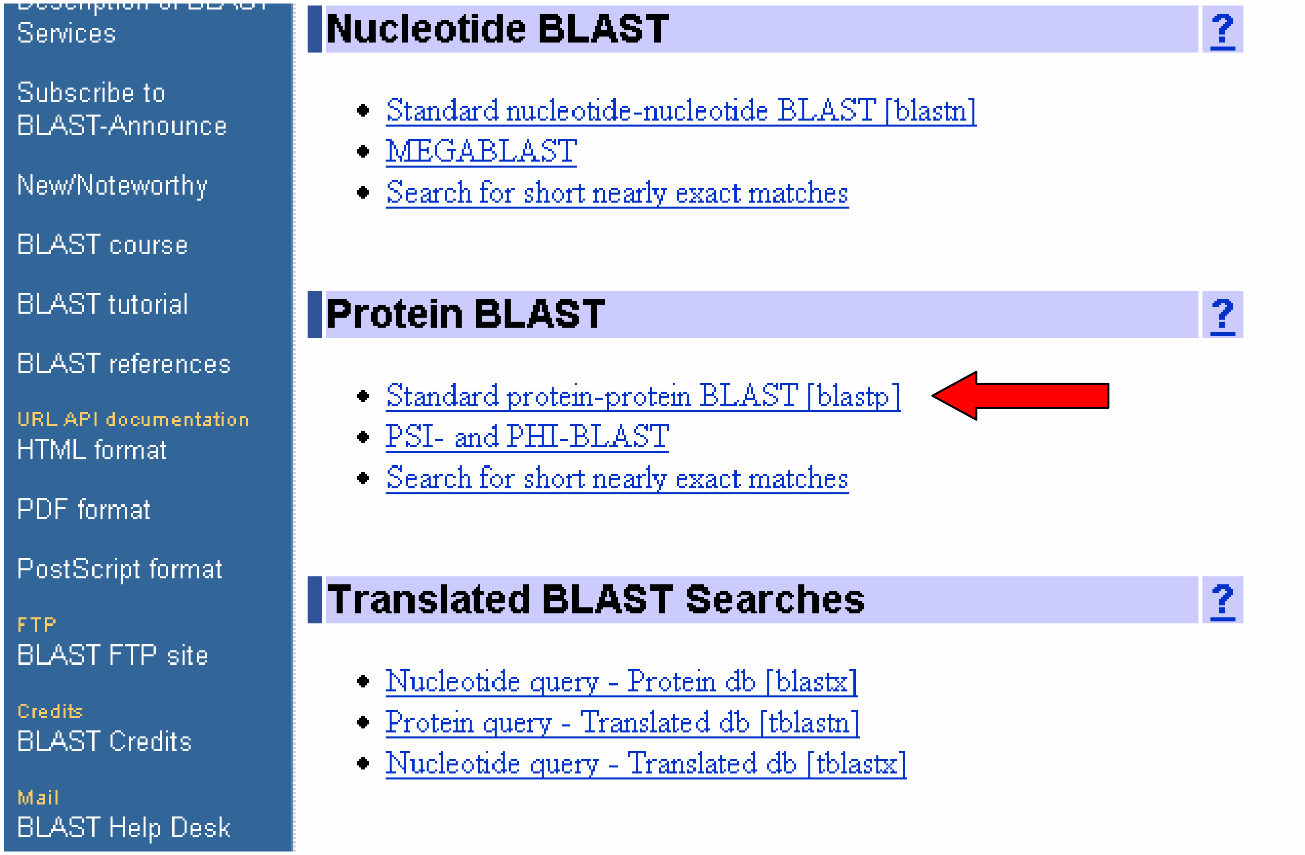
BLASTing protein sequences

Two of the most popular **blastp** online services:

- NCBI (National Center for Biotechnology Information) server
- Swiss EMBnet server (European Molecular Biology network)

BLASTing protein sequences: NCBI blastp server

- URL: http://www.ncbi.nlm.nih.gov/BLAST



Nucleotide BLAST

- Standard nucleotide-nucleotide BLAST [blastn]
- MEGABLAST
- Search for short nearly exact matches

