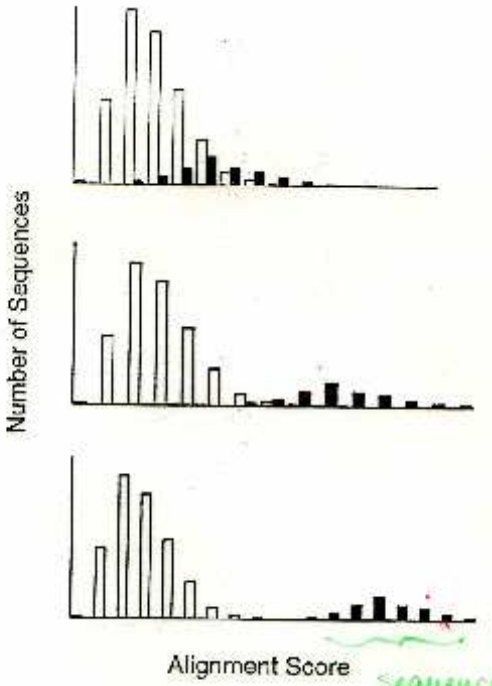
Bioinformatique M1: Lecture 5

P. Derreumaux

ALIGNER UNE SEQUENCE CONTRE UNE BANQUE DE SEQUENCES: FASTA, BLAST, PSI-BLAST

Alignment of your query sequence with database



overlap

perfect case

sequences known to be structurally related to the query sequence.

Two conflicting properties:

Sensitivity: detect 'true positive' matches

Specificity: reject 'false positive' matches.

Programmes de recherche de gènes

★ **2 Mots importants**

- ***Sensibilité***: La capacité à détecter les vraies instances de l'objet recherché (« vrais positifs »).
- ***Spécificité***: La capacité à rejeter les fausses instances (« faux positifs »).

$$\text{Sensibilité: SN} = \frac{TP}{TP+FN}$$

Total « vrais » objets

$$\text{Specificité : SP} = \frac{TP}{TP+FP}$$

Total prédictions

TP: “Vrai positif”
FP: “Faux positif”
FN: “Faux négatif”

Heuristic Search Methods for LOCAL alignment

- Dynamic programming returns best score between 2 Seq
- DP not used to search databases
  - ↑ number of sequences in DB
  - ↑ number of query seq
  - small number of seq in DB related to query
  - speed of computers

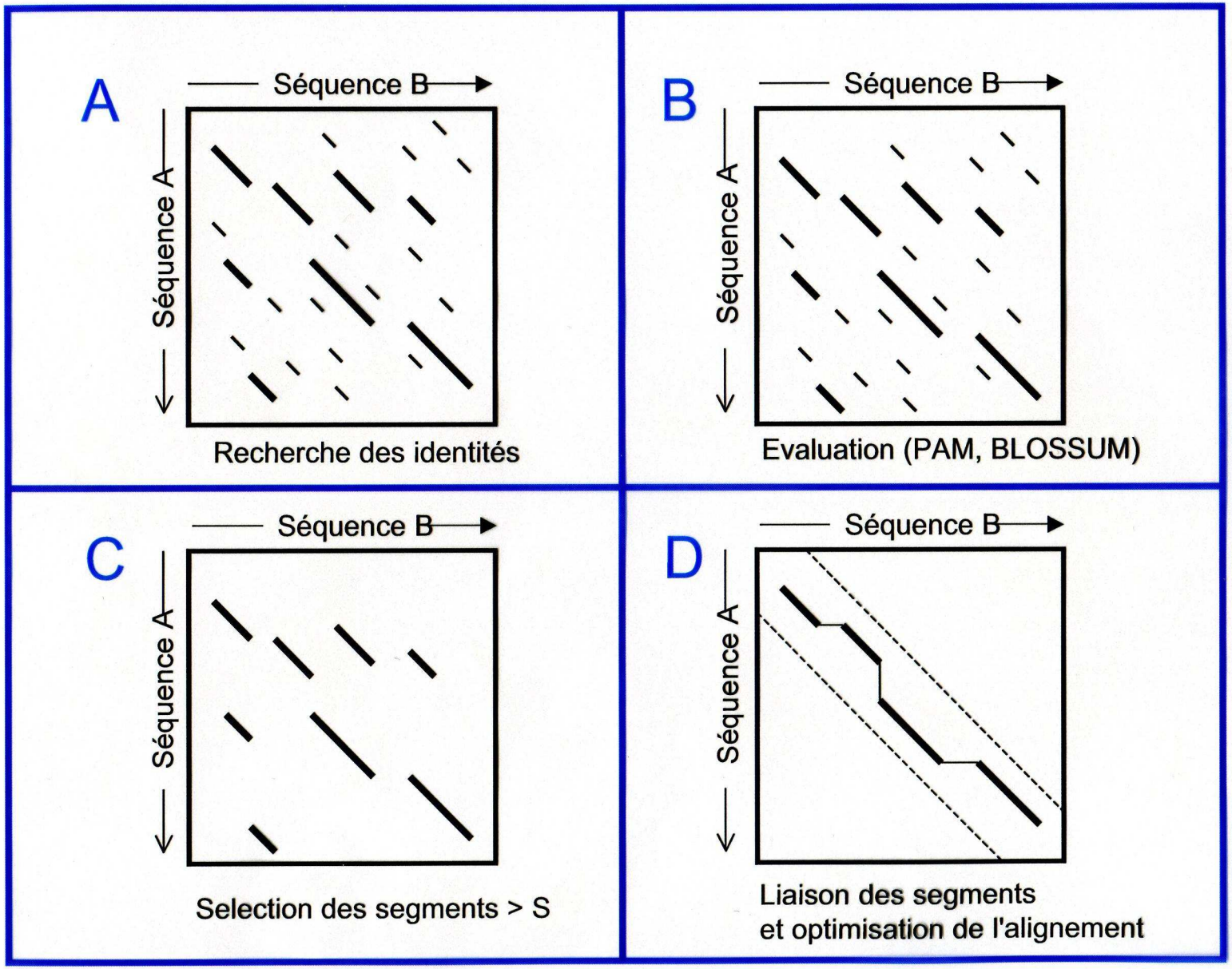
- Methodes heuristiques : succession d'approximations

local alignment
- FASTA : approximation of SW
- BLAST : approximation to a simplified version of SW with effects of gaps removed.

Both programs = 100 faster than SW

- FASTA and BLAST's results must be confirmed

Algorithme FASTA (Pearson & Lipman, 1985)



FASTA : prot query / prot DB

TFASTA : prot query / DNA DB

other DNA query / DNA DB.

Inputs parameters

1. Matrix (BLOSUM.
2. Bank
3. Gap Penalty : P
4. ktup = 2 prot] by default.
   = 6 DNA]
5. Editing parameters (Score min or Evalue min, alignment format)

Séquence brute

```
GGGTTTTATACGGATCTCCCTTCTCGTTTGATAATTATGCCATTAAAGGTTTTACCAGTTCATAAATTTA
GTAAAAATGAACCCCATAAAAAACAAAAGAGGTTCATCTACTTTTAAGGAATTAAACCAAGGAATTTAAT
TCATATTAAATAGCCATGGTTTCCAGTTTTTTACTGGCAGAGTACAAAAACCTAATAGTGAATCCCTCTG
AGCATTTCAAAATCTCAGTGAATGAAGATAATTTGACTGAATGGGATGTCATCTTAAAAGGCCCACCTGA
CACTCTTTATGAGGGAGGCTTATTCAAAGCAAAGATTGTCTTTCCTCCAAAATACCCATATGAACCACCC
AGATTAACATTCACCTCTGAAATGTGGCATCCCAATATCTACTCTGATGGGAAATTATGTATTTCTATCT
TGCATGGAGACAATGCTGAAGAACAGGGAATGACTTGGTCTCCGGCTCAAAAGATTGATACCGTACTTCT
TAGTGTAATTTCTCTGCTCAATGAGCCAAATCCAGATTCTCCAGCAAATGTAGATGCAGCTAAAAGCTAC
CGTAAATATCTATATAAAGAGGATTTAGAATCATACCCCATGGAAGTTAAAAAGACTGTCAAAAAATCAT
TGGATGAGTGTTCAGCGGAAGACATAGAATATTTTAAAAATGTTCCAGTGAATGTTCTACCAGTACCCAG
TGATGATTATGAAGATGAAGAAATGGAGGATGGCACCTATATCTTAACCTATGATGATGAGGATGAAGAA
GAGGATGAAGAGATGGATGATGAGTAGTGCTGATTTTAATGCATAACATATTAGTTACTTACACTTTAGT
GCTTAGATTTTAGTGTTTAAACTTTAGTGATTAGATTTTAGTGCTTAGATTTTAGTGTTTAAACTTTAGT
GATTAGATTTTAGTGCTTAGATTTTAATGTTTAAACTTTAGTGATTAGATTTTAGTGCTTAGATTTTAA
```

