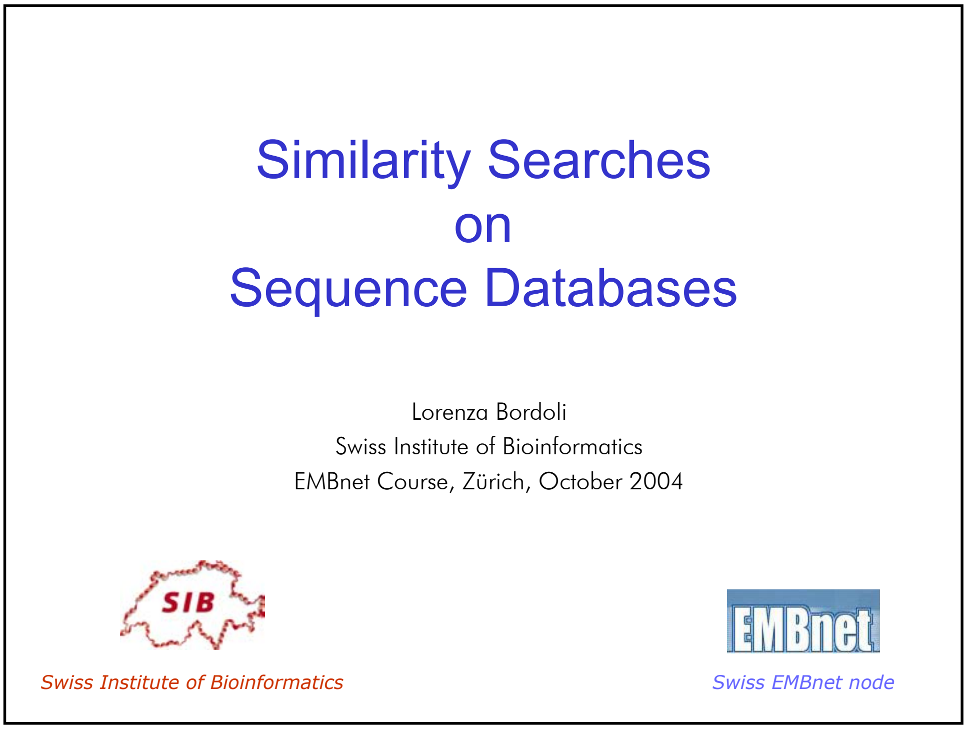
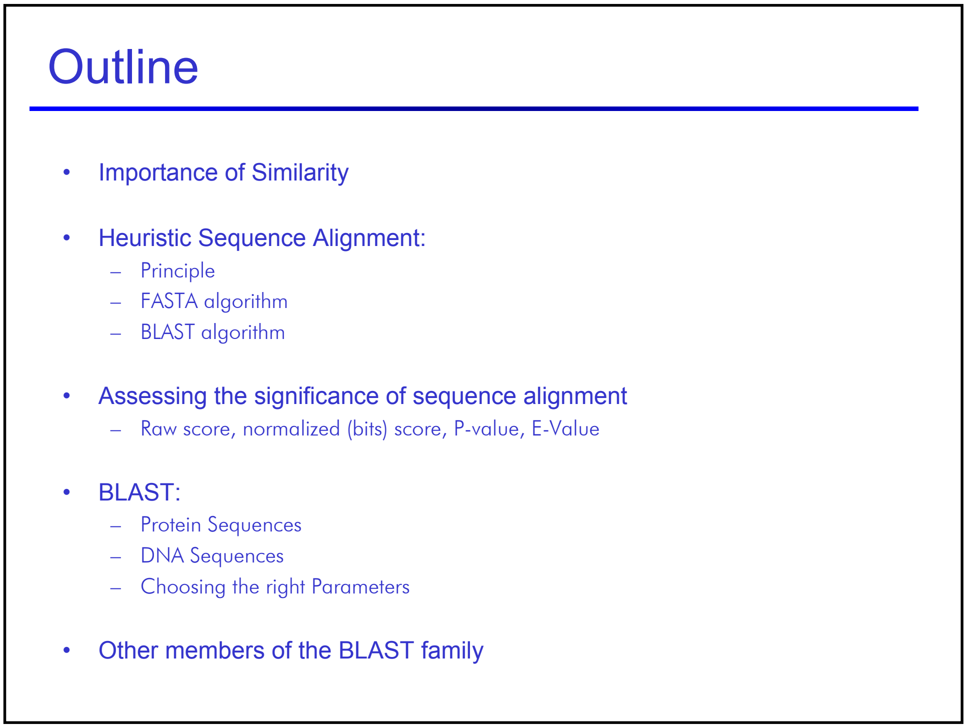


Similarity Searches on Sequence Databases

Lorenza Bordoli
Swiss Institute of Bioinformatics
EMBnet Course, Zürich, October 2004

Swiss Institute of Bioinformatics

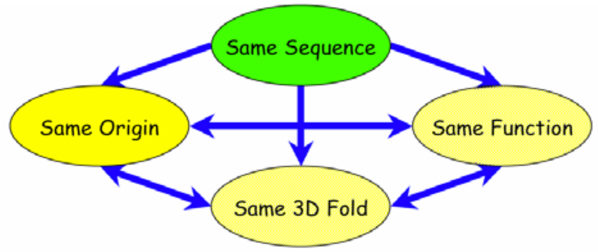
Swiss EMBnet node



Outline

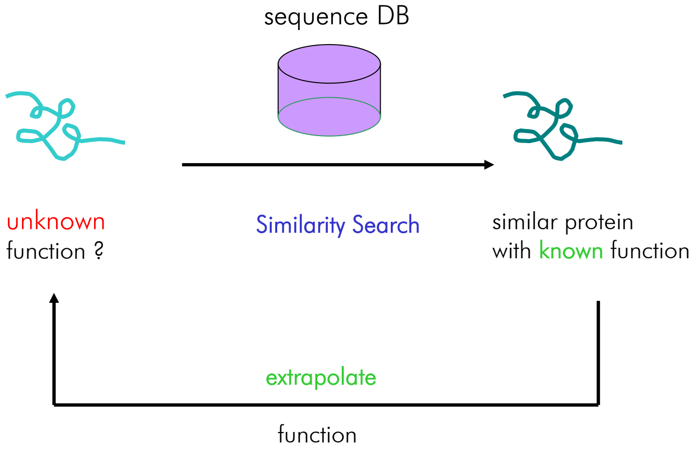
- Importance of Similarity
- Heuristic Sequence Alignment:
  - Principle
  - FASTA algorithm
  - BLAST algorithm
- Assessing the significance of sequence alignment
  - Raw score, normalized (bits) score, P-value, E-Value
- BLAST:
  - Protein Sequences
  - DNA Sequences
  - Choosing the right Parameters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

Importance of Similarity



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

Importance of Similarity



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.

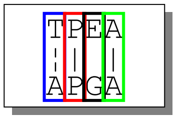
Beware:
- E-value (*Expectation value*)
- Length of the segments similar between the two sequences
- The number of insertions/deletions

Alignment score

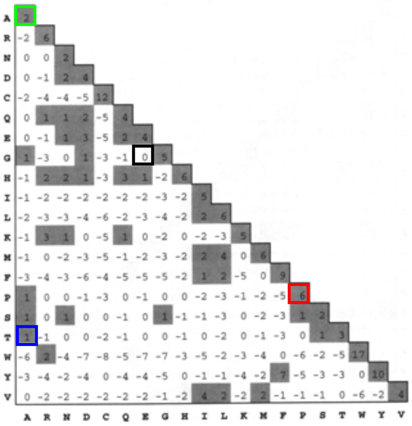
Amino acid substitution matrices

- Example: PAM250
- Most used: Blosum62

Raw score of an alignment



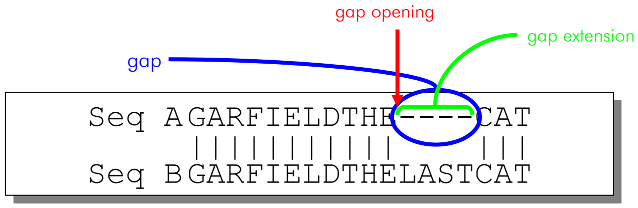
Score = 1 + 6 + 0 + 2 = 9



	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2																			
R	-2	6																		
N	0	0	2																	
D	0	-1	2	4																
C	-2	-4	-4	-5	12															
Q	0	1	1	2	-5	4														
E	0	-1	1	3	-5	2	4													
G	1	-3	0	1	-3	-1	0	5												
H	-1	2	2	1	-3	3	1	-2	6											
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5										
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6									
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5								
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6							
F	-3	-4	-3	-6	-4	-5	-5	-5	-2	1	2	-5	0	9						
P	1	0	0	-1	-3	0	-1	0	0	-2	-3	-1	-2	-5	6					
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2				
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	3			
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17		
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4

Insertions and deletions

Gap penalties



```
Seq A GARFIELDTHE----CAT
      |||||||||||    |||
Seq B GARFIELDTHELASTCAT
```

- Opening a gap penalizes an alignment score
- Each extension of a gap penalizes the alignment's score
- The gap opening penalty is in general higher than the gap extension penalties (simulating evolutionary behavior)

- The raw score of a gapped alignment is the sum of all amino acid substitutions from which we subtract the gap opening and extension penalties.