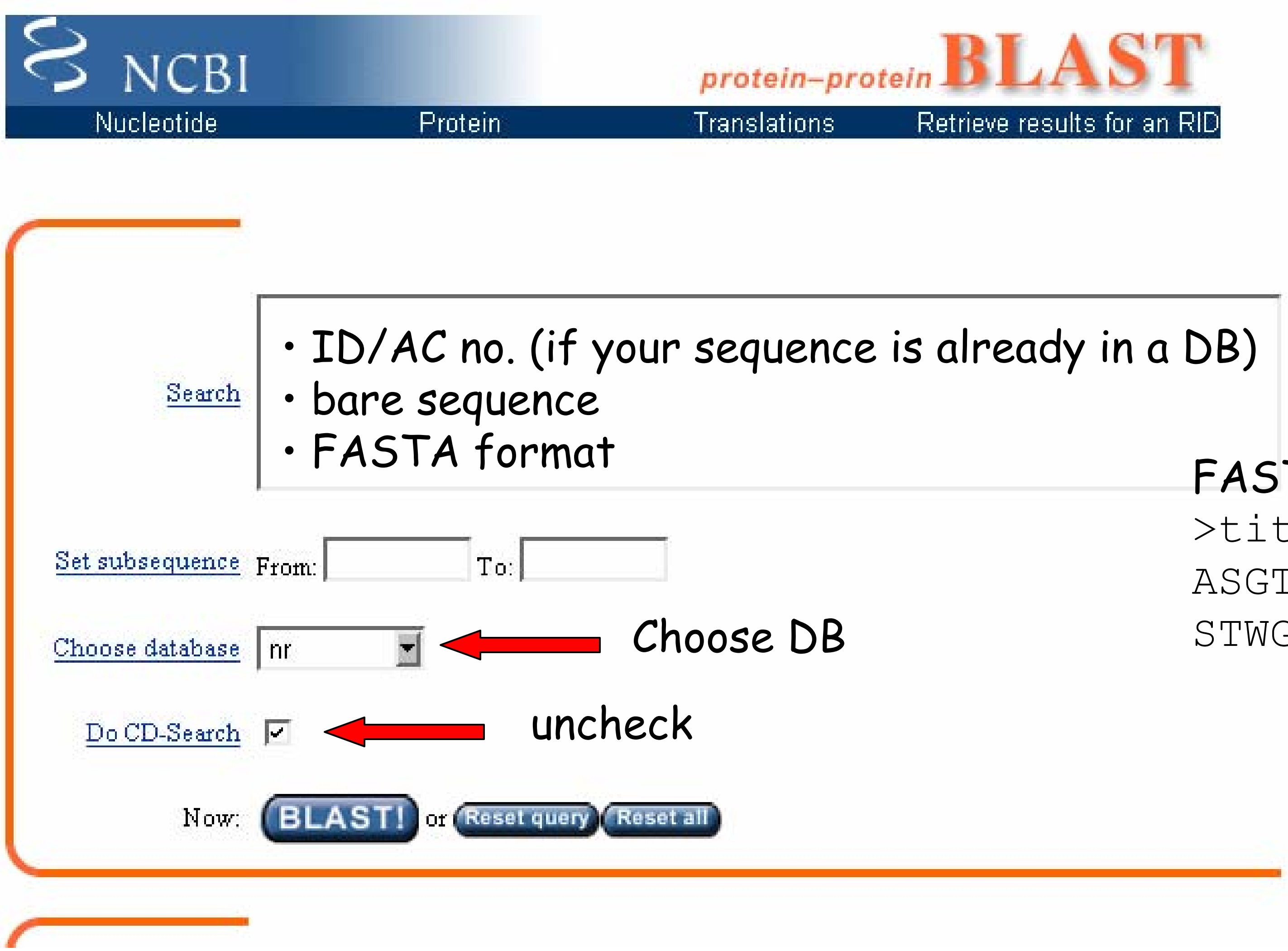
Protein BLAST

- Standard protein-protein BLAST [blastp]
- PSI- and PHI-BLAST
- Search for short nearly exact matches

Translated BLAST Searches

- Nucleotide query - Protein db [blastx]
- Protein query - Translated db [tblastn]
- Nucleotide query - Translated db [tblastx]

BLASTing protein sequences: NCBI blastp server



Search

- ID/AC no. (if your sequence is already in a DB)
- bare sequence
- FASTA format

Set subsequence From: To:

Choose database nr ← Choose DB

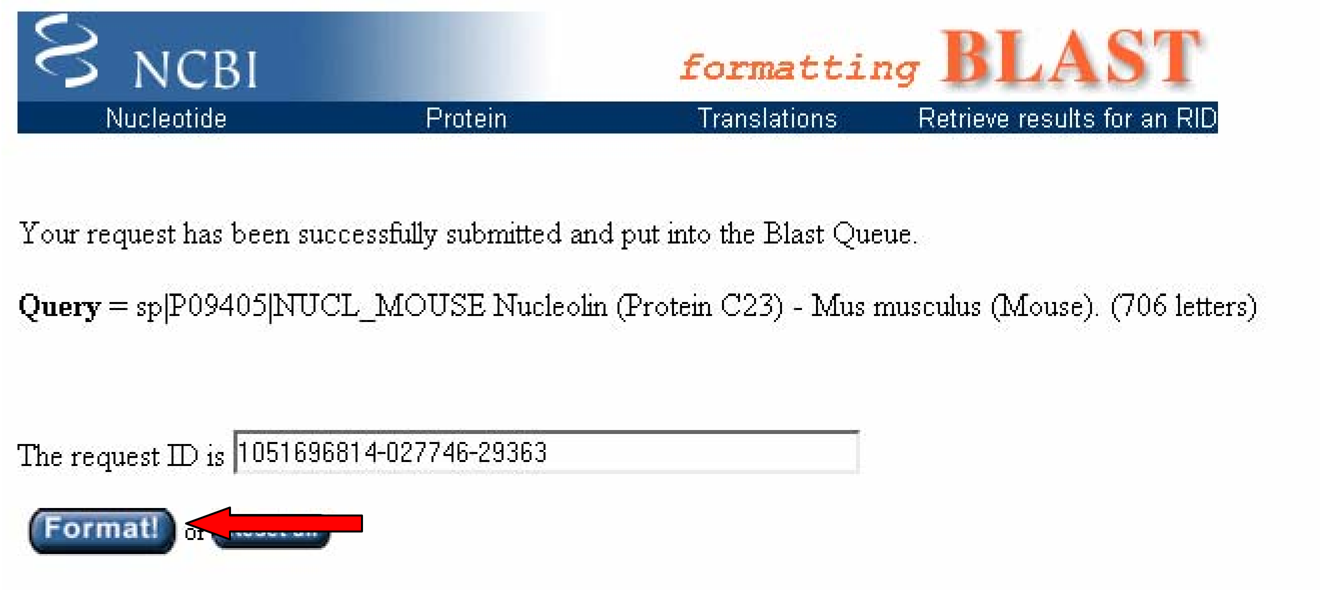
Do CD-Search ☑ ← uncheck

Now: BLAST! or Reset query Reset all

FASTA format:

```
>titel
ASGTRCVKDQQG
STWGPPFRTS
```

BLASTing protein sequences: NCBI blastp server



Your request has been successfully submitted and put into the Blast Queue.

Query = sp|P09405|NUCL_MOUSE Nucleolin (Protein C23) - Mus musculus (Mouse). (706 letters)