Format FASTA

```
>X62440 African swine fever virus DNA for ubiquitin conjugating enzyme
GGGTTTTATACGGATCTCCCTTCTCGTTTGATAATTATGCCATTAAAGGTTTTACCAGTTCATAAATTTA
GTAAAAATGAACCCCATAAAAAACAAAAGAGGTTCATCTACTTTTAAGGAATTAAACCAAGGAATTTAAT
TCATATTAAATAGCCATGGTTTCCAGTTTTTTACTGGCAGAGTACAAAAACCTAATAGTGAATCCCTCTG
AGCATTTCAAAATCTCAGTGAATGAAGATAATTTGACTGAATGGGATGTCATCTTAAAAGGCCCACCTGA
CACTCTTTATGAGGGAGGCTTATTCAAAGCAAAGATTGTCTTTCCTCCAAAATACCCATATGAACCACCC
AGATTAACATTCACCTCTGAAATGTGGCATCCCAATATCTACTCTGATGGGAAATTATGTATTTCTATCT
TGCATGGAGACAATGCTGAAGAACAGGGAATGACTTGGTCTCCGGCTCAAAAGATTGATACCGTACTTCT
TAGTGTAATTTCTCTGCTCAATGAGCCAAATCCAGATTCTCCAGCAAATGTAGATGCAGCTAAAAGCTAC
CGTAAATATCTATATAAAGAGGATTTAGAATCATACCCCATGGAAGTTAAAAAGACTGTCAAAAAATCAT
TGGATGAGTGTTCAGCGGAAGACATAGAATATTTTAAAAATGTTCCAGTGAATGTTCTACCAGTACCCAG
TGATGATTATGAAGATGAAGAAATGGAGGATGGCACCTATATCTTAACCTATGATGATGAGGATGAAGAA
GAGGATGAAGAGATGGATGATGAGTAGTGCTGATTTTAATGCATAACATATTAGTTACTTACACTTTAGT
GCTTAGATTTTAGTGTTTAAACTTTAGTGATTAGATTTTAGTGCTTAGATTTTAGTGTTTAAACTTTAGT
GATTAGATTTTAGTGCTTAGATTTTAATGTTTAAACTTTAGTGATTAGATTTTAGTGCTTAGATTTTAA
```

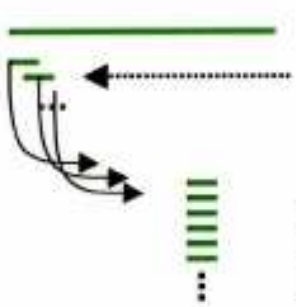
BLAST

BLAST (Basic Local Alignment Search Tool) allows rapid sequence comparison of a query sequence against a database.

The BLAST algorithm is fast, accurate, and web-accessible.

Algorithme BLAST (Altschul *et al.*, 1990)

(1) Recherche de la liste des mots de longueurs w à hauts scores



séquence source de longueur L

Au maximum L-W+1 mots (w~3 pour protéines)

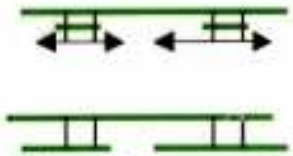
Utilisation d'une matrice
(PAM ou BLOSSUM, etc.)
-> liste des mots qui ont un score ≥ T

(2) Comparaison liste de mots / banque de données -> appariements exacts

Séquences
Banque de données

Appariements exacts

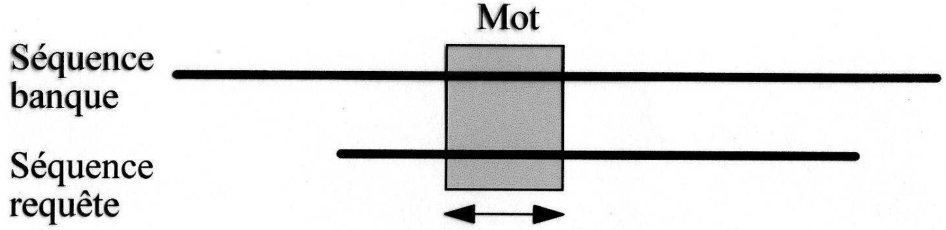
(3) Extension des appariements pour trouver les alignements de score ≥ seuil S



Paires de segments maximaux

BLAST

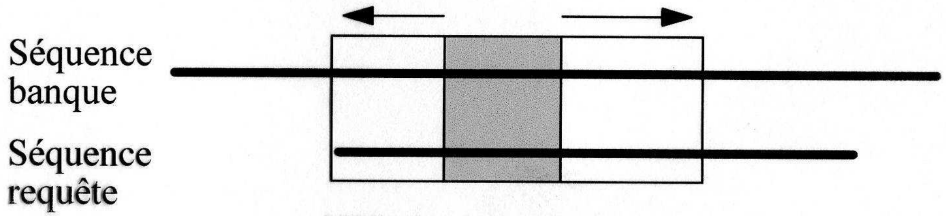
Etape 1: détection de "mots" similaires



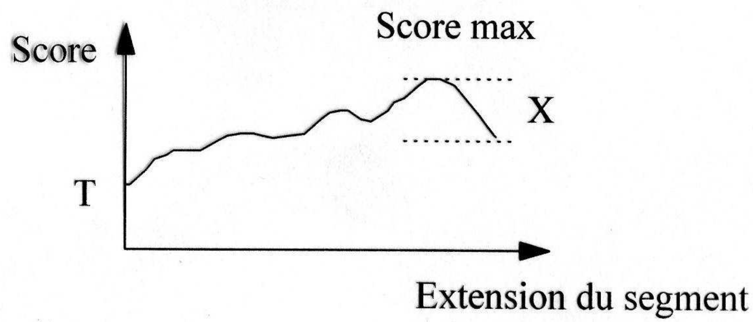
Longueur du mot = W

Score ≥ T (Threshold score)

Etape 2 : extension du segment similaire

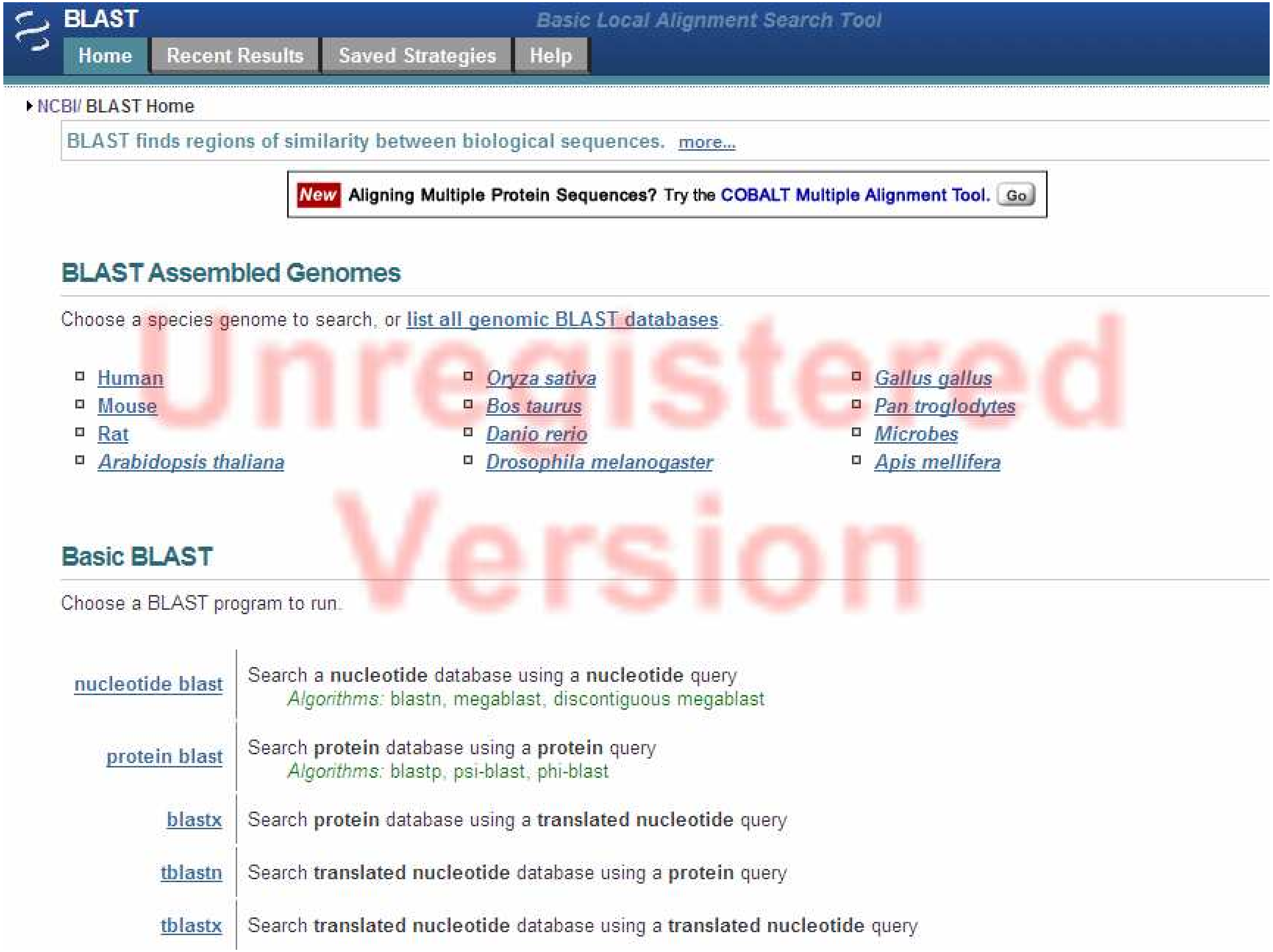


HSP: high score segment pair)