Alignment

Alignement types:

- Global Alignment between the complete sequence A and the complete sequence B
- Local Alignment between a sub-sequence of A and a sub-sequence of B

Computer implementation (Algorithms):

Dynamic programing (exact algorithm)

- Global Needleman-Wunsch
- Local Smith-Waterman

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

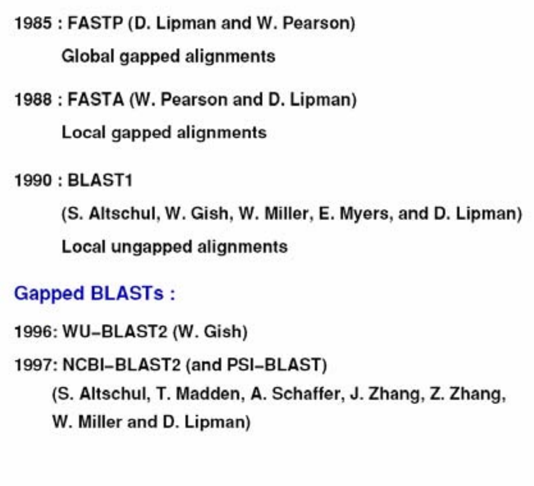
=>Restricting the alignment process:

- 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 (But good ones often do).
- Advantage: This methods that are least 50-100 times faster than dynamic programming therefore better suited to search databases.

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: algorithm (4 steps)

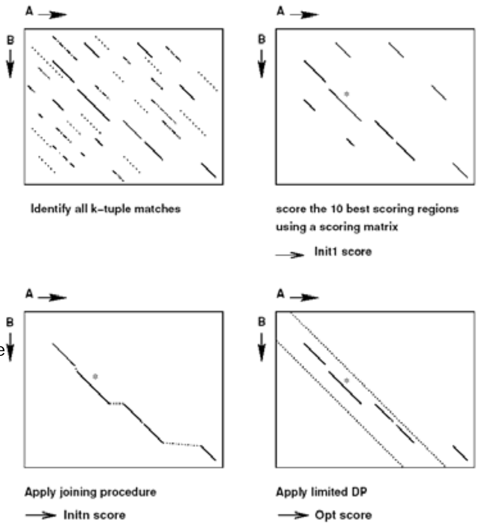
Localize the 10 best regions of similarity between the two seq. Each identity between two "word" is represented by a dot

Each diagonal: ungapped alignment

The smaller the k, The sensitive the method but slower

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

DP applied around the best scoring diagonal.



Identify all k–tuple matches

score the 10 best scoring regions using a scoring matrix → Init1 score

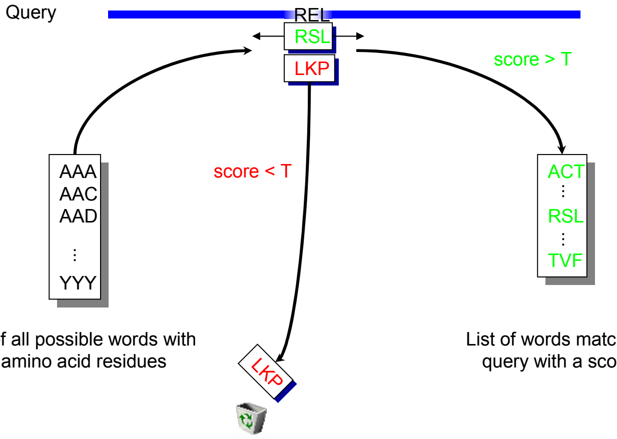
Apply joining procedure → Initn score

Apply limited DP → Opt score

BLAST: algorithm

1. Blast algorithm: creating a list of similar words

⇨A substitution matrix is used to compute the word scores

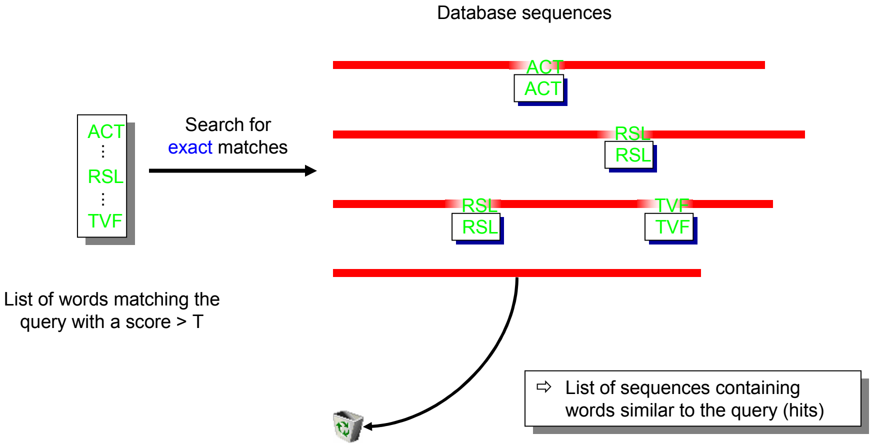


List of all possible words with 3 amino acid residues

List of words matching the query with a score > T

BLAST: algorithm

2. Blast algorithm: eliminating sequences without word hits



Database sequences

Search for exact matches

List of words matching the query with a score > T

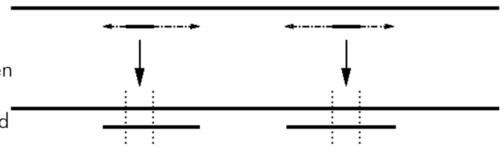
⇨ List of sequences containing words similar to the query (hits)

BLAST: Algorithm

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.

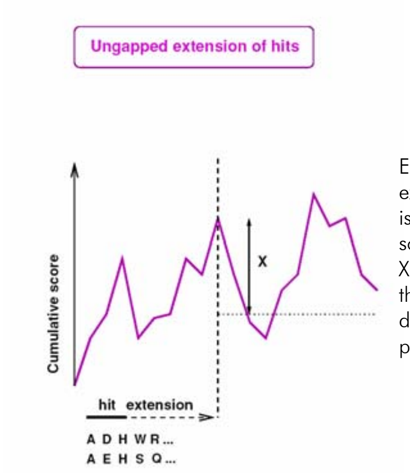
Each match is then extended. The extension is stopped as soon as the score decreases more then X when compared with the highest value obtained during the extension process