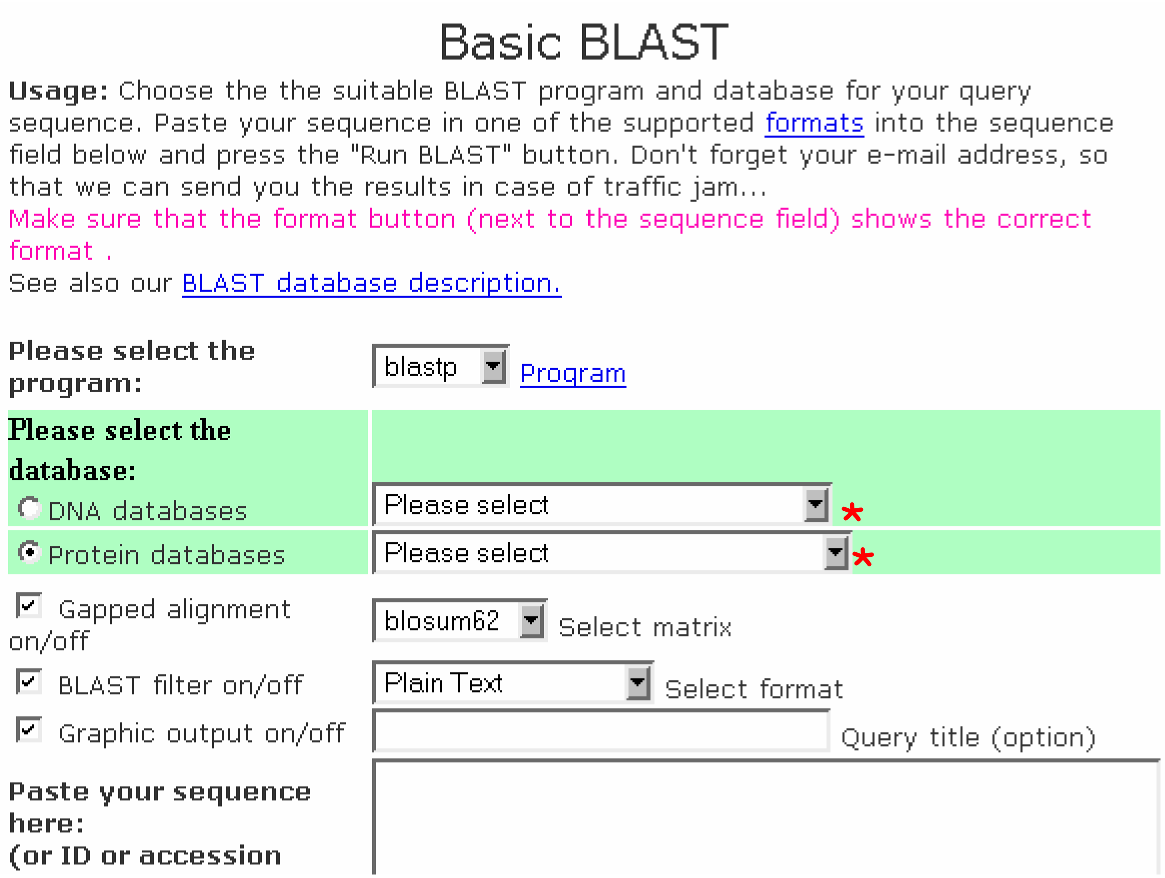
The request ID is 1051696814-027746-29363

If you get no reply, DO NOT resubmit the same query several times in a row - it will only make things worse for everybody (including you)!

BLASTing protein sequences: Swiss EMBnet blastp server

- URL: http://www.ch.embnet.org/software/bBLAST.html

The EMBnet interface gives you many more choices *:



Basic BLAST

Usage: Choose the the suitable BLAST program and database for your query sequence. Paste your sequence in one of the supported formats into the sequence field below and press the "Run BLAST" button. Don't forget your e-mail address, so that we can send you the results in case of traffic jam...

Make sure that the format button (next to the sequence field) shows the correct format .

See also our BLAST database description.

Please select the program: blastp ▾ Program

Please select the database:

○ DNA databases — Please select ▾ *

● Protein databases — Please select ▾ *

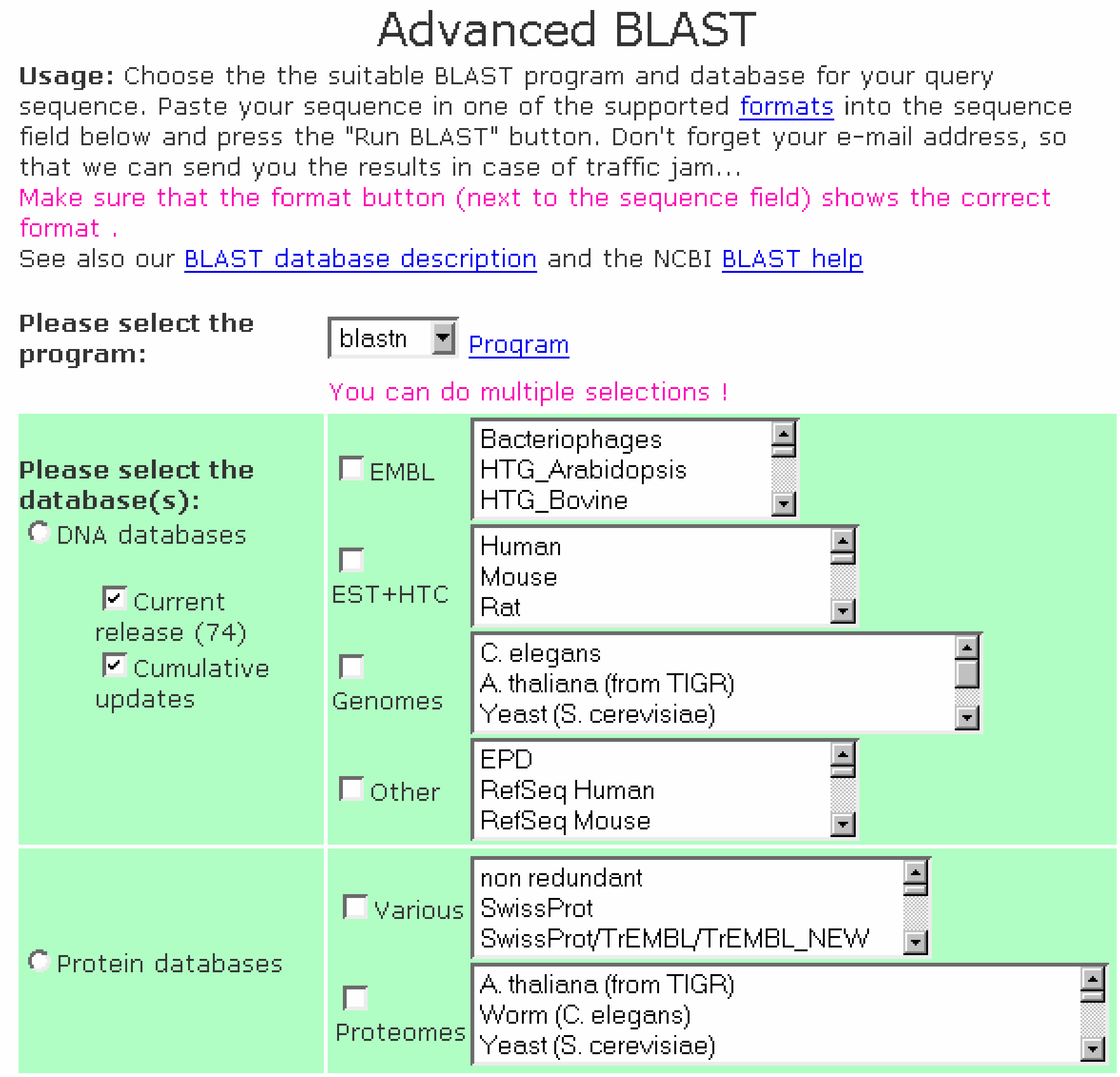
☑ Gapped alignment on/off — blosum62 ▾ Select matrix

☑ BLAST filter on/off — Plain Text ▾ Select format

☑ Graphic output on/off — Query title (option)

Paste your sequence here:
(or ID or accession

BLASTing protein sequences: Swiss EMBnet blasp server



Advanced BLAST

Usage: Choose the the suitable BLAST program and database for your query sequence. Paste your sequence in one of the supported formats into the sequence field below and press the "Run BLAST" button. Don't forget your e-mail address, so that we can send you the results in case of traffic jam...