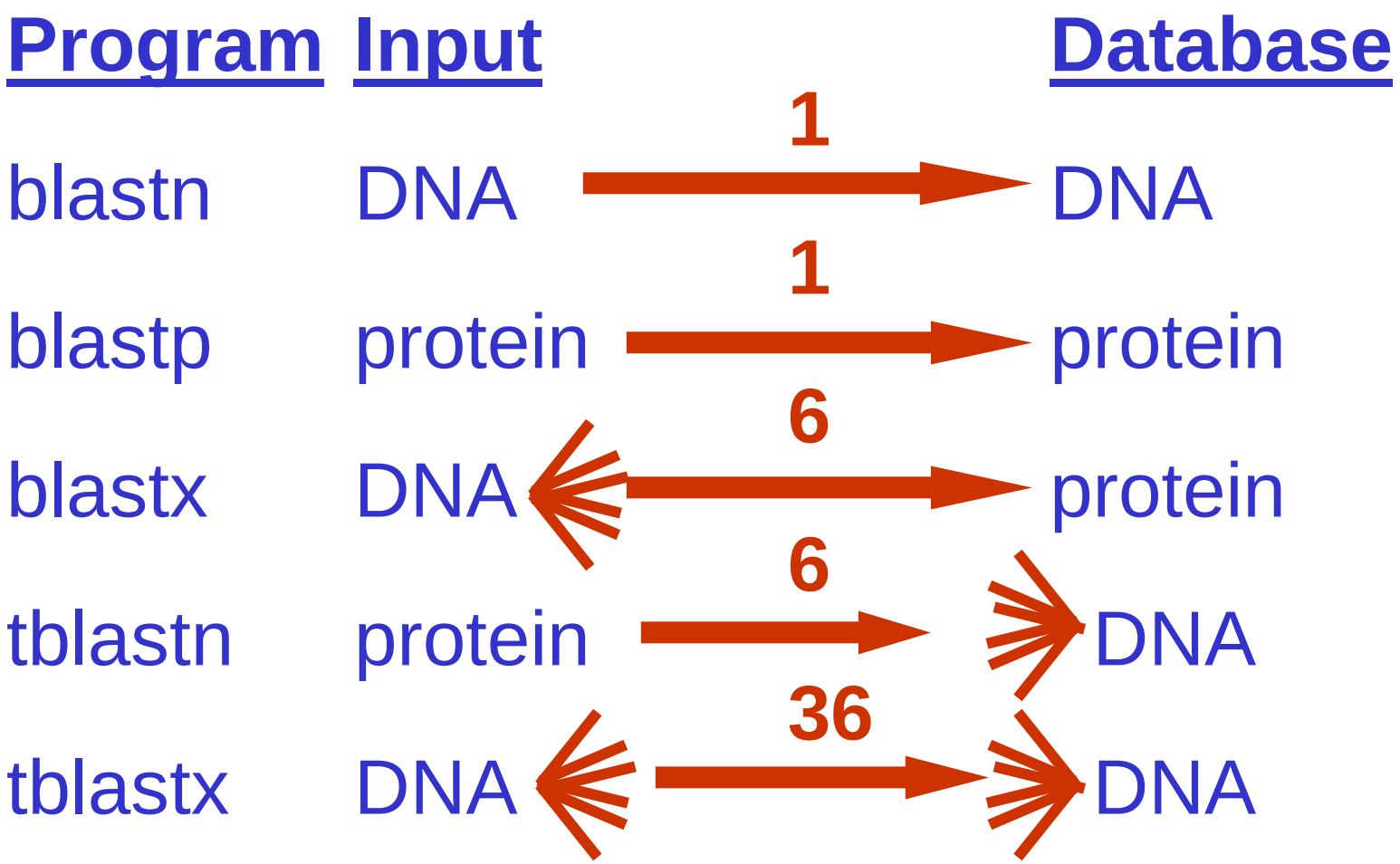


Extension est stoppée quand :

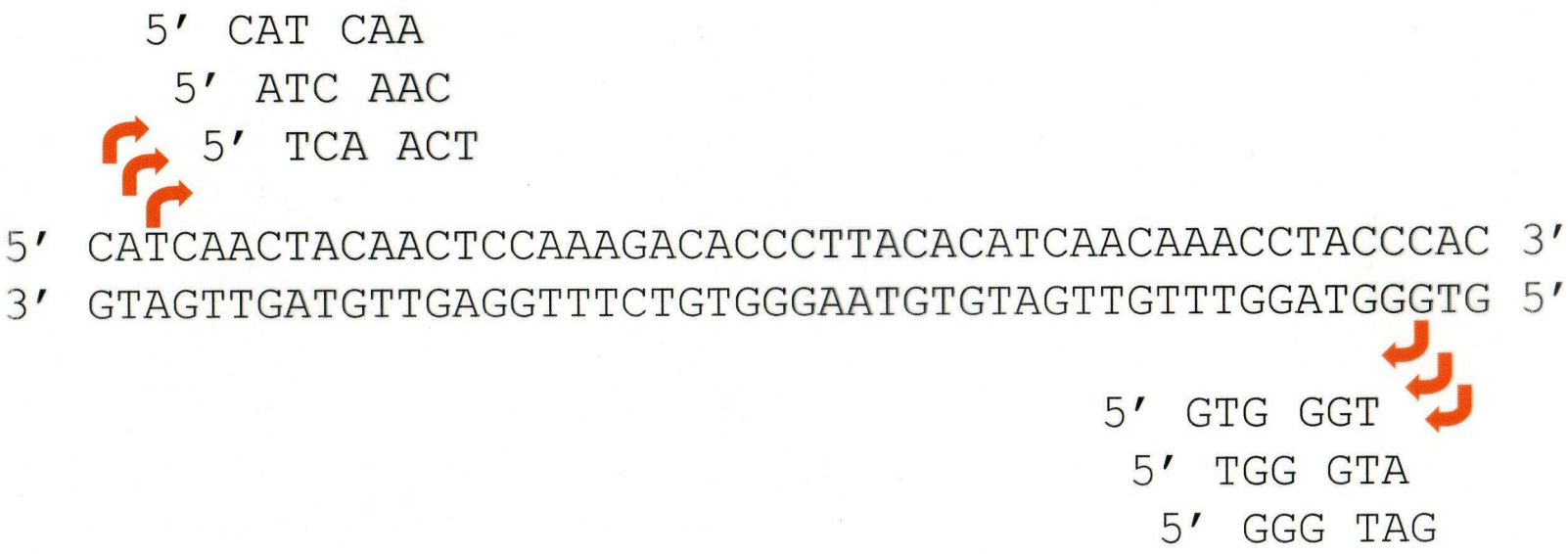
- la fin d'une des deux séquences est atteinte
- [redacted]
- score ≤ score_max - X