HSP = High Scoring Segment Pair

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

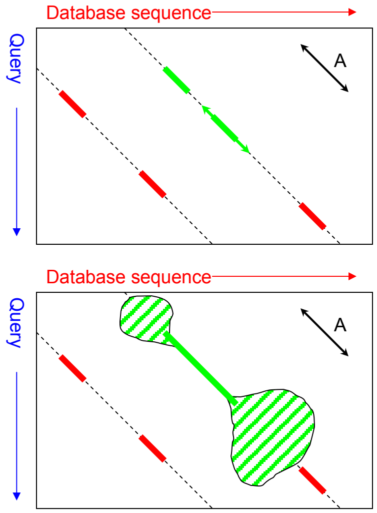
BLAST: Algorithm



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

BLAST: algorithm

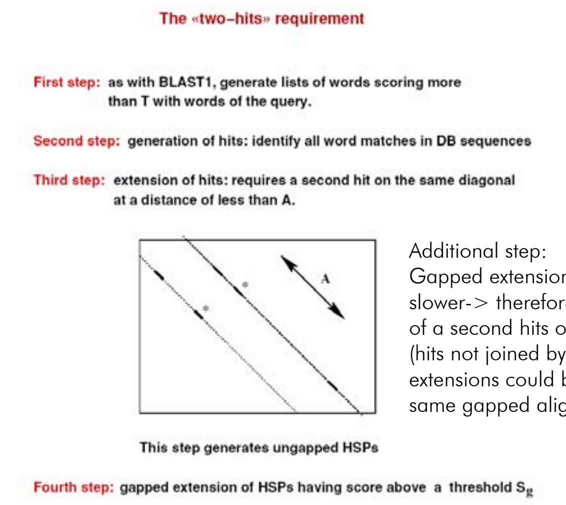
3. Blast algorithm: extension of hits



Ungapped extension if:
- 2 "Hits" are on the same diagonal but at a distance less than A

Extension using dynamic programming
- limited to a restricted region

BLAST: Algorithm



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

- Scoring System:
  - 1. Scoring (Substitution) matrix (or match mismatch for DNA): 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 (bit) score to compare different alignments, based on different scoring systems, e.g. different substitution matrices.

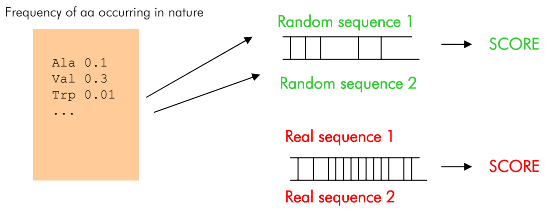
2. A method to asses the statistical significance of the alignment is needed (is an alignment biological relevant?) : E-value

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

Assessing the significance of sequence alignment

Statistics derived from the scores:

- P-value
  - ⇨ Probability that an alignment with this score occurs by chance in a database of this size
  - ⇨ The closer the P-value is towards 0, the better the alignment

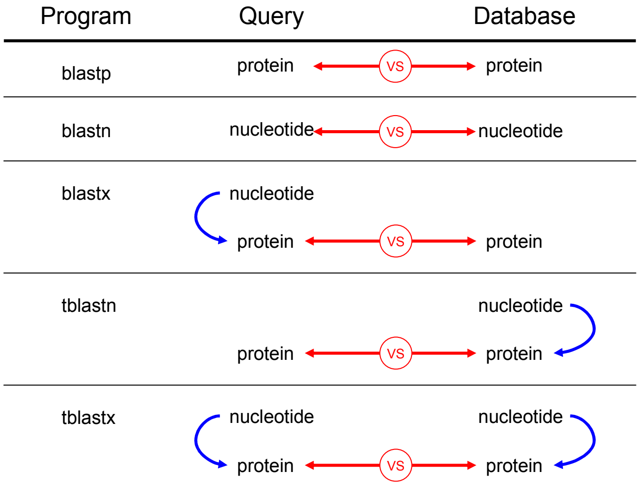
- E-value
  - ⇨ Number of matches with this score one can expect to find by chance in a database of size N
  - ⇨ The closer the e-value is towards 0, the better the alignment
- Relationship between E-value and P-value:
  - ⇨ In a database containing *N* sequences
    E = P x *N*

BLAST

Basic Local Alignment Search Tool

A Blast for each query

Different programs are available according to the type of query



Program	Query		Database
blastp	protein	vs	protein
blastn	nucleotide	vs	nucleotide
blastx	nucleotide protein	vs	protein
tblastn	protein	vs	nucleotide protein
tblastx	nucleotide protein	vs	nucleotide protein

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; identify common regions between proteins, or collect related proteins (phylogenetic analysis;

tblastn = Compares a protein sequence with a nucleotide database

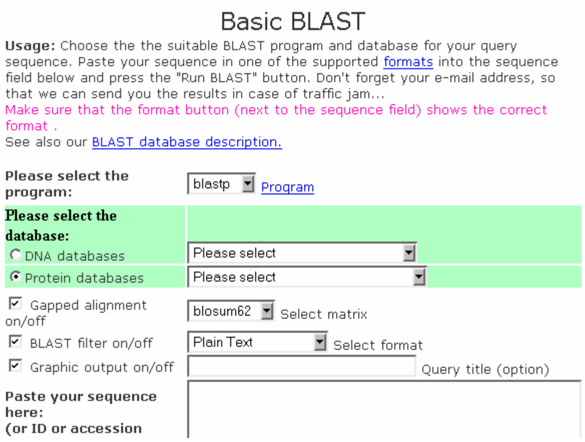
If you want to discover new genes encoding proteins (from multiple organisms), use **tblastn** to compare your protein with DNA sequences translated into their six possible reading frames; map a protein to genomic DNA;

BLASTing protein sequences

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

- NCBI (National Center for Biotechnology Information) server: http://www.ncbi.nlm.nih.gov/BLAST
- ExPASy server: http://www.expasy.org/tools/blast/
- Swiss EMBnet server (European Molecular Biology network):
  http://www.ch.embnet.org/software/bBLAST.html (basic)
  http://www.ch.embnet.org/software/aBLAST.html (advanced)

BLASTing protein sequences: Swiss EMBnet blastp server



Basic BLAST

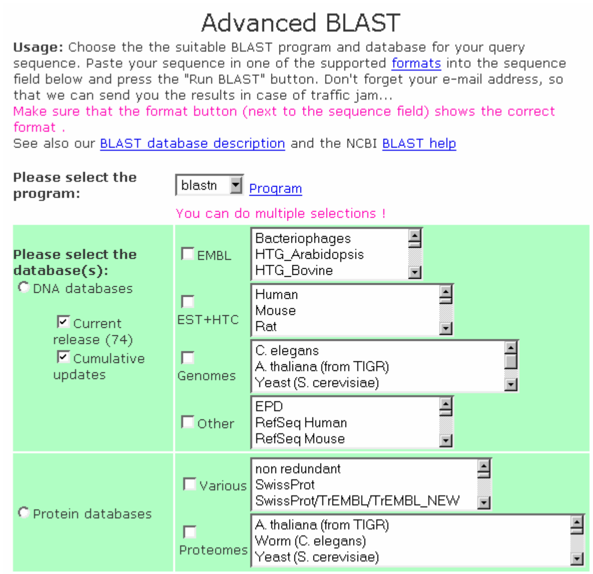
Usage: Choose the the suitable BLAST program and database for your query sequence. Paste your sequence in one of the supported formats into the sequence field below and press the "Run BLAST" button. Don't forget your e-mail address, so that we can send you the results in case of traffic jam...

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

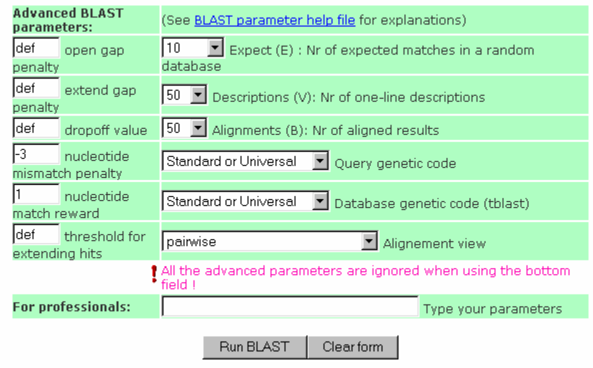
Please select the program: blastn Program

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)

BLASTing protein sequences: Swiss EMBnet blasp server

- Greater choice of databases to search
- Advanced Blast parameter modification



Advanced BLAST parameters:	(See BLAST parameter help file for explanations)
def open gap penalty	10 Expect (E) : Nr of expected matches in a random database
def extend gap penalty	50 Descriptions (V): Nr of one-line descriptions
def dropoff value	50 Alignments (B): Nr of aligned results
-3 nucleotide mismatch penalty	Standard or Universal Query genetic code
1 nucleotide match reward	Standard or Universal Database genetic code (tblast)
def threshold for extending hits	pairwise Alignement view
	! All the advanced parameters are ignored when using the bottom field !
For professionals:	Type your parameters