Make sure that the format button (next to the sequence field) shows the correct format .

See also our BLAST database description and the NCBI BLAST help

Please select the program: blastn ▾ Program

You can do multiple selections !

Please select the database(s):

○ DNA databases

☑ Current release (74)

☑ Cumulative updates

□ EMBL: Bacteriophages, HTG_Arabidopsis, HTG_Bovine

□ EST+HTC: Human, Mouse, Rat

□ Genomes: C. elegans, A. thaliana (from TIGR), Yeast (S. cerevisiae)

□ Other: EPD, RefSeq Human, RefSeq Mouse

○ Protein databases

□ Various: non redundant, SwissProt, SwissProt/TrEMBL/TrEMBL_NEW

□ Proteomes: A. thaliana (from TIGR), Worm (C. elegans), Yeast (S. cerevisiae)

Understanding your BLAST output

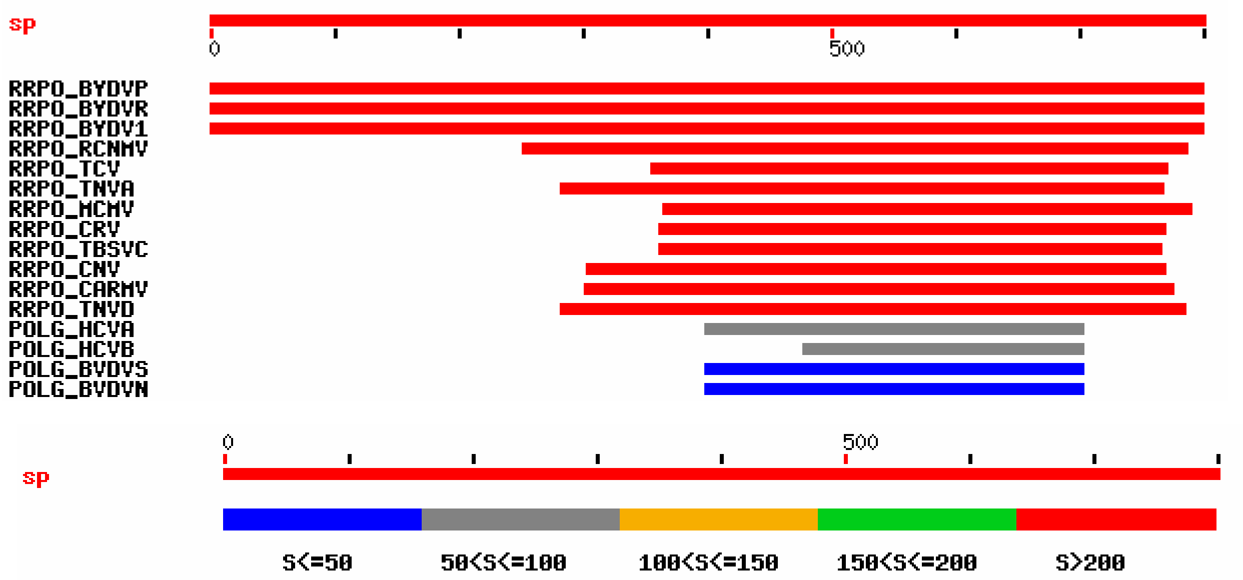
1. Graphic display:
   shows you where your query is similar to other sequences

2. Hit list:
   the name of sequences similar to your query, ranked by similarity

3. The alignment:
   every alignment between your query and the reported hits

4. The parameters:
   a list of the various parameters used for the search

Understanding your BLAST output: 1. Graphic display



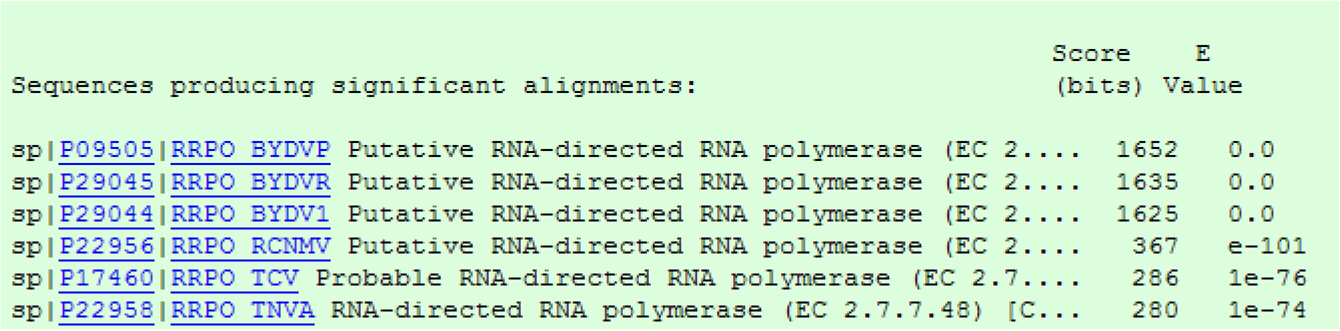
query sequence

Portion of another sequence similar to your query sequence:

red, green, ochre, matches: good
grey matches: intermediate
blue: bad, (twilight zone)

The display can help you see that some matches do not extend over the entire length of your sequence => useful tool to discover domains.