Choose the BLAST program

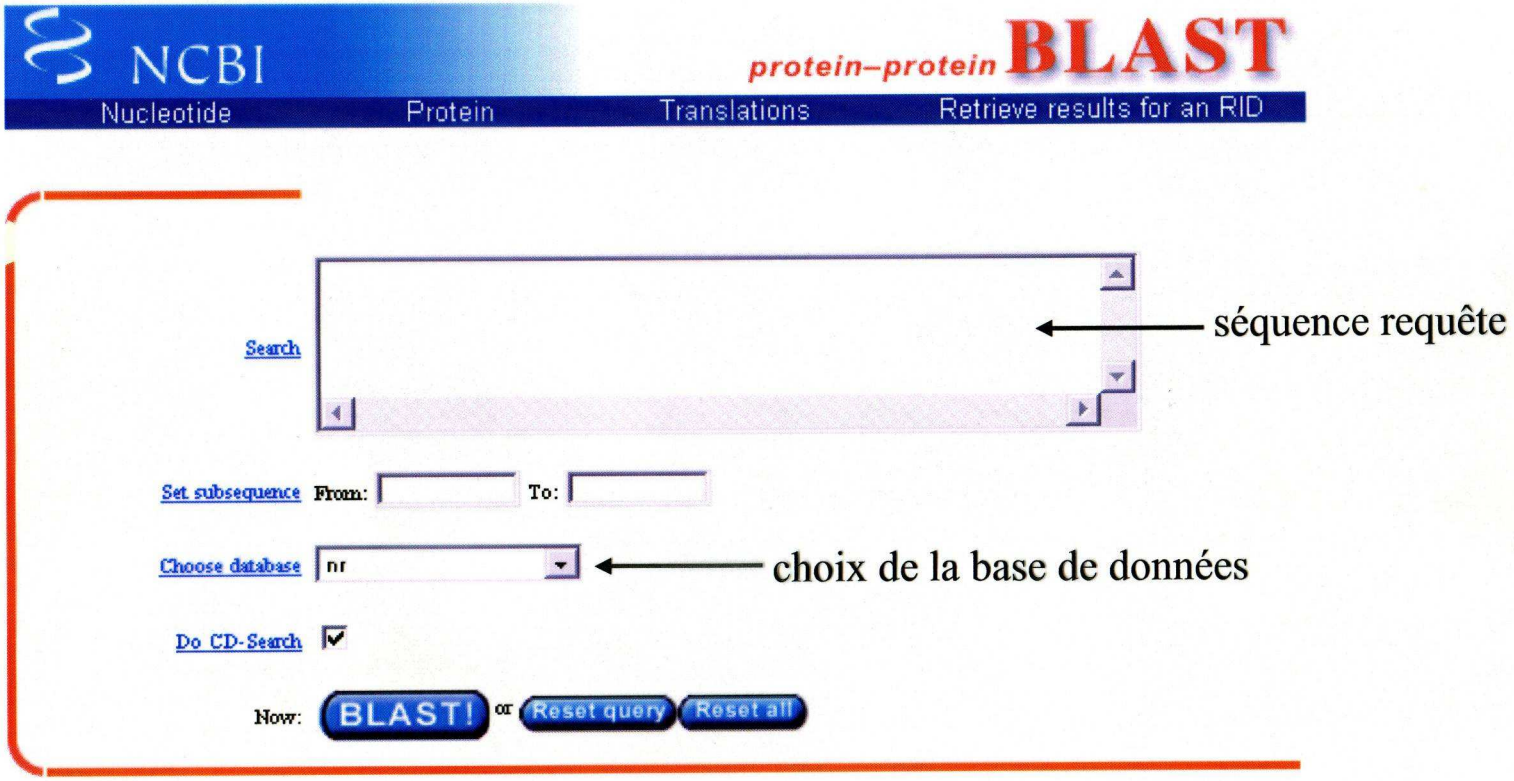


DNA potentially encodes six proteins



BLAST

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



Databases at NCBI

Protein sequence databases

nr All non-redundant GenBank CDS translations
+PDB+SwissProt+PIR+PRF

swissprot the last major release of SWISS-PROT

DNA sequence Databases

nr All Non-redundant GenBank+EMBL+DDBJ+PDB sequences
(but no EST, STS, GSS, or phase 0, 1 or 2 HTGS sequences)

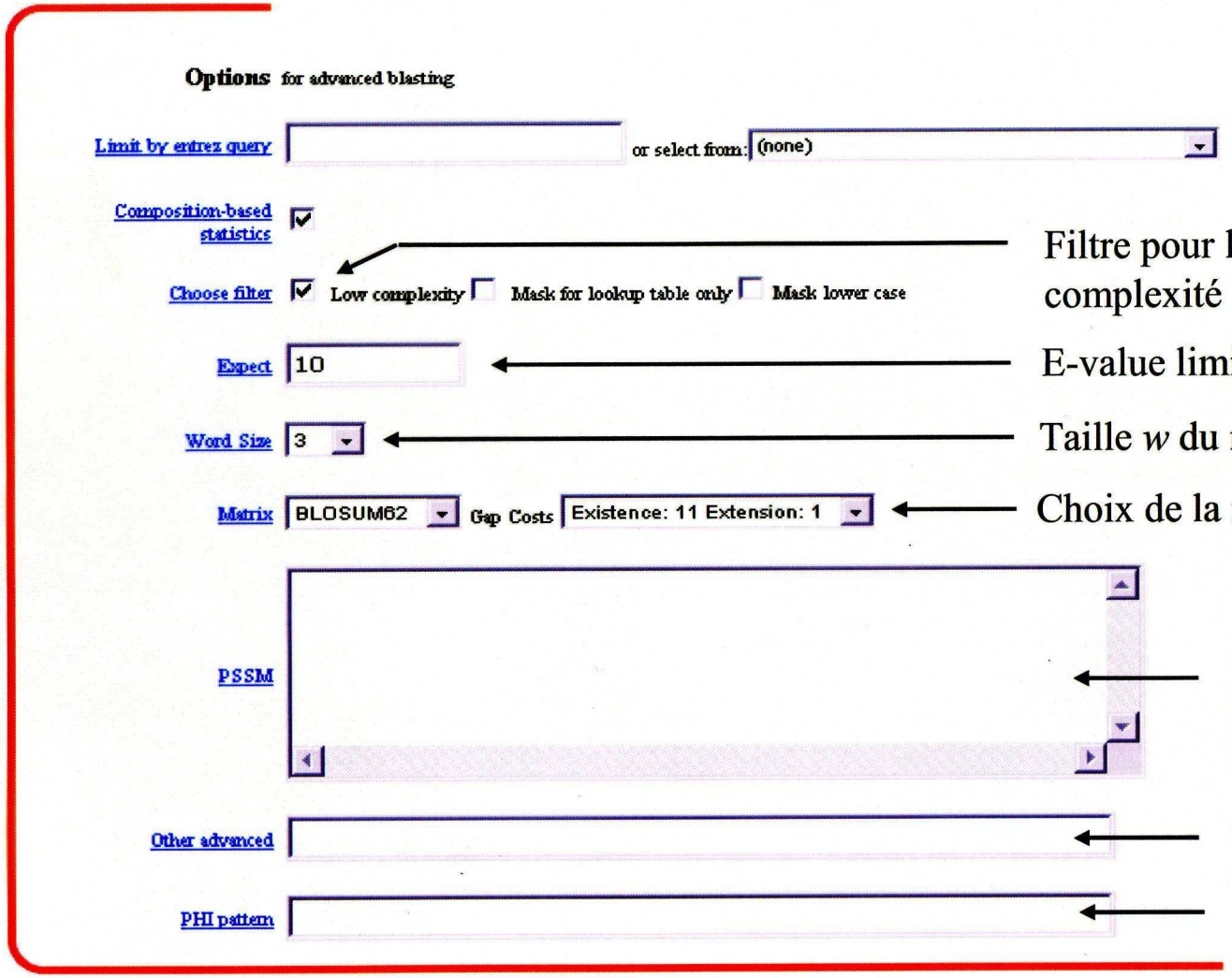
dbest Non-redundant Database of GenBank+EMBL+DDBJ EST Divisions

dbsts Non-redundant Database of GenBank+EMBL+DDBJ STS Divisions

htgs htgs unfinished High Throughput Genomic Sequences

BLAST

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



Limiter la recherche à une espèce

Filtre pour les séquences de faible complexité

E-value limite

Taille *w* du mot *m*

Choix de la matrice et gestion des indels

Position Specific Score Matrix

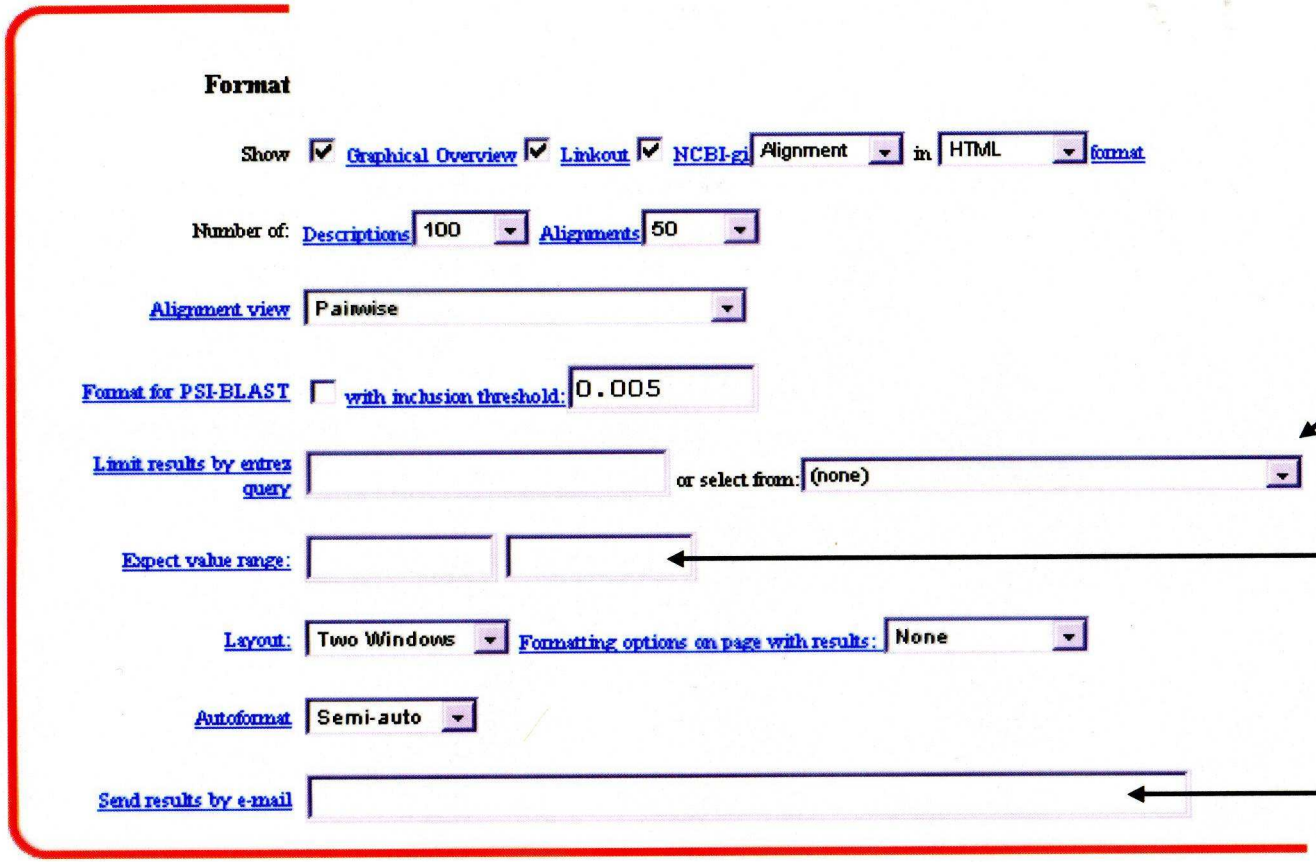
PSI - BLAST

Options supplémentaires

Motif PHI - BLAST

BLAST : Format de la sortie

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



Limiter l 'affichage des résultats à une espèce

Limiter l'affichage des résultats à une plage de valeur d'E-value

Pour recevoir les résultats par e-mail

Filtrage de donnés de séquences avec BLAST

Les sorties de BLAST peuvent être traitées par 3 types de programmes :