Run BLAST Clear form

Understanding your BLAST output

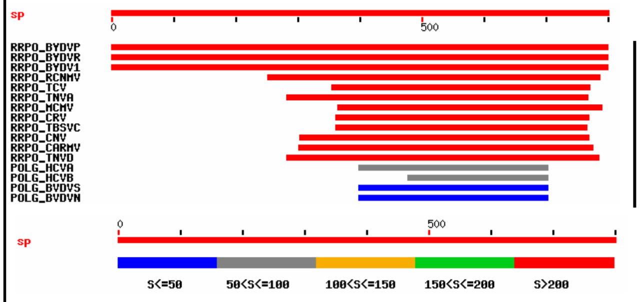
1. Graphic display:
   shows you where your query is similar to other sequences

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



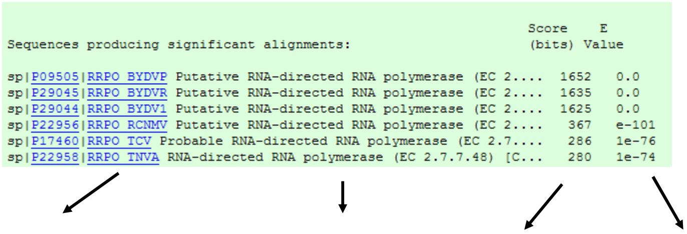
query sequence

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
```

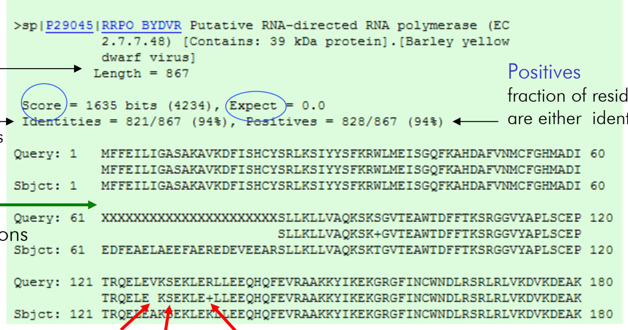
Sequence ac number and name Description Bit score E-value

- 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

Positives
fraction of residues that are either identical or similar

XXX: low complexity regions masked

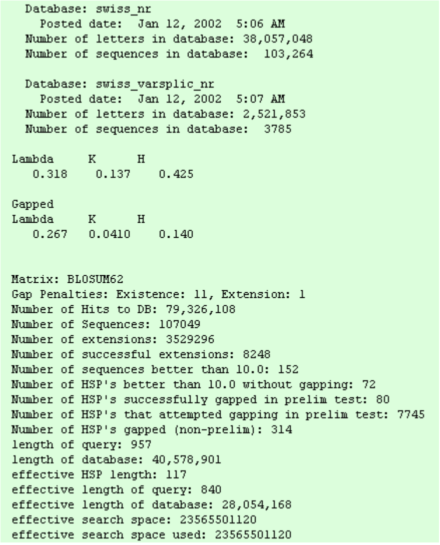
mismatch

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

Understanding your BLAST output: 4. Parameters



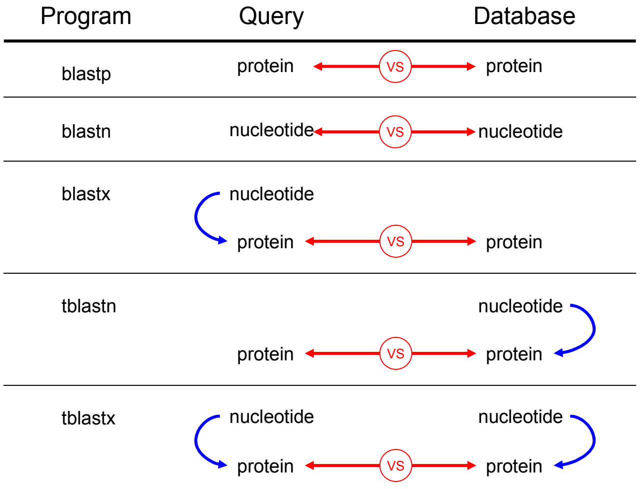
```
  Database: swiss_nr
    Posted date:  Jan 12, 2002  5:06 AM
  Number of letters in database: 38,057,048
  Number of sequences in database:  103,264

  Database: swiss_varsplic_nr
    Posted date:  Jan 12, 2002  5:07 AM
  Number of letters in database: 2,521,853
  Number of sequences in database:  3785

Lambda     K      H
   0.318    0.137    0.425

Gapped
Lambda     K      H
   0.267   0.0410    0.140


Matrix: BLOSUM62
Gap Penalties: Existence: 11, Extension: 1
Number of Hits to DB: 79,326,108
Number of Sequences: 107049
Number of extensions: 3529296
Number of successful extensions: 8248
Number of sequences better than 10.0: 152
Number of HSP's better than 10.0 without gapping: 72
Number of HSP's successfully gapped in prelim test: 80
Number of HSP's that attempted gapping in prelim test: 7745
Number of HSP's gapped (non-prelim): 314
length of query: 957
length of database: 40,578,901
effective HSP length: 117
effective length of query: 840
effective length of database: 28,054,168
effective search space: 23565501120
effective search space used: 23565501120
```

Search details (at the bottom of the results)

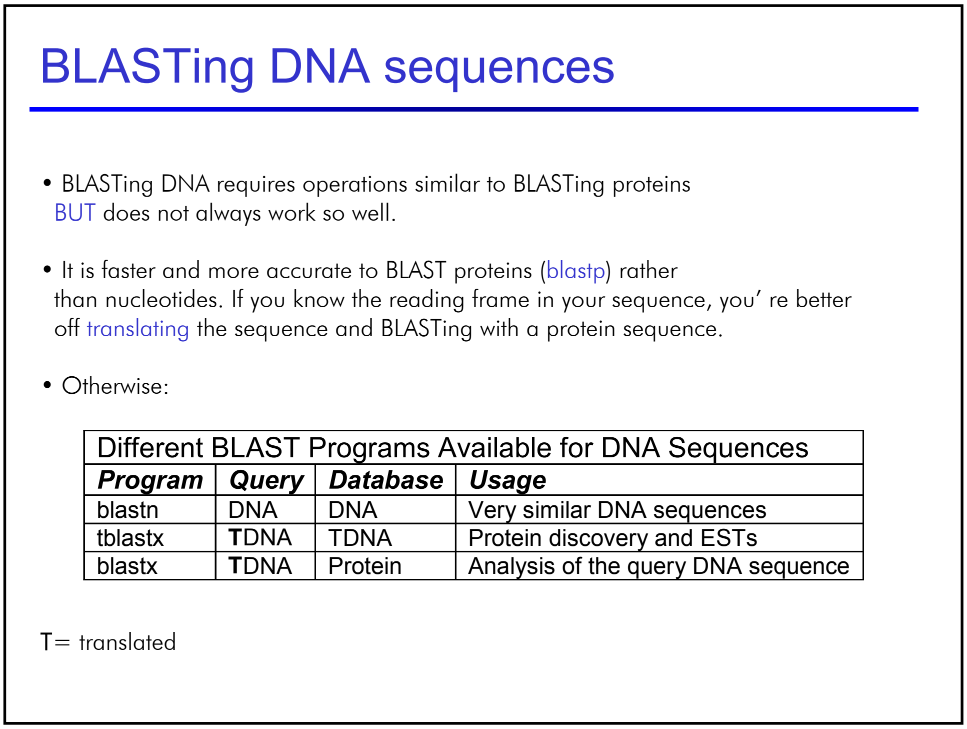
- Size of the database searched
- Scoring system parameters
- Details about the number of hits found

A Blast for each query

Different programs are available according to the type of query



Program	Query		Database
blastp	protein	vs	protein
blastn	nucleotide	vs	nucleotide
blastx	nucleotide protein	vs	protein
tblastn	protein	vs	nucleotide protein
tblastx	nucleotide protein	vs	nucleotide protein



BLASTing DNA sequences

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

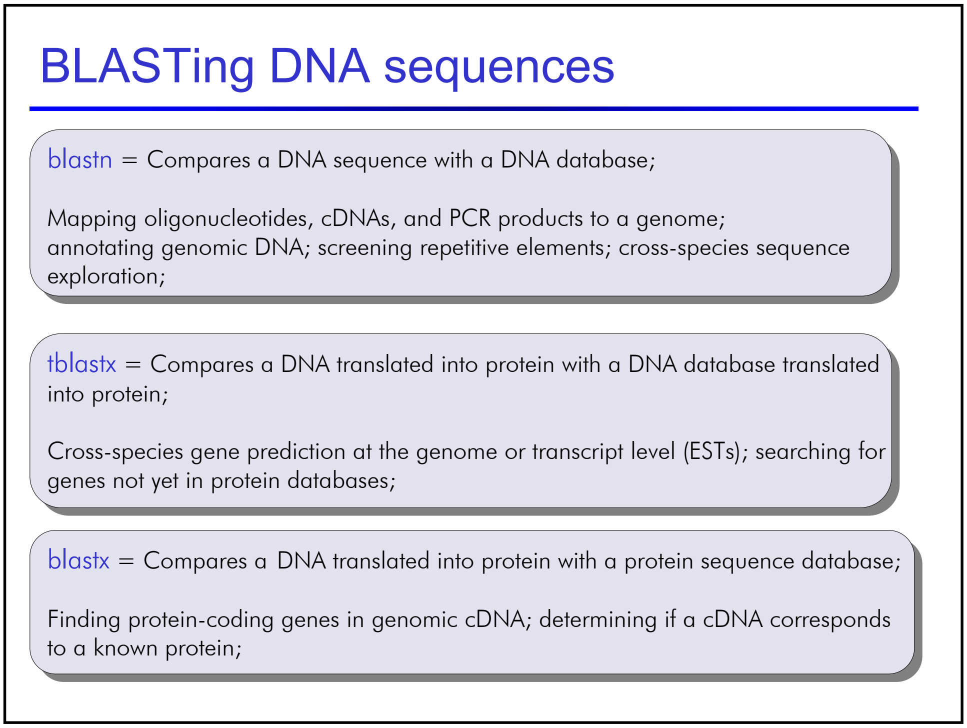
- It is faster and more accurate to BLAST proteins (blastp) rather than nucleotides. If you know the reading frame in your sequence, you' re better off translating the sequence and BLASTing with a protein sequence.

- 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

blastn = Compares a DNA sequence with a DNA database;

Mapping oligonucleotides, cDNAs, and PCR products to a genome; annotating genomic DNA; screening repetitive elements; cross-species sequence exploration;

tblastx = Compares a DNA translated into protein with a DNA database translated into protein;

Cross-species gene prediction at the genome or transcript level (ESTs); searching for genes not yet in protein databases;

blastx = Compares a DNA translated into protein with a protein sequence database;

Finding protein-coding genes in genomic cDNA; determining if a cDNA corresponds to a known protein;

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 (in general!)

- 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

BLASTting DNA: BLASTN output

- DNA double-stranded molecule => genes may occur on either strand
- *plus* strand (the query sequence), *minus* strand (reverse complement)
- If the similarity between query and subject is on the same strand: *plus/plus*
- If the minus strand of the query sequence is similar to a database sequence: *plus/minus* with the subject sequence in reverse coordinates (flipped)