Understanding your BLAST output: 2. Hit list



```
                                                                 Score    E
Sequences producing significant alignments:                      (bits)  Value

sp|P09505|RRPO_BYDVP Putative RNA-directed RNA polymerase (EC 2....  1652   0.0
sp|P29045|RRPO_BYDVR Putative RNA-directed RNA polymerase (EC 2....  1635   0.0
sp|P29044|RRPO_BYDV1 Putative RNA-directed RNA polymerase (EC 2....  1625   0.0
sp|P22956|RRPO_RCNMV Putative RNA-directed RNA polymerase (EC 2....   367   e-101
sp|P17460|RRPO_TCV Probable RNA-directed RNA polymerase (EC 2.7....   286   1e-76
sp|P22958|RRPO_TNVA RNA-directed RNA polymerase (EC 2.7.7.48) [C...   280   1e-74
```

Sequence ac number and name

Description

Bit score

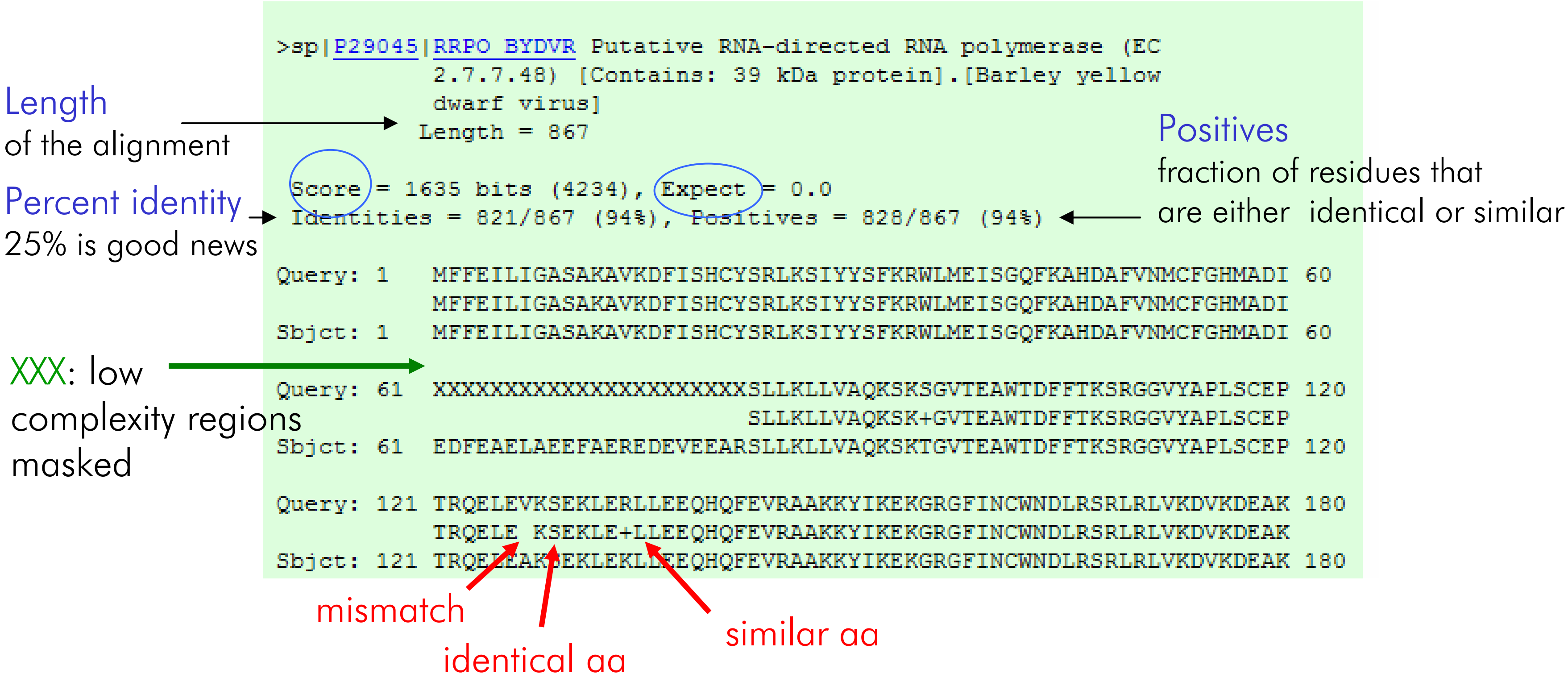
E-value

- Sequence ac number and name: Hyperlink to the database entry: useful annotations
- Description: better to check the full annotation

- Bit score (normalized score) : A measure of the similarity between the two sequences: the higher the better (matches below 50 bits are very unreliable)

- E-value: The lower the E-value, the better. Sequences identical to the query have an E-value of 0. Matches above 0.001 are often close to the twilight zone. As a rule-of-thumb an E-value above 10-4 (0.0001) is not necessarily interesting. If you want to be certain of the homology, your E-value must be lower than 10^{-4}

Understanding your BLAST output: 3. Alignment



```
>sp|P29045|RRPO_BYDVR Putative RNA-directed RNA polymerase (EC
          2.7.7.48) [Contains: 39 kDa protein].[Barley yellow
          dwarf virus]
         Length = 867

 Score = 1635 bits (4234), Expect = 0.0
 Identities = 821/867 (94%), Positives = 828/867 (94%)

Query: 1   MFFEILIGASAKAVKDFISHCYSRLKSIYYSFKRWLMEISGQFKAHDAFVNMCFGHMADI 60
            MFFEILIGASAKAVKDFISHCYSRLKSIYYSFKRWLMEISGQFKAHDAFVNMCFGHMADI
Sbjct: 1   MFFEILIGASAKAVKDFISHCYSRLKSIYYSFKRWLMEISGQFKAHDAFVNMCFGHMADI 60

Query: 61  XXXXXXXXXXXXXXXXXXXXXXSLLKLLVAQKSKSGVTEAWTDFFTKSRGGVYAPLSCEP 120
                                  SLLKLLVAQKSK+GVTEAWTDFFTKSRGGVYAPLSCEP
Sbjct: 61  EDFEAELAEEFAEREDEVEEARSLLKLLVAQKSKTGVTEAWTDFFTKSRGGVYAPLSCEP 120

Query: 121 TRQELEVKSEKLERLLEEQHQFEVRAAKKYIKEKGRGFINCWNDLRSRLRLVKDVKDEAK 180
            TRQELE KSEKLE+LLEEQHQFEVRAAKKYIKEKGRGFINCWNDLRSRLRLVKDVKDEAK
Sbjct: 121 TRQELEAKSEKLEKLLEEQHQFEVRAAKKYIKEKGRGFINCWNDLRSRLRLVKDVKDEAK 180
```

Length
of the alignment

Percent identity
25% is good news

XXX: low
complexity regions
masked

Positives
fraction of residues that
are either identical or similar

mismatch

identical aa

similar aa

A good alignment should not contain too many gaps and should have a few patches of high similarity, rather than isolated identical residues spread here and there

BLASTing DNA sequences

BLASTing DNA sequences

- BLASTing DNA requires operations similar to BLASTing proteins BUT does not always work so well.