- **XNU**
  XNU lit un fichier de séquence au format FASTA et recherche les séquences répétées en tandem ayant statistiquement une signification.
  La motivation de cette fonction est de filtrer les séquences pour éliminer les courtes sous-séquences répétées pouvant fausser les scores des recherches. S'applique aux séquences de protéines

- **SEG**
  SEG lit un fichier de séquence au format FASTA et recherche les zones de faible complexité. Ces dernières sont masqués dans le fichier de sortie. ex proteines anti gel, listeria (alyn)

- DUST lit un fichier de séquence au format FASTA et recherche les zones répétées (dans une fenetre)
  syntaxe : dust fasta-file [cut-off]

- **XBLAST**
  XBLAST lit un fichier de séquence au format FASTA et masque les séquences appartenant à une sortie de BLAST. Ceci permet de filtrer une séquence sur la base de similitudes trouvées dans les banques de données.

Rmq : ces fragments sont remplacés par 'XXXX'

XNU

-a Eliminate only the ascending half of an alignment.

-d Eliminate only the descending half of an alignment.

-r Reverse the output and print the repeats while eliminating the unique portion of the sequence.

-v Turns the verbose flag on. Default is off. Verbose output is sent to stdout.

EXAMPLE

Assuming a file named 'seq' contains the following:

```
>Sample sequence with internal repeats
ACDEFGHIKLMNPQRQRQRQRQRQRQRQRSTVWY
```

```
xnu seq
```

will print the sequence with the repeats eliminated:

```
>Sample sequence with internal repeats
ACDEFGHIKLMNPXXXXXXXXXXXXXXXXXXXSTVWY
```

```
xnu seq -r
```

will print the repetitive part of the sequence:

```
>Sample sequence with internal repeats
XXXXXXXXXXXXXQRQRQRQRQRQRQRQRQRXXXXX
```

SEG

EXAMPLES
The following is a file named 'prion' in FASTA format:

```
>PRIO_HUMAN MAJOR PRION PROTEIN PRECURSOR
MANLGCWMLVLFVATWSDLGLCKKRPKPGGWNTGGSRYPGQGSPGGNRYPPQGGGGWGQP
HGGGWGQPHGGGWGQPHGGGWGQPHGGGWGQGGGTHSQWNKPSKPKTNMKHMAGAAAAGA
VVGGLGGYMLGSAMSRPIIHFGSDYEDRYYRENMHRYPNQVYYRPMDEYSNQNNFVHDCV
NITIKQHTVTTTTKGENFTETDVKMMERVVEQMCITQYERESQAYYQRGSSMVLFSSPPV
ILLISFLIFLIVG
```

The command line:

```
  seg prion
```

gives the standard output below

```
>PRIO_HUMAN MAJOR PRION PROTEIN PRECURSOR

                                    1-49   MANLGCWMLVLFVATWSDLGLCKKRPKPGG
                                           WNTGGSRYPGQGSPGGNRY
ppqggggwgqphgggwgqphgggwgqphgg     50-94
                gwgqphgggwgqggg
                                   95-112  THSQWNKPSKPKTNMKHM
       agaaaagavvgglggymlgsams     113-135
                                   136-187 RPIIHFGSDYEDRYYRENMHRYPNQVYYRP
                                           MDEYSNQNNFVHDCVNITIKQH
                tvtttkgenftet      188-201
                                   202-236 DVKMMERVVEQMCITQYERESQAYYQRGSS
                                           MVLFS
              sppvillisflifliv     237-252
                                   253-253 G
```

The low-complexity sequences are on the left (lower case) and high-complexity sequences are on the right (upper case). All sequence segments read from left to right and their order in the sequence is from top to bottom, as shown by the central column of residue numbers.

The command line:

```
  seg prion -x
```

gives the following FASTA-formatted file:-

```
>PRIO_HUMAN MAJOR PRION PROTEIN PRECURSOR
MANLGCWMLVLFVATWSDLGLCKKRPKPGGWNTGGSRYPGQGSPGGNRYxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxTHSQWNKPSKPKTNMKHMxxxxxxxxx
xxxxxxxxxxxxxxxxRPIIHFGSDYEDRYYRENMHRYPNQVYYRPMDEYSNQNNFVHDCV
NITIKQHxxxxxxxxxxxxxxxxDVKMMERVVEQMCITQYERESQAYYQRGSSMVLFSxxxx
xxxxxxxxxxxxG
```

Why filters ?

- PAM and Blosum work for standard AA composition.

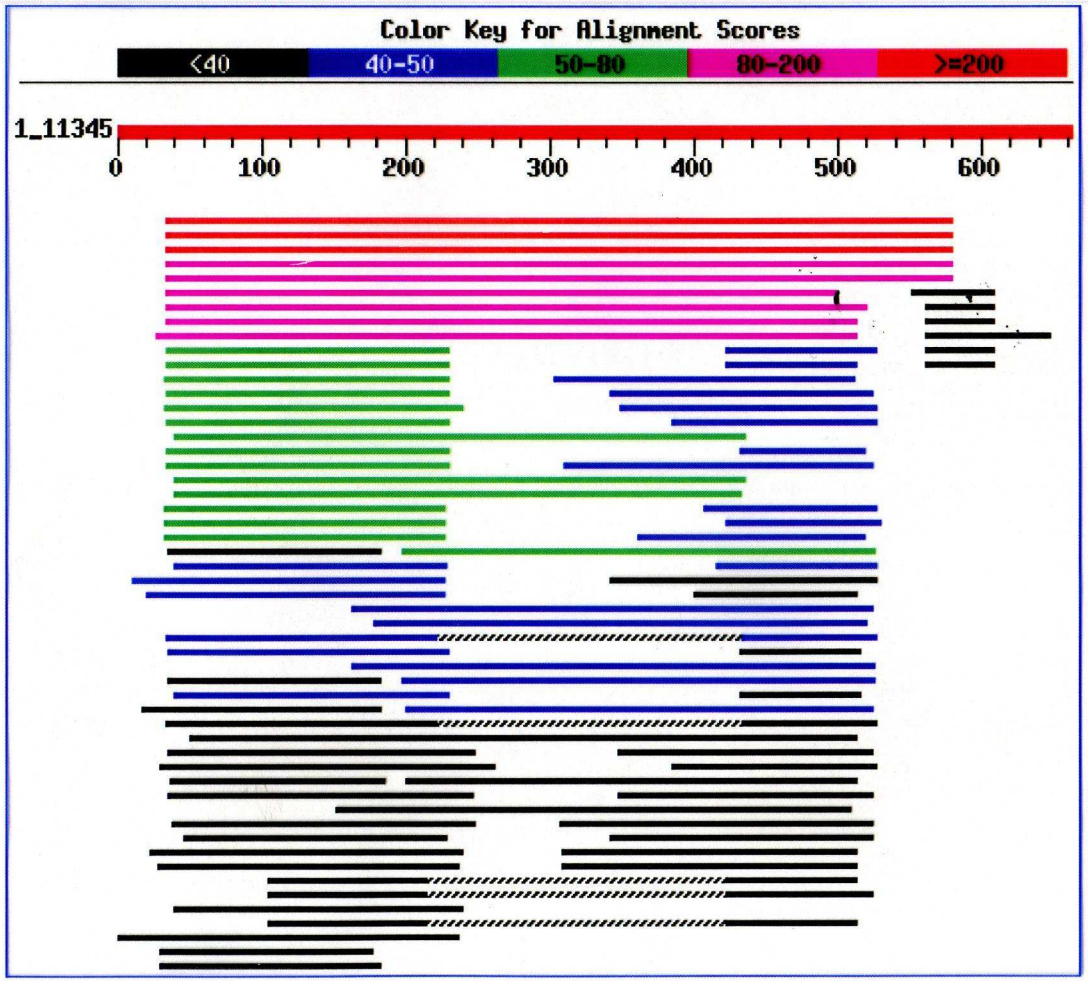
- These low entropy (sequences or fragments) often gives anomalous matches in a database search and thus obscures more meaningful hits.