```
Score = 87.7 bits (44), Expect = 2e-15 Identities = 57/60 (95%), Gaps = 1/60 (1%)
Strand = Plus / Plus
Query: 1     ggtggtttagaacgatctggtcttaccctgctaccaactgttcatcggttattgttggag 60
             |||| ||||||||||| ||||||||||| |||||||||||||||||||||||||||||||
Sbjct: 96694 ggtgttttagaacgat-tggtcttacccggctaccaactgttcatcggttattgttggag 96752


Score = 52.0 bits (26), Expect = 1e-04 Identities = 26/26 (100%)
Strand = Plus / Minus
Query: 18    tggtcttaccctgctaccaactgttc 43
             ||||||||||||||||||||||||||
Sbjct: 40758 tggtcttaccctgctaccaactgttc 40733
```

BLASTting DNA: BLASTX output

- Query sequence: translated in the 3 reading frames, on both *plus* and *minus* strand: +1,+2,+3 (plus strand) and -1, -2, -3 (minus strand)
- Matches on the plus strand: +1,+2,+3
- Matches on the minus strand: query coordinates are inverted

```
Score = 790 bits (2040), Expect = 0.0
Identities = 520/1381 (37%), Positives = 745/1381 (53%), Gaps = 36/1381 (2%)
Frame = +3
Query: 156 SEMNVNMKYQLPNFTAETPIQNVVLHKHH--IYLGAVNYIYVLNDKDLQKVAEYKTGPVL 329
           S +N ++ Y +P F A PIQN+V + + +Y+ + N I +N + L+KV E +TGPV
Sbjct: 31  SPVNFSVVYTMPFFQAGGPIQNIVNNSFYQEVYVASQNVIEAVN-QSLEKVWELRTGPV- 88
```

```
Score = 64.5 bits (169), Expect = 1.7e-258
Identities = 30/34 (88%), Positives = 34/34 (100%), Gaps = 3/34 (2%)
Frame = -1
Query: 1071 SEMNVNMKYQLPNFTAETPIQNVVLHKHH--IYLGAVNYIYVLNDKDLQKVAEYKTGPVL 970
            S +N ++ Y +P F A PIQN+V + + +Y+ + N I +N + L+KV E +TGPV
Sbjct: 722  SPVNFSVVYTMPFFQAGGPIQNIVNNSFYQEVYVASQNVIEAVN-QSLEKVWELRTGPV- 755
```

BLASTting DNA: TBLASTN output

- Alignments similar to BLASTX, except that the database and query are exchanged (e.g. on minus strand the database sequence has flipped coordinates)

```
Score = 47.8 bits (112), Expect = 5e-04
Identities = 20/21 (95%), Positives = 21/21 (99%)
Frame = +2
Query:       1 SQITRIPLNGLGCEHFQSCSQ 21
               SQIT+IPLNGLGCEHFQSCSQ
Sbjct: 108872  SQITKIPLNGLGCEHFQSCSQ 108934
```

```
Score = 45.8 bits (107), Expect = 0.002
Identities = 19/21 (90%), Positives = 20/21 (94%)
Frame = -2
Query:      1 SQITRIPLNGLGCEHFQSCSQ 21
              SQIT+IPLNGLGC HFQSCSQ
Sbjct: 28239  SQITKIPLNGLGCRHFQSCSQ 28177
```

BLASTting DNA: TBLASTX output

- Both query and database have strand and frame
- Alignments may have any combination of frames

```
Score = 790 bits (2040), Expect = 0.0
Identities = 520/1381 (37%), Positives = 745/1381 (53%), Gaps = 36/1381 (2%)
Frame = +3/+3
Query: 156 SEMNVNMKYQLPNFTAETPIQNVVLHKHH--IYLGAVNYIYVLNDKDLQKVAEYKTGPVL 329
           S +N ++ Y +P F A PIQN+V + + +Y+ + N I +N + L+KV E +TGPV
Sbjct: 31  SPVNFSVVYTMPFFQAGGPIQNIVNNSFYQEVYVASQNVIEAVN-QSLEKVWELRTGPV- 88
```