- It is faster and more accurate to BLAST proteins (blastp) rather than nucleotides. If you know the reading frame in your sequence, you' re better off translating the sequence and BLASTing with a protein sequence.

- Otherwise:

Different BLAST Programs Available for DNA Sequences			
Program	***Query***	***Database***	***Usage***
blastn	DNA	DNA	Very similar DNA sequences
tblastx	**T**DNA	TDNA	Protein discovery and ESTs
blastx	**T**DNA	Protein	Analysis of the query DNA sequence

T= translated

BLASTing DNA sequences: choosing the right BLAST

Question	***Answer***
Am I interested in non-coding DNA?	Yes: Use **blastn**. Never forget that blastn is only for closely related DNA sequences (more than 70% identical)
Do I want to discover new proteins?	Yes: Use **tblastx**.
Do I want to discover proteins encoded in my query DNA sequence?	Yes: Use **blastx**.
Am I unsure of the quality of my DNA?	Yes: Use **blastx** if you suspect your DNA sequence is the coding for a protein but it may contain sequencing errors.

- Pick the right database: choose the database that's compatible with the BLAST program you want to use

- Restrict your search: Database searches on DNA are slower. When possible, restrict your search to the subset of the database that you're interested in (e.g. only the Drosophila genome)

- Shop around: Find the BLAST server containing the database that you're interested in

- Use filtering: Genomic sequences are full of repetitions: use some filtering

Choosing the Right Parameters

Choosing the right Parameters

- The default parameters that BLAST uses are quite optimal and well tested. However for the following reasons you might want to change them:

Some Reasons to Change BLAST Default Parameters	
Reason	***Parameters to Change***
The sequence you're interested in contains many identical residues; it has a biased composition.	Sequence filter (automatic masking)
BLAST doesn't report any results	Change the substitution matrix or the gap penalties.
Your match has a borderline E-value	Change the substitution matrix or the gap penalties to check the match robustness.
BLAST reports too many matches	Change the database you're searching OR filter the reported entries by keyword OR increase the number of reported matches OR increase Expect, the E-value threshold.

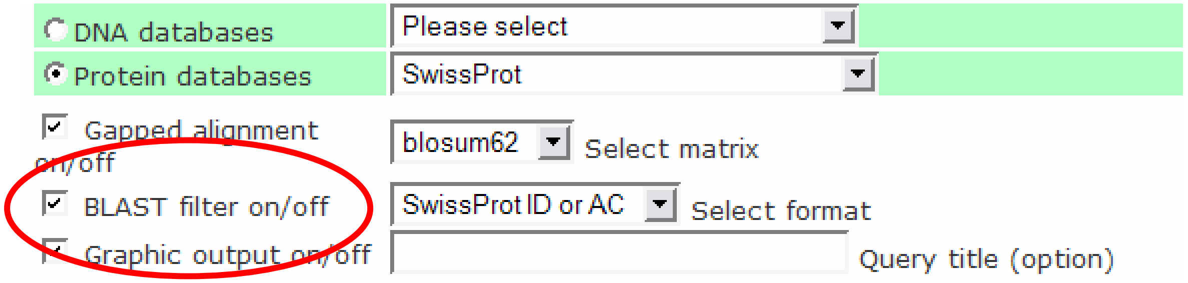
Choosing the right Parameters: sequence masking

- When BLAST searches databases, it makes the assumption that the average composition of any sequence is the same as the average composition of the whole database.

- However this assumption doesn't hold all the time, some sequences have biased compositions, e.g. many proteins contain patches known as low-complexity regions: such as segments that contain many prolines or glutamic acid residues.

- If BLAST aligns two proline-rich domains, this alignment gets a very good E-value because of the high number of identical amino acids it contains. BUT there is a good chance that these two proline-rich domains are not related at all.

- In order to avoid this problem, sequence masking can be applied.



○ DNA databases — Please select

● Protein databases — SwissProt

☑ Gapped alignment on/off — blosum62 — Select matrix

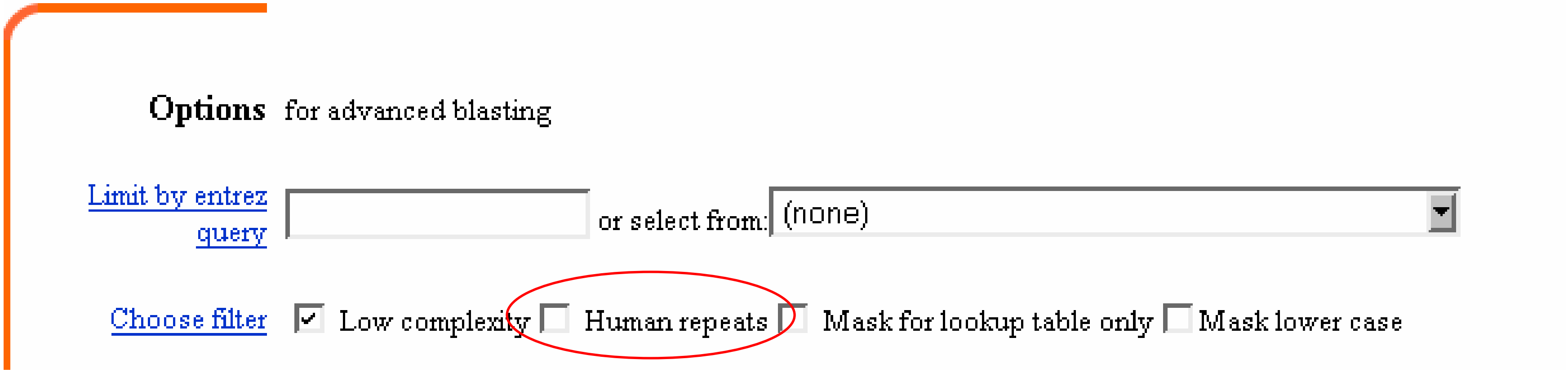
☑ BLAST filter on/off — SwissProt ID or AC — Select format

☑ Graphic output on/off — Query title (option)

Choosing the right Parameters: DNA masking

- DNA sequences are full of sequences repeated many times: most of genomes contain many such repeats, especially the human genome (60% are repeats).

- If you want to avoid the interference of that many repeats, select the Human Repeats check box that appears in the blastn page of NCBI.