BLAST: Fichier de sortie

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

Distribution of 134 Blast Hits on the Query Sequence ← Nombres de hits

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



Répartition des hits en fonction du score

BLAST sur Internet : Fichier de sortie (7/8)

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

Sequences producing significant alignments:	Score (bits)	E Value
gi|729833|sp|P40189|IL6B_HUMAN Interleukin-6 receptor beta ...	229	1e-59
gi|729834|sp|Q00560|IL6B_MOUSE Interleukin-6 receptor beta ...	223	8e-58
gi|729835|sp|P40190|IL6B_RAT Interleukin-6 receptor beta ch...	218	3e-56
gi|729564|sp|Q99062|GCSR_HUMAN Granulocyte colony stimulati...	147	9e-35
gi|729565|sp|P40223|GCSR_MOUSE GRANULOCYTE COLONY STIMULATI...	134	4e-31
gi|1170784|sp|P42702|LIFR_HUMAN Leukemia inhibitory factor ...	108	2e-23
gi|1170785|sp|P42703|LIFR_MOUSE Leukemia inhibitory factor ...	106	1e-22
gi|12229836|sp|Q99665|I12S_HUMAN Interleukin-12 receptor be...	103	1e-21
gi|12229832|sp|P97378|I12S_MOUSE Interleukin-12 receptor be...	103	1e-21
gi|2494727|sp|Q90374|PRLR_COLLI PROLACTIN RECEPTOR PRECURSO...	67	7e-11
gi|730343|sp|Q08501|PRLR_MOUSE Prolactin receptor precursor...	65	4e-10
gi|548527|sp|Q04594|PRLR_CHICK PROLACTIN RECEPTOR PRECURSOR...	64	1e-09
gi|2494725|sp|Q28235|PRLR_CEREL PROLACTIN RECEPTOR PRECURSO...	63	1e-09
gi|2494728|sp|Q91094|PRLR_MELGA PROLACTIN RECEPTOR PRECURSO...	62	2e-09
gi|2506457|sp|P05710|PRLR_RAT Prolactin receptor precursor ...	62	3e-09
gi|6166563|sp|Q62959|LEPR_RAT Leptin receptor precursor (LE...	62	4e-09
gi|2494724|sp|Q28172|PRLR_BOVIN PROLACTIN RECEPTOR PRECURSO...	60	1e-08
gi|130323|sp|P14787|PRLR_RABIT PROLACTIN RECEPTOR PRECURSOR...	59	4e-08
gi|1352612|sp|P48356|LEPR_MOUSE Leptin receptor precursor (...	57	7e-08
gi|1352611|sp|P48357|LEPR_HUMAN Leptin receptor precursor (...	53	2e-06
gi|1168985|sp|Q08406|CNTR_RAT Ciliary neurotrophic factor r...	52	4e-06
gi|2494726|sp|Q91513|PRLR_ORENI PROLACTIN RECEPTOR PRECURSO...	52	5e-06
gi|1352099|sp|P26992|CNTR_HUMAN Ciliary neurotrophic factor...	51	8e-06
gi|125978|sp|P10586|PTPF_HUMAN LAR protein precursor (Leuko...	51	8e-06
gi|2494719|sp|O09030|I131_MOUSE Interleukin-13 receptor alp...	49	4e-05
gi|12229808|sp|Q91735|EPB3_XENLA Ephrin type-B receptor 3 p...	48	6e-05
gi|130321|sp|P16471|PRLR_HUMAN Prolactin receptor precursor...	47	2e-04
gi|1705964|sp|P51641|CNTR_CHICK CILIARY NEUROTROPHIC FACTOR...	45	5e-04
gi|12229805|sp|Q07498|EPB3_CHICK Ephrin type-B receptor 3 (...	44	9e-04
gi|3182940|sp|Q99715|CA1C_HUMAN Collagen alpha 1(XII) chain...	44	0.001
gi|12229834|sp|Q60837|I12R_MOUSE Interleukin-12 receptor be...	43	0.002

BLAST : fichier de sortie

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

```
>gi|729833|sp|P40189|IL6B_HUMAN  Interleukin-6 receptor beta chain precursor (IL-6R-beta)
          (Interleukin 6 signal transducer) (Membrane glycoprotein
          130) (GP130) (Oncostatin M receptor) (CDw130) (CD130
          antigen)
         Length = 918

 Score =  229 bits (583), Expect = 1e-59
 Identities = 154/550 (28%), Positives = 250/550 (45%), Gaps = 12/550 (2%)

Query: 35  PAKPENISCVYYYRKNLTCTWSPGKETSY-TQYTVKRTYAFGEKHDNCTTNSSTSENRAS 93
           P KP+N+SC+    K + C W  G+ET   T +T+K  +A   K  +C     T     S
Sbjct: 126 PEKPKNLSCIVNEGKKMRCEWDGGRETHLETNFTLKSEWA-THKFADCKAKRDTP---TS 181

Query: 94  CSFFLPRITIPDNYTIEVEAENGDGVIKSHMTYWRLENIAKTEPPKIFRVKPVLGIKRMI 153
           C+     +    N  + VEAEN  G + S    +       K  PP    V     +  ++
Sbjct: 182 CTVDYSTVYFV-NIEVWVEAENALGKVTSDHINFDPVYKVKPNPPHNLSVINSEELSSIL 240

Query: 154 QIEWIKPELAPVSSDLKYTLRFRTVNSTSWMEVNFAKNRKDKNQTYNLTGLQPFTEYVIA 213
           ++ W  P +  V   LKY +++RT ++++W ++   ++       ++ +  L+PFTEYV
Sbjct: 241 KLTWTNPSIKSVII-LKYNIQYRTKDASTWSQIP-PEDTASTRSSFTVQDLKPFTEYVFR 298

Query: 214 LRCAVKESK-FWSDWSQEKMGMTEEEAPC-GLELWRVLKPAEADGRRPVRLLWKKARGAP 271
           +RC  ++ K +WSDWS+E  G+T E+ P      W  + P+   G R V+L+WK
Sbjct: 299 IRCMKEDGKGYWSDWSEEASGITYEDRPSKAPSFWYKIDPSHTQGYRTVQLVWKTLPPFE 358

Query: 272 VLEKTLGYNIWYYPESXXXXXXXXXXXXXXXLHLGGESFWVSMISYNSLGKSPVATLRI 331
              K L  Y +                        ++L  + +  ++   N +GKS  A L I
Sbjct: 359 ANGKILDYEVTL--TRWKSHLQNYTVNATKLTVNLTNDRYLATLTVRNLVGKSDAAVLTI 416

Query: 332 PAIQEKSFQCIEVMQACVAEDQLVVKWQSSALDVNTWMIEWFPDVDSEPTTLSWESVSQA 391
           PA   ++   +  ++A   ++ L V+W +    V  +++EW    D  P    W+
Sbjct: 417 PACDFQATHPVMDLKAFPKDNMLWVEWTTPRESVKKYILEWCVLSDKAPCITDWQQEDGT 476
```

Sometimes a real match has an E value > 1

Sequences producing significant alignments:		Score (bits)	E Value
gi|5803139|ref|NP_006735.1|	retinol-binding protein 4, inte...	378	e-105
gi|230284|pdb|1RBP|	Retinol Binding Protein >gi|493897|p...	371	e-103
gi|88364|pir||A27786	plasma retinol-binding protein - human	370	e-103
gi|4558179|pdb|1QAB|E	Chain E, The Structure Of Human Retin...	363	e-100
gi|7770173|gb|AAF69622.1|AF119917_30	(AF119868) PRO2222 [Ho...	324	5e-89
gi|13645517|ref|XP_005907.2|	retinol-binding protein 4, int...	233	9e-62
gi|296672|emb|CAA26553.1|	(X02775) RBP [Homo sapiens]	207	8e-54
gi|5419892|emb|CAB46489.1|	(X02824) RBP (aa 101-172) [Homo ...	149	2e-36
gi|2895204|gb|AAC02945.1|	(AF025334) mutant retinol binding...	90	2e-18
gi|2895206|gb|AAC02946.1|	(AF025335) mutant retinol binding...	73	2e-13
gi|4502163|ref|NP_001638.1|	apolipoprotein D precursor [Hom...	55	4e-08
gi|619383|gb|AAB32200.1|	apolipoprotein D, apoD [human, pla...	55	5e-08
gi|1246096|gb|AAB35919.1|	(S80440) apolipoprotein D, apoD (...	43	3e-04
gi|223373|prf||0801163A	complex-forming glycoprotein HC [Ho...	37	0.011
gi|4884164|emb|CAB43305.1|	(AL050169) hypothetical protein ...	35	0.043
gi|13639329|ref|XP_005360.3|	61620 [Homo sapiens] >gi|13639...	35	0.043
gi|4502067|ref|NP_001624.1|	alpha-1-microglobulin/bikunin p...	35	0.068
gi|14735821|ref|XP_029964.1|	progestagen-associated endomet...	35	0.070
gi|4557393|ref|NP_000597.1|	complement component 8, gamma p...	34	0.14
gi|4505583|ref|NP_002562.1|	progestagen-associated endometr...	32	0.49
gi|13639651|ref|XP_005430.2|	complement component 8, gamma ...	31	1.1