```
Score = 64.5 bits (169), Expect = 1.7e-258
Identities = 30/34 (88%), Positives = 34/34 (100%), Gaps = 3/34 (2%)
Frame = -1/+2
Query: 1071 SEMNVNMKYQLPNFTAETPIQNVVLHKHH--IYLGAVNYIYVLNDKDLQKVAEYKTGPVL 970
            S +N ++ Y +P F A PIQN+V + + +Y+ + N I +N + L+KV E +TGPV
Sbjct: 722  SPVNFSVVYTMPFFQAGGPIQNIVNNSFYQEVYVASQNVIEAVN-QSLEKVWELRTGPV- 755
```

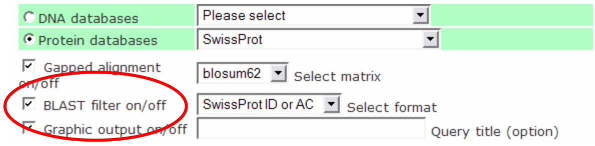
Choosing the right Parameters

- 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

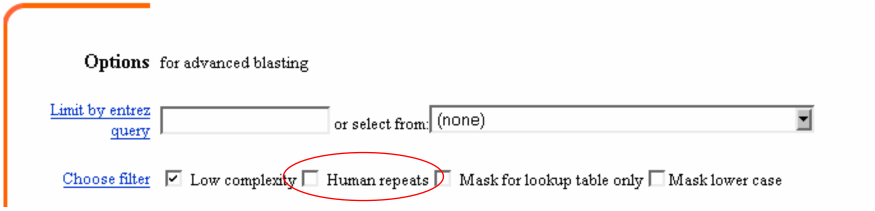
- When BLAST searches databases, it makes the assumption that the average composition of any sequence is the same as the average composition of the whole database.
- However this assumption doesn't hold all the time, some sequences have biased compositions, e.g. many proteins contain patches known as 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 or the Xblast-repsim filter

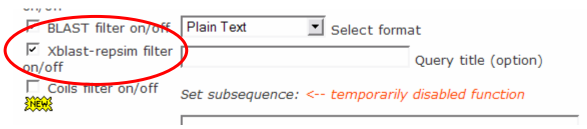


Options for advanced blasting

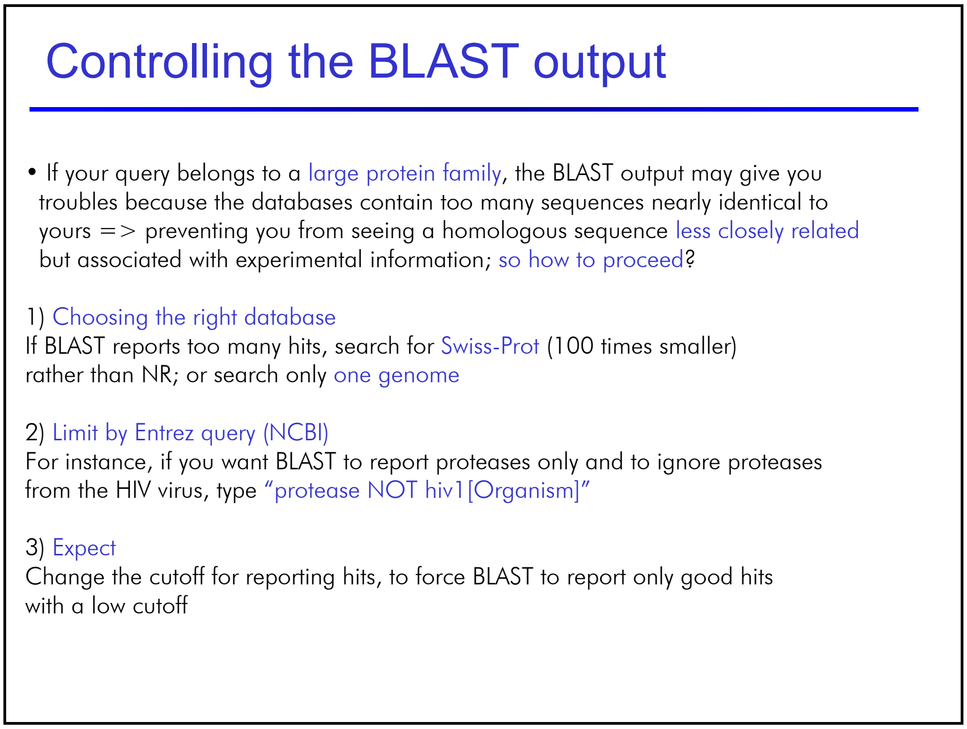
Limit by entrez query — or select from: (none)

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

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



☑ BLAST filter on/off — Plain Text Select format
☑ Xblast-repsim filter on/off — Query title (option)
☐ Coils filter on/off — *Set subsequence: <-- temporarily disabled function*



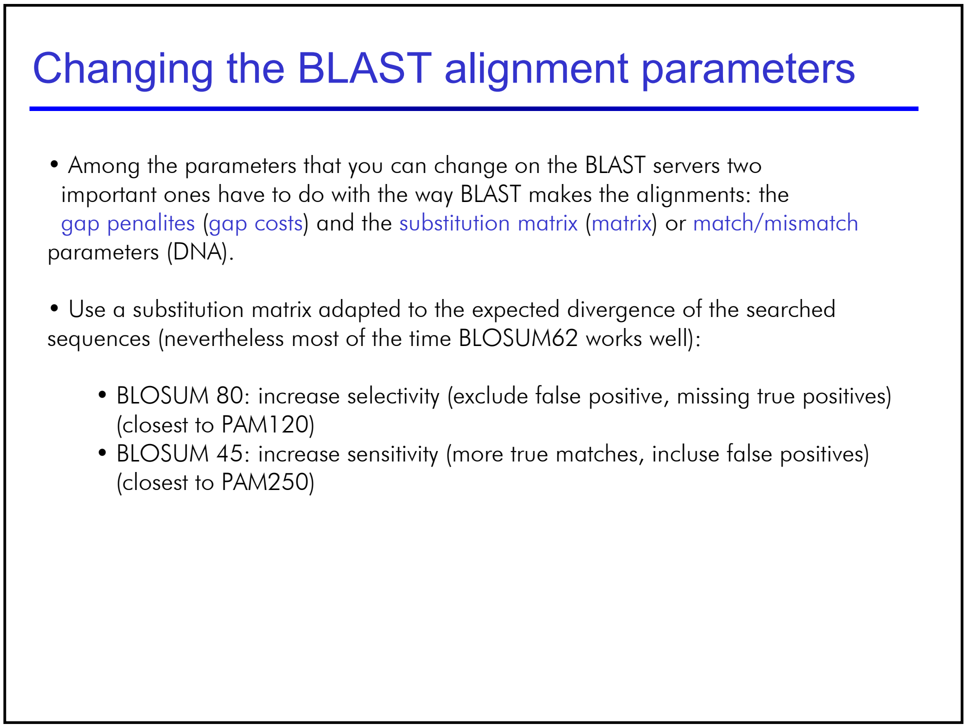
Controlling the BLAST output

- 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

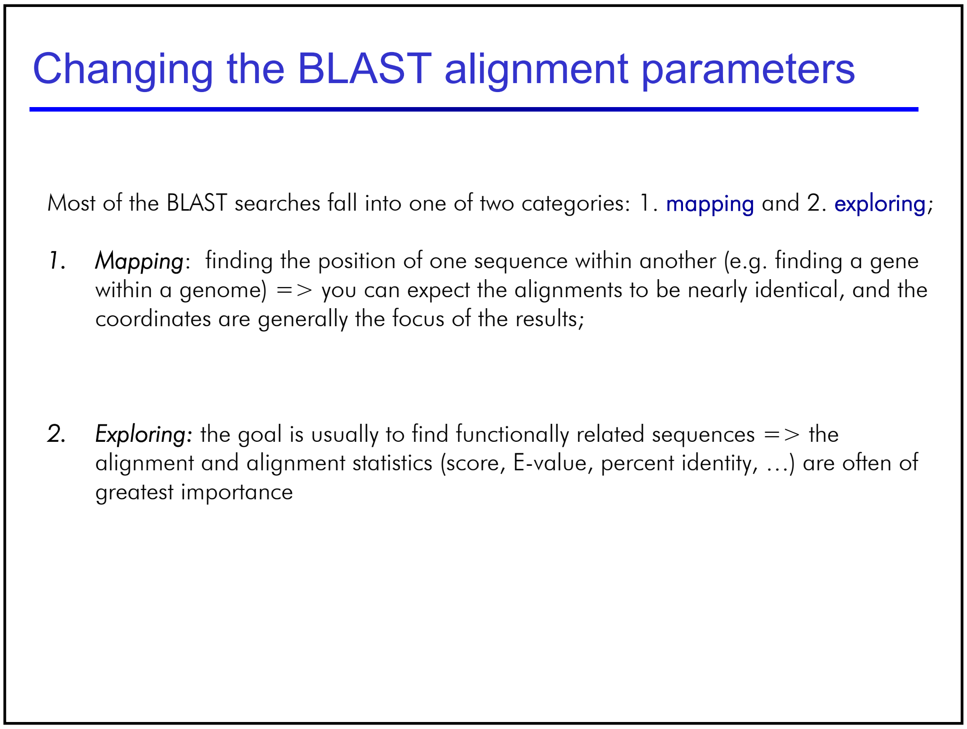


Changing the BLAST alignment parameters

- Among the parameters that you can change on the BLAST servers two important ones have to do with the way BLAST makes the alignments: the gap penalites (gap costs) and the substitution matrix (matrix) or match/mismatch parameters (DNA).

- Use a substitution matrix adapted to the expected divergence of the searched sequences (nevertheless most of the time BLOSUM62 works well):

  - BLOSUM 80: increase selectivity (exclude false positive, missing true positives) (closest to PAM120)
  - BLOSUM 45: increase sensitivity (more true matches, incluse false positives) (closest to PAM250)

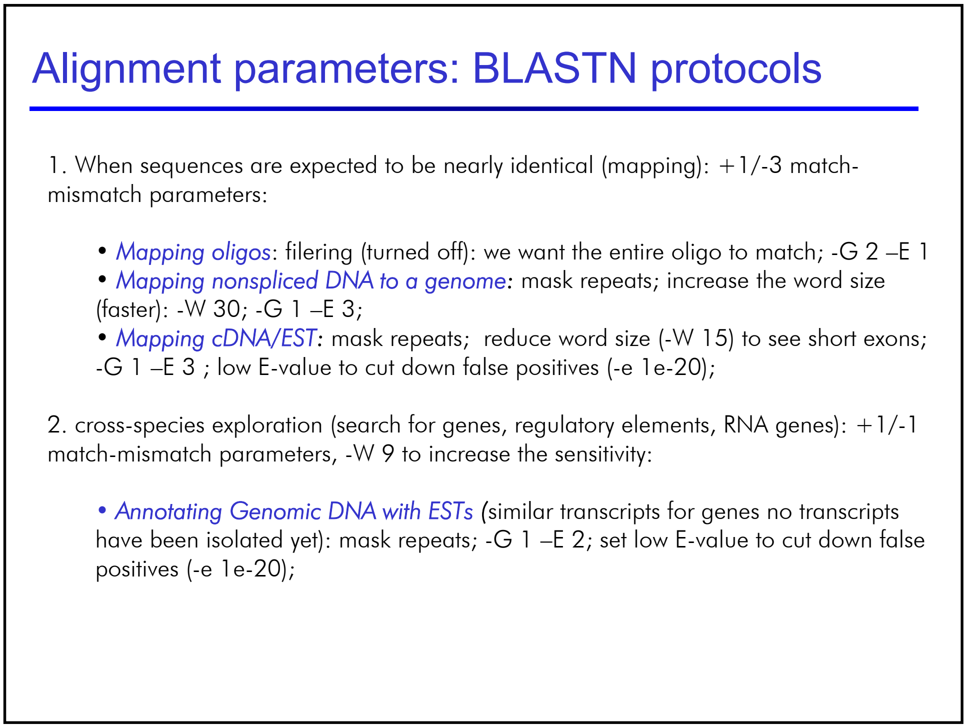


Changing the BLAST alignment parameters

Most of the BLAST searches fall into one of two categories: 1. mapping and 2. exploring;