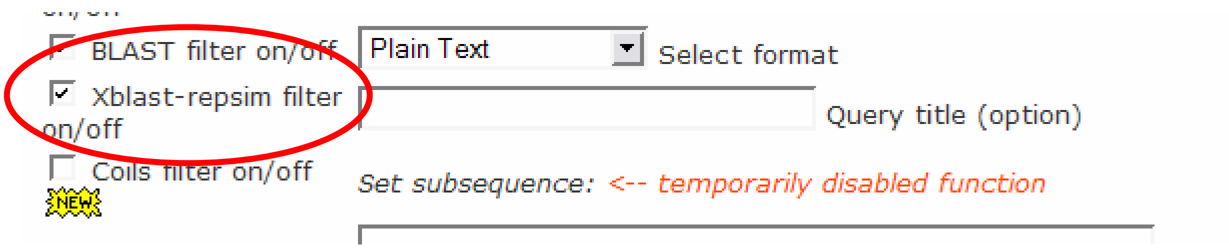


Options for advanced blasting

Limit by entrez query [   ] or select from: (none) ▾

Choose filter ☑ Low complexity □ Human repeats □ Mask for lookup table only □ Mask lower case

- Or at the swiss EMBnet server (advanced BLAST):



on/off

☑ BLAST filter on/off

☑ Xblast-repsim filter on/off

□ Coils filter on/off

Plain Text ▾ Select format

[   ] Query title (option)

Set subsequence: <-- temporarily disabled function

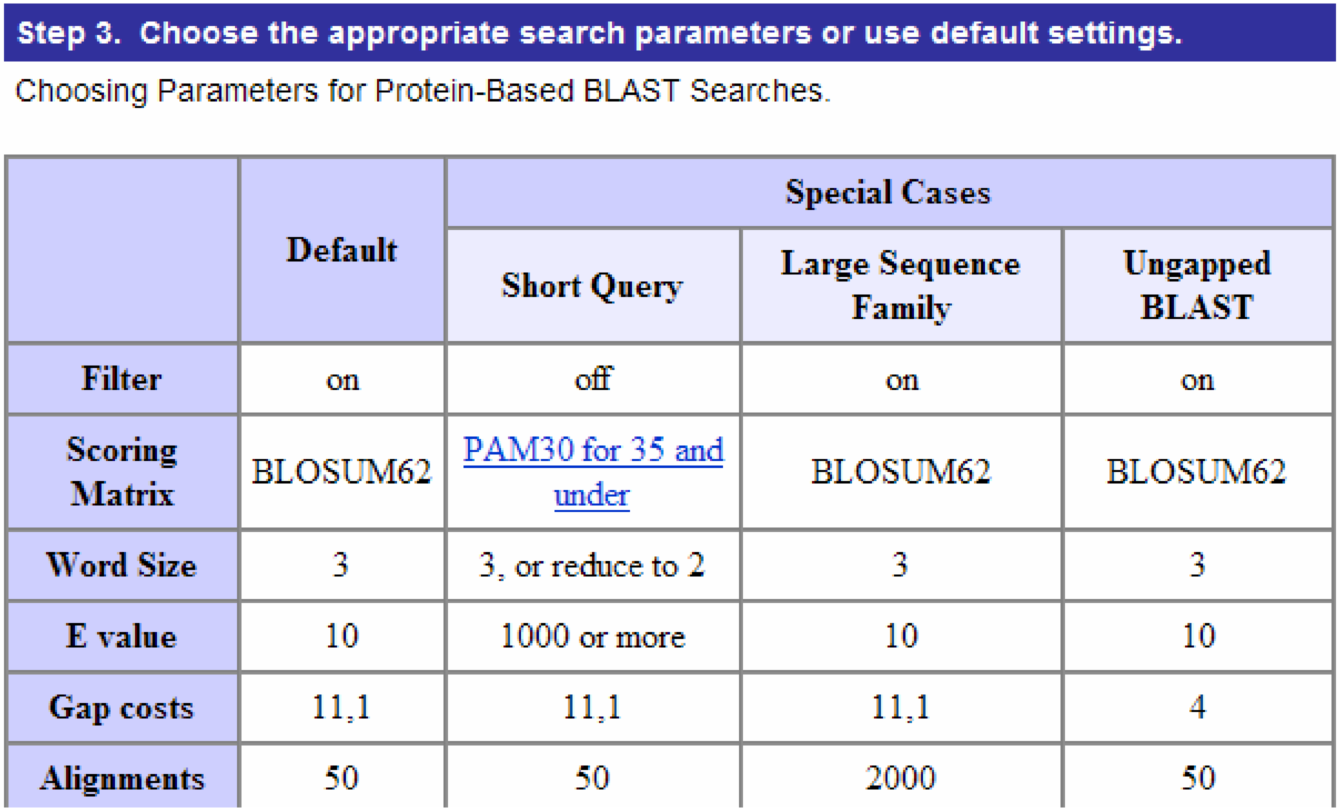
Changing the BLAST alignment parameters

- Among the parameters that you can change on the NCBI BLAST server two important ones have to do with the way BLAST makes the alignments: the gap penalites (gap costs) and the substitution matrix (matrix).

- The best reason to play with them is to check the robustness of a hit that's borderline. If this match does not go away when you change the substitution matrix or the gap penalties, then it has better chances of being biologically meaningful