…try a reciprocal BLAST to confirm

Sometimes a similar *E* value occurs for a short exact match and long less exact match

```
>gi|2895206|gb|AAC02946.1| (AF025335) mutant retinol binding protein [Homo sapiens]
          Length = 36

 Score = 72.8 bits (177), Expect = 2e-13
 Identities = 34/36 (94%), Positives = 35/36 (96%)

Query: 82  NWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGN 117
           NWDVCADMV TFTDTEDPAKFKMKYWGVASFLQKG+
Sbjct: 1   NWDVCADMVDTFTDTEDPAKFKMKYWGVASFLQKGS 36

>gi|4502163|ref|NP_001638.1| apolipoprotein D precursor [Homo sapiens]
 gi|13646407|ref|XP_003067.3| apolipoprotein D precursor [Homo sapiens]
 gi|14727745|ref|XP_049984.1| apolipoprotein D precursor [Homo sapiens]
 gi|114034|sp|P05090|APD_HUMAN APOLIPOPROTEIN D PRECURSOR
 gi|72088|pir||LPHUD apolipoprotein D precursor [validated] - human
 gi|178841|gb|AAB59517.1| (J02611) apolipoprotein D precursor [Homo sapiens]
 gi|178847|gb|AAA51764.1| (M16696) apolipoprotein D precursor [Homo sapiens]
 gi|13938509|gb|AAH07402.1|AAH07402 (BC007402) apolipoprotein D [Homo sapiens]
          Length = 189

 Score = 55.5 bits (132), Expect = 4e-08
 Identities = 47/151 (31%), Positives = 78/151 (51%), Gaps = 39/151 (25%)

Query: 27  VKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQMSATAKGRVRLLNNWDVC 86
           V+ENFD  ++ G WY + +K P         I A +S+ E          G++++LN  ++
Sbjct: 33  VQENFDVNKYLGRWYEI-EKIPTTFENGRCIQANYSLMEN--------GKIKVLNQ-ELR 82

Query: 87  ADMVGTFTDTE---------DPAKFKMKY-WGVASFLQKGNDDHWIVDTDYDTYAVQYSC 136
           AD  GT    E          +PAK ++K+ W + S         +WI+ TDY+ YA+ YSC
Sbjct: 83  AD--GTVNQIEGEATPVNLTEPAKLEVKFSWFMPS------APYWILATDYENYALVYSC 134

Query: 137 ----RLLNLDGTCADSYSFVFSRDPNGLPPE 163
               +L ++D      ++++ +R+PN LPPE
Sbjct: 135 TCIIQLFHVD------FAWILARNPN-LPPE 158
```

Many applications of BLAST

- Blastx cDNA / Prot
- Blastn cDNA / Gb
- Blastn cDNA / dbEST

- ★ Caractériser le gène correspondant à un cDNA
- ★ Trouver d'autres cDNA ou mRNA semblables
- ★ Trouver des EST correspondant à un cDNA (p.ex. pour profil d'expression)

EST = Expressed Sequence Tags
= Small pieces of messager RNA

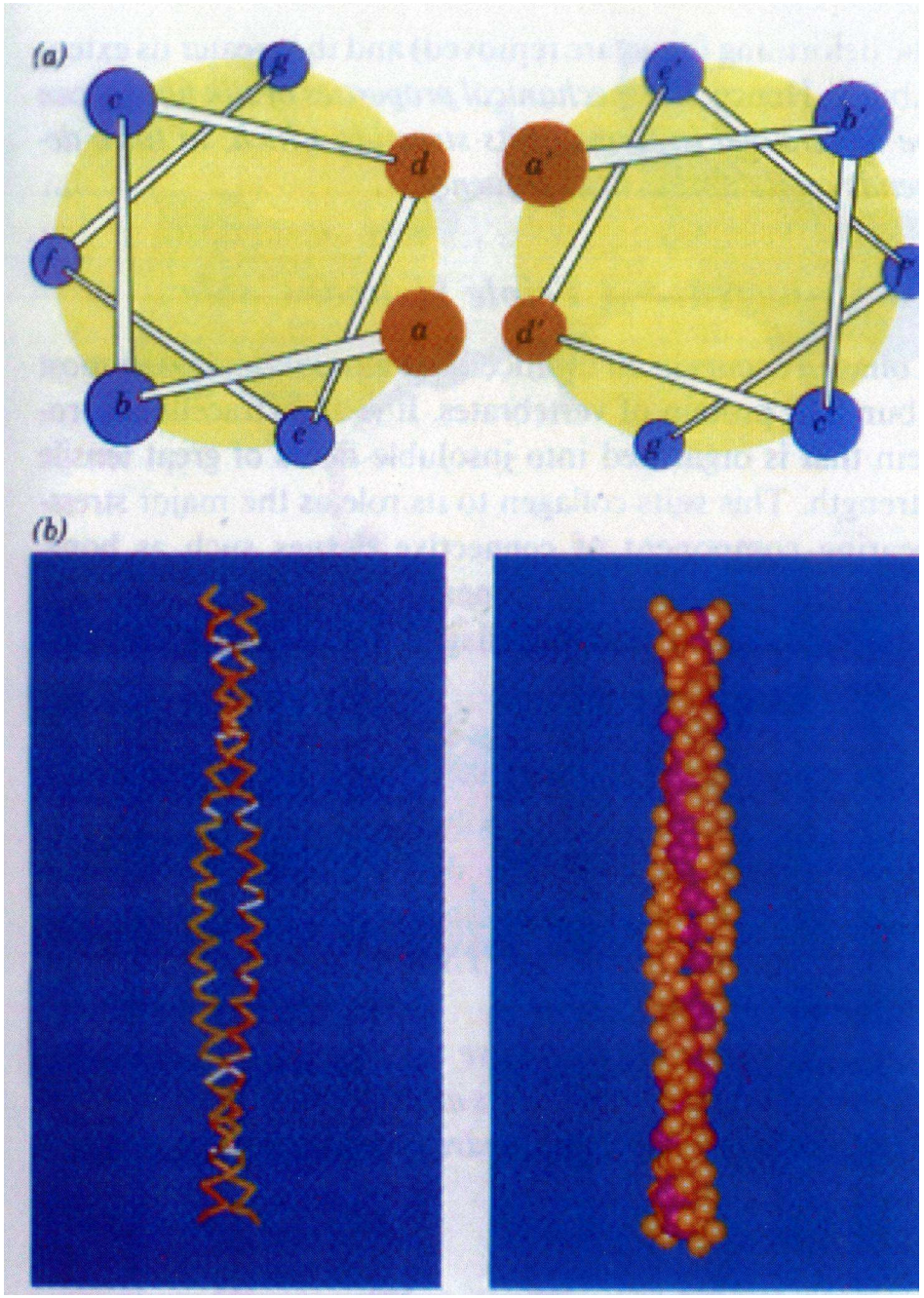
→ Two final remarks on BLAST

Rmk 1

Other Filters.

- Some types of low complexity sequences are not detected.
  - transmembrane regions
  - coiled-coil regions

Use appropriate programs to find them and mask these regions (XXXX) before using Blast or other softwares.


(a)
g
c
d
f
a
b
e
c'
b'
a'
f'
d'
g'
e'
(b)

Rmk2

- For genomes, BLAT (BLAST-Like Alignment Tool) is often used.

  BLAT is 500 times faster than BLAST because the entire genomes are not conserved.

  BLAT's speed stems from the use of all nonoverlapping k-mers in the genome.

  k-mer size: typically 8-16 for nucleotide comparisons and 3-7 for amino acid comparison

- For detection of Orthologs (e.g. COGs) Reciprocal BLAST is used.

PHI-BLAST: Pattern hit initiated BLAST

Launches from the same page as PSI-BLAST

Combines matching of regular expressions
with local alignments surrounding the match.

Given a protein sequence S and a regular expression pattern P occurring in S, PHI-BLAST helps answer the question: What other protein sequences both contain an occurrence of P and are homologous to S in the vicinity of the pattern occurrences? PHI-BLAST may be preferable to just searching for pattern occurrences because it filters out those cases where the pattern occurrence is probably random and not indicative of homology.

NCBI

Nucleotide Protein formatting BLAST Translations Retrieve results for an RID

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

Query = (76 letters)

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

The request ID is 1092244432-18962-139331594840.BLASTQ4

Format! or Reset all

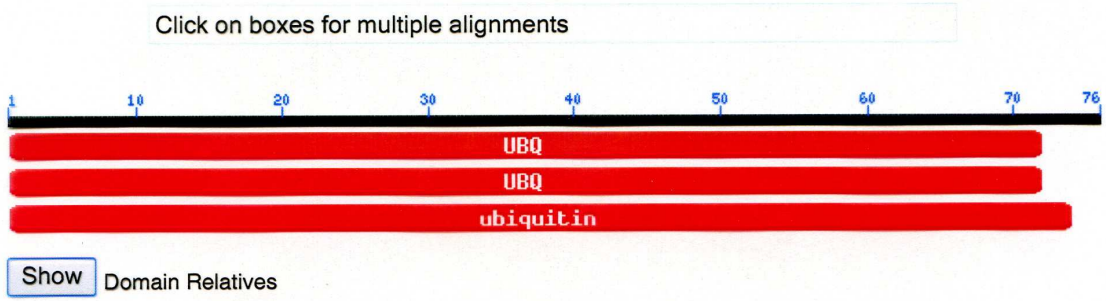
The results are estimated to be ready in 22 seconds but may be done sooner.

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

NCBI Conserved Domain Search



.. This CD alignment includes 3D structure. To display structure, download **Cn3D**!