1. *Mapping*: finding the position of one sequence within another (e.g. finding a gene within a genome) => you can expect the alignments to be nearly identical, and the coordinates are generally the focus of the results;

2. *Exploring*: the goal is usually to find functionally related sequences => the alignment and alignment statistics (score, E-value, percent identity, …) are often of greatest importance



Alignment parameters: BLASTN protocols

1. When sequences are expected to be nearly identical (mapping): +1/-3 match-mismatch parameters:

- *Mapping oligos*: filering (turned off): we want the entire oligo to match; -G 2 –E 1
- *Mapping nonspliced DNA to a genome*: mask repeats; increase the word size (faster): -W 30; -G 1 –E 3;
- *Mapping cDNA/EST*: mask repeats; reduce word size (-W 15) to see short exons; -G 1 –E 3 ; low E-value to cut down false positives (-e 1e-20);

2. cross-species exploration (search for genes, regulatory elements, RNA genes): +1/-1 match-mismatch parameters, -W 9 to increase the sensitivity:

- *Annotating Genomic DNA with ESTs* (similar transcripts for genes no transcripts have been isolated yet): mask repeats; -G 1 –E 2; set low E-value to cut down false positives (-e 1e-20);

Alignment parameters: BLASTP protocols

Most BLASTP searches fall under the exploring category: try to learn about your query sequence by comparing it to other proteins:

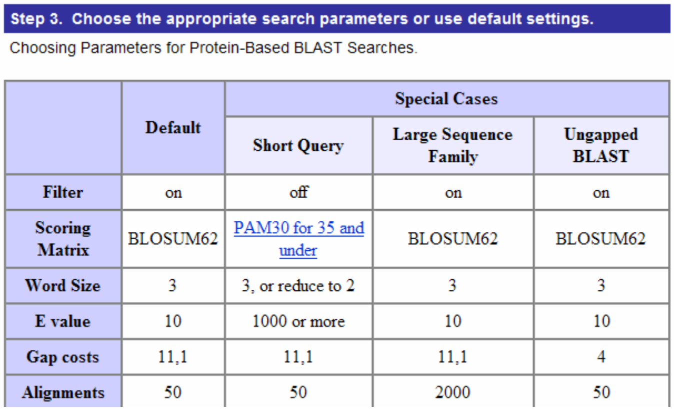
- *Standard search (default parameters)*: balances speed and sensitivity; not ideal for very distant proteomes;

- *Fast insensitive search*: when performing multiple searches (but not for sequences that have less than 50 percent identity); sequences are expected to be very similar: BLOSUM80, set low E-value (-e 1e-5); -G 9 –E 2;

- *Slow, sensitive search*: looking for distant relatives; set E higher (-e 100); BLOSUM45;

Changing the BLAST alignment parameters

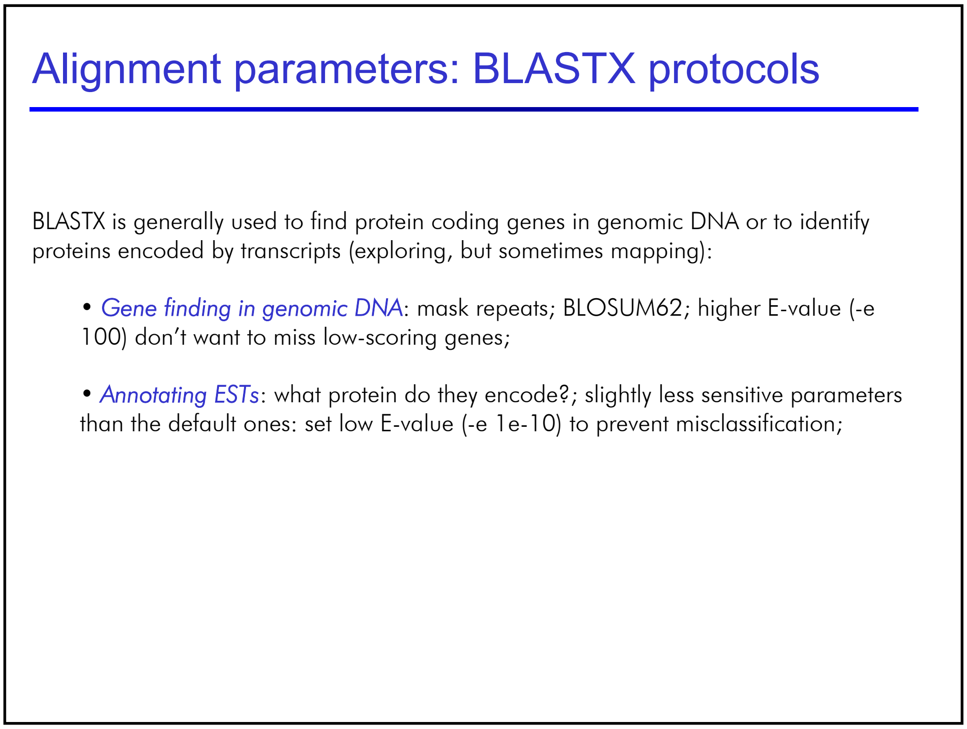
- Guidelines from BLAST tutorial at NCBI (http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/information3.html)



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

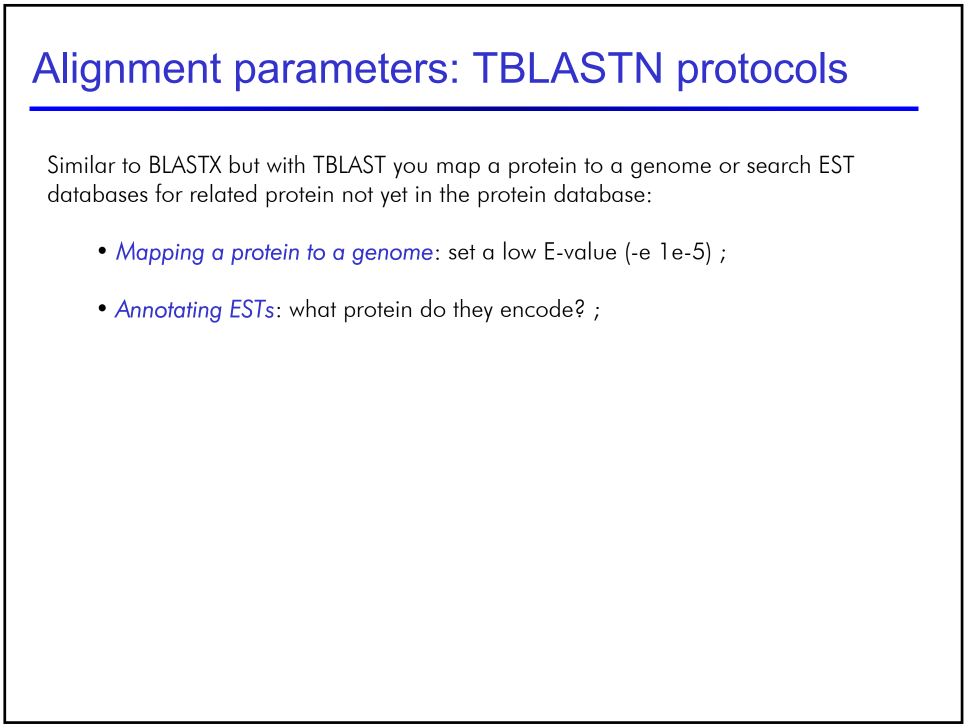


Alignment parameters: BLASTX protocols

BLASTX is generally used to find protein coding genes in genomic DNA or to identify proteins encoded by transcripts (exploring, but sometimes mapping):

- *Gene finding in genomic DNA*: mask repeats; BLOSUM62; higher E-value (-e 100) don't want to miss low-scoring genes;