Changing the BLAST alignment parameters

- Guidelines from BLAST tutorial at NCBI



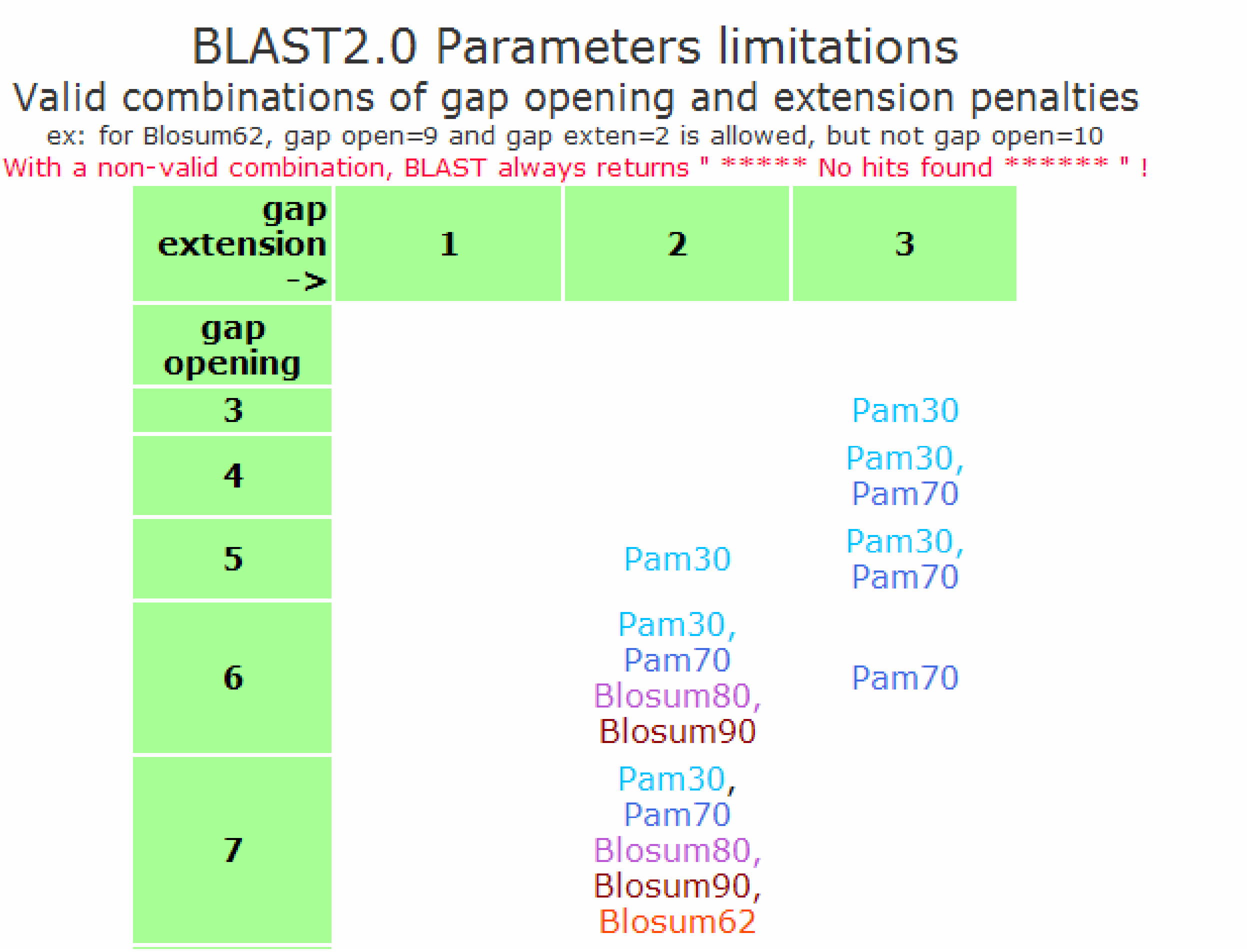
Step 3. Choose the appropriate search parameters or use default settings.

Choosing Parameters for Protein-Based BLAST Searches.

	Default	Special Cases		
		Short Query	Large Sequence Family	Ungapped BLAST
Filter	on	off	on	on
Scoring Matrix	BLOSUM62	PAM30 for 35 and under	BLOSUM62	BLOSUM62
Word Size	3	3, or reduce to 2	3	3
E value	10	1000 or more	10	10
Gap costs	11,1	11,1	11,1	4
Alignments	50	50	2000	50

Changing the BLAST alignment parameters

- Guidelines from BLAST tutorial at the swiss EMBnet server



BLAST2.0 Parameters limitations

Valid combinations of gap opening and extension penalties

ex: for Blosum62, gap open=9 and gap exten=2 is allowed, but not gap open=10

With a non-valid combination, BLAST always returns " ***** No hits found ****** " !

gap extension ->	1	2	3
gap opening			
3			Pam30
4			Pam30, Pam70
5		Pam30	Pam30, Pam70
6		Pam30, Pam70 Blosum80, Blosum90	Pam70
7		Pam30, Pam70 Blosum80, Blosum90, Blosum62	

Controlling the BLAST output

- If your query belongs to a large protein family, the BLAST output may give you troubles because the databases contain too many sequences nearly identical to yours => preventing you from seeing a homologous sequence less closely related but associated with experimental information; so how to proceed?

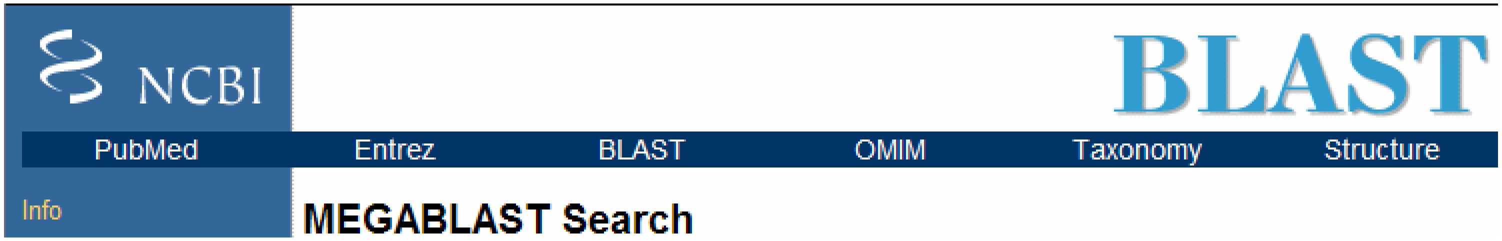
1) Choosing the right database
If BLAST reports too many hits, search for Swiss-Prot (100 times smaller) rather than NR; or search only one genome

2) Limit by Entrez query (NCBI)
For instance, if you want BLAST to report proteases only and to ignore proteases from the HIV virus, type "protease NOT hiv1[Organism]"

3) Expect
Change the cutoff for reporting hits, to force BLAST to report only good hits with a low cutoff

BLAST Family

- Faster algorithm for genomic search: MEGABLAST (NCBI) and SSAHA (Ensembl): This program is optimized for aligning sequences that differ slightly as a result of sequencing or other similar "errors". (larger word size is used as default)



MEGABLAST Search

- PSI-BLAST and PHI-BLAST-> tomorrow

Acknowledgments

Frédérique Galisson, for the pictures about the FASTA and BLAST algorithms.

References

- David W. Mount, Bioinformatics, Cold Spring Harbor Laboratory Press
- Jean-Michel Claverie & Cedric Notredame, Bioinformatics for Dummies, Wiley Publishing