- *Annotating ESTs*: what protein do they encode?; slightly less sensitive parameters than the default ones: set low E-value (-e 1e-10) to prevent misclassification;



Alignment parameters: TBLASTN protocols

Similar to BLASTX but with TBLAST you map a protein to a genome or search EST databases for related protein not yet in the protein database:

- *Mapping a protein to a genome*: set a low E-value (-e 1e-5) ;

- *Annotating ESTs*: what protein do they encode? ;

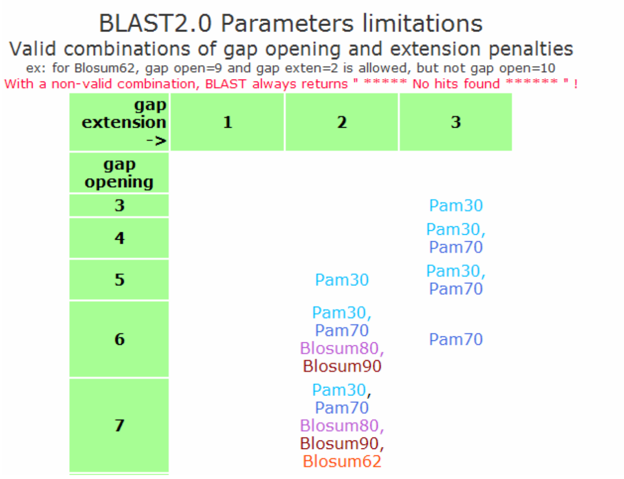
Alignment parameters: TBLASTX protocols

Coding sequences evolve slowly compared to the DNA: TBLASTX for gene-prediction for genomes that are appropriately diverged: not too much (human vs. E.coli) or not enough (human vs. chimpanzees)

- *Finding undocumented genes in genomic DNA*: mask repeats;
- *Transcript of unknown function*: first BLASTX and then (if no results) TBLASTX with ESTs databases;

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	

Conclusions

Blast: the most used database search tool

- Fast and very reliable even for a heuristic algorithm
- Does not necessarily find the best alignment, but most of the time it finds the best matching sequences in the database
- Easy to use with default parameters
- Solid statistical framework for the evaluation of scores

but...

- The biologist's expertise is still essential to the analysis of the results !

Tips and tricks

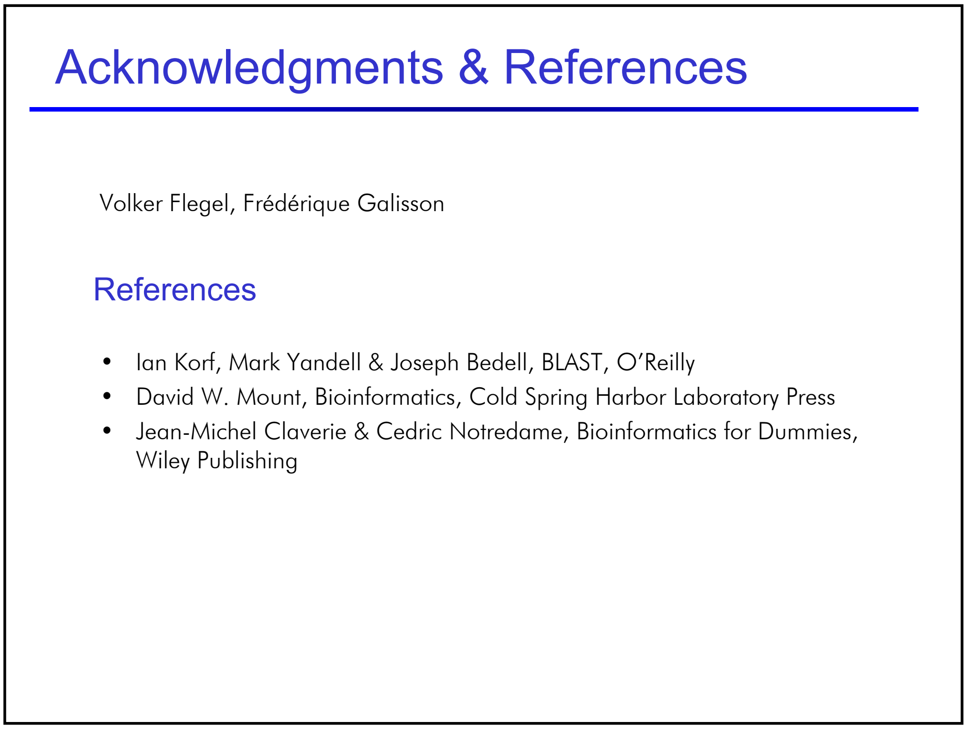
- For coding sequences always search at the protein level
- Mask low complexity regions
- Use a substitution matrix adapted to the expected divergence of the searched sequences (nevertheless most of the time BLOSUM62 works well)
- If there are only matches to a limited region of your query, cut out that region and rerun the search with the remaining part of your query

BLAST Family

- Faster algorithm for genomic search:
  - MegaBLAST (NCBI): http://www.ncbi.nih.gov/BLAST/
  - and SSAHA (Ensembl): http://www.ensembl.org/

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 to speed up the search)

- PSI-BLAST and PHI-BLAST-> Thursday



Acknowledgments & References

Volker Flegel, Frédérique Galisson

References

- Ian Korf, Mark Yandell & Joseph Bedell, BLAST, O'Reilly
- David W. Mount, Bioinformatics, Cold Spring Harbor Laboratory Press
- Jean-Michel Claverie & Cedric Notredame, Bioinformatics for Dummies, Wiley Publishing