PSSMs producing significant alignments:		Score (bits)	E value
gnl|CDD|5394	cd00196, UBQ, Ubiquitin homologs; Includes ubiquitin and ubiqu...	98.7	1e-22
gnl|CDD|189	smart00213, UBQ, Ubiquitin homologues; Ubiquitin-mediated prot...	95.3	1e-21
gnl|CDD|25454	pfam00240, ubiquitin, Ubiquitin family. This family contains a...	99.6	6e-23

---

gnl|CDD|5394, cd00196, UBQ, Ubiquitin homologs; Includes ubiquitin and ubiquitin-like proteins. Ubiquitin-mediated proteolysis is part of the regulated turnover of proteins required for controlling cell cycle progression. Other family members are protein modifiers that perform a wide range of functions. Ubiquitination usually results in a covalent bond between the c-terminus of ubiquitin and the epsilon-amino group of a substrate lysine. The three-step mechanism requires an activating enzyme (E1) that forms a thiol ester with the c-terminal carboxy group, a conjugating enzyme (E2) that transiently carries the activated ubiquitin molecule as a thiol ester, and a ligase (E3) that transfers the activated ubiquitin from the E2 to the substrate lysine residue. In poly-ubiquitination ubiquitin itself is the substrate.

```
CD-Length = 72 residues, 100.0% aligned
Score = 98.7 bits (246), Expect = 1e-22

Query:  1   MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYN  60
Sbjct:  1   ISLTVKTLDGKTITLEVKPSDTVSELKAKIEEKEGVPPEQQRLIFKGKVLEDEQTLADYG  60

Query:  61  IQKESTLHLVLR  72
Sbjct:  61  IQDGSTIHLVLR  72
```

---

gnl|CDD|189, smart00213, UBQ, Ubiquitin homologues; Ubiquitin-mediated proteolysis is involved in the regulated turnover of proteins required for controlling cell cycle progression

```
CD-Length = 72 residues, 100.0% aligned
```

Definition: Consensus-String

The consensus-String S_M of a multiple alignment is the concatenation of all consensus symbols.

A	**T**	**A**	**A**	**G**	**C**
A	**-**	**A**	**T**	**G**	**C**
A	**-**	**A**	**A**	**G**	**C**
A	**-**	**A**	**A**	**G**	**C**

Consensus Sequence S_M if the majority rule is applied

Profils ou PPSM

- plus subtils que les consensus ou les patterns PROSITE.
- Pour chaque position de l'alignement, on détermine la fréquence d'observation des différents résidus.
- conversion matrice fréquences → PSSM.

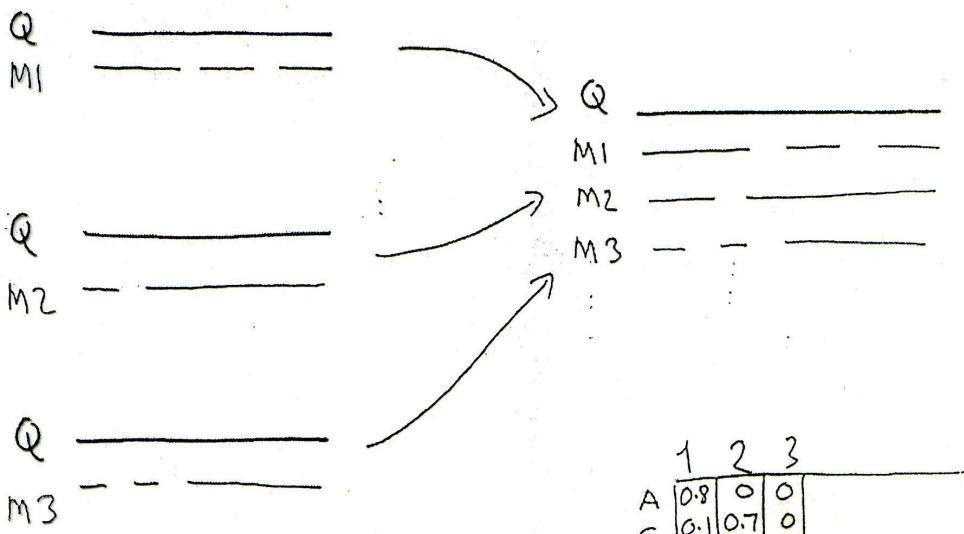
$$\text{Score}(b,i) = \log \frac{F_{obs\,b,i}}{F_{exp\,b}}$$

- b : character b
- i : position i
- $F_{exp\,b}$ → from large data set (genomes)

PSI-BLAST

(1) Search query sequence against large database using scoring matrix (1st step) or profile (2nd, ...)
Keep significant hits

(2) Align them all to query



(3) Make a sequence profile

	1	2	3	...	103
A	0.8	0	0		0.1
C	0.1	0.7	0		0.1
...					
W	0	0.1	1.0		0.1
Y	0	0.2	0		0.1

Search with the profile (return to (1)) until the threshold value for inclusion in the position specific matrix is satisfied → (2nd E-value parameter)

S. Altschul et al.

Nucleic Acids Research, 1997, Vol. 25, No. 17

3. The number of SWISS-PROT sequences yielding alignments with *E*-value ≤0.01, and relative running times, for Smith–Waterman and various version of BLAST

Protein family	Query	Smith–Waterman	Original BLAST	Gapped BLAST	PSI-BLAST
Serine protease	P00762	275	273	275	286
Serine protease inhibitor	P01008	108	105	108	111
Ras	P01111	255	249	252	375
Globin	P02232	28	26	28	623
Hemagglutinin	P03435	128	114	128	130
Interferon α	P05013	53	53	53	53
Alcohol dehydrogenase	P07327	138	128	137	160
Histocompatibility antigen	P10318	262	241	261	338
Cytochrome P450	P10635	211	197	211	224
Glutathione transferase	P14942	83	79	81	142
H^+-transporting ATP synthase	P20705	198	191	197	207
Normalized running time		36	1.0	0.34	0.87

To score and evaluate the significance of the alignments found, the original BLAST program uses BLOSUM-62 substitution scores (18) and sum-statistics (21,22). The Smith–Waterman and gapped BLAST programs use BLOSUM-62 substitution scores, $10 + k$ gap costs, and the statistics of equations 1 and 2, in conjunction with the experimentally determined parameters $\lambda_g = 0.255$ and $K_g = 0.035$ (3). PSI-BLAST uses the same gap costs and λ_g, but applied to the position-specific score matrix constructed from the output of the gapped BLAST run. Only one PSI-BLAST iteration is executed. All three BLAST programs use the same parameter settings as in Table 2, except that T is set to 11. Normalized running times are the mean ratio of program running time to that for the original BLAST. The time for PSI-BLAST includes the time for the initial BLAST search.

Proteins with one or multiple copies of the BRCT domain form a superfamily many of whose members are involved in DNA damage-responsive cell cycle checkpoints.

Table 4. PSI-BLAST protein database search results using the C-terminus of BRCA1 as query

Protein	Species	GenBank ID number	PSI-BLAST iteration	*E*-value
BARD	*Homo sapiens*	1710175	0	2e-06
T10M13.12[a]	*Arabidopsis thaliana*	2104545	1	4e-06
F26D2.b[b]	*Caenorhabditis elegans*	1914176	1	4e-04
KIAA0259[a]	*H.sapiens*	1665785	1	0.001
F37D6.1	*C.elegans*	1418521	2	4e-06
C19G10.07	*Schizosaccharomyces pombe*	1723501	2	6e-05
KIAA0170	*H.sapiens*	1136400	2	0.002
53BP1	*H.sapiens*	488592	2	0.008
T13F2.3[a]	*C.elegans*	1667334	3	2e-07
K04C2.4	*C.elegans*	470351	3	3e-07
T19E10.1	*C.elegans*	1067065	4	7e-04
Rad4/Cut5	*S.pombe*	730470	4	0.002
REV1	*Saccharomyces cerevisiae*	132409	4	0.003
ECT2	*Mus musculus*	423597	5	1e-04
XRCC1	*M.musculus*	627867	5	6e-04
Crb2	*S.pombe*	1449177	5	0.002
RAP1	*S.cerevisiae*	173558	5	0.006
TcEST030[c]	*Trypanosoma cruzi*	1536857	6	0.001
DPB11	*S.cerevisiae*	1352999	6	0.001
L8543.18	*S.cerevisiae*	1078075	6	0.010
SPAC6G9.12[a]	*S.pombe*	1644324	7	4e-04
YM8021.03	*S.cerevisiae*	1078533	7	0.005
YHR154w	*S.cerevisiae*	731729	7	0.008
C36A4.8[a]	*C.elegans*	1657667	7	0.010
UNE452	*S.cerevisiae*	1151000	8	8e-04
DNA ligase IV	*H.sapiens*	1706482	8	0.008
CDC9	*Candida albicans*	1706483	9	0.006
DNA ligase	*Thermus scotoductus*	1352293	10	0.010
GNF1	*Drosophila melanogaster*	544404	11	0.004
mutT[c]	*M.jannaschii*	2129134	15	0.008
RAD9	*S.cerevisiae*	131817	7	0.74
RAP1 homolog	*K.lactis*	422087	9	0.21
ZK675.2	*C.elegans*	599712	13	3.5
D90904[a]	*Synechocystis* sp.	1652299	15	0.17

DNA ligase	*Thermus scotoductus*	1352293	10	0.010
GNF1	*Drosophila melanogaster*	544404	11	0.004
mutT[c]	*M.jannaschii*	2129134	15	0.008
RAD9	*S.cerevisiae*	131817	7	0.74
RAP1 homolog	*K.lactis*	422087	9	0.21
ZK675.2	*C.elegans*	599712	13	3.5
D90904[a]	*Synechocystis* sp.	1652299	15	0.17
TDT	*Mus domestica*	2149634	15	0.46
YGR103w	*S.cerevisiae*	1723693	16	0.017
Pescadillo[a]	*H.sapiens*	2194203	16	0.017
PPOL	*Sarcophaga peregrina*	1709741	16	0.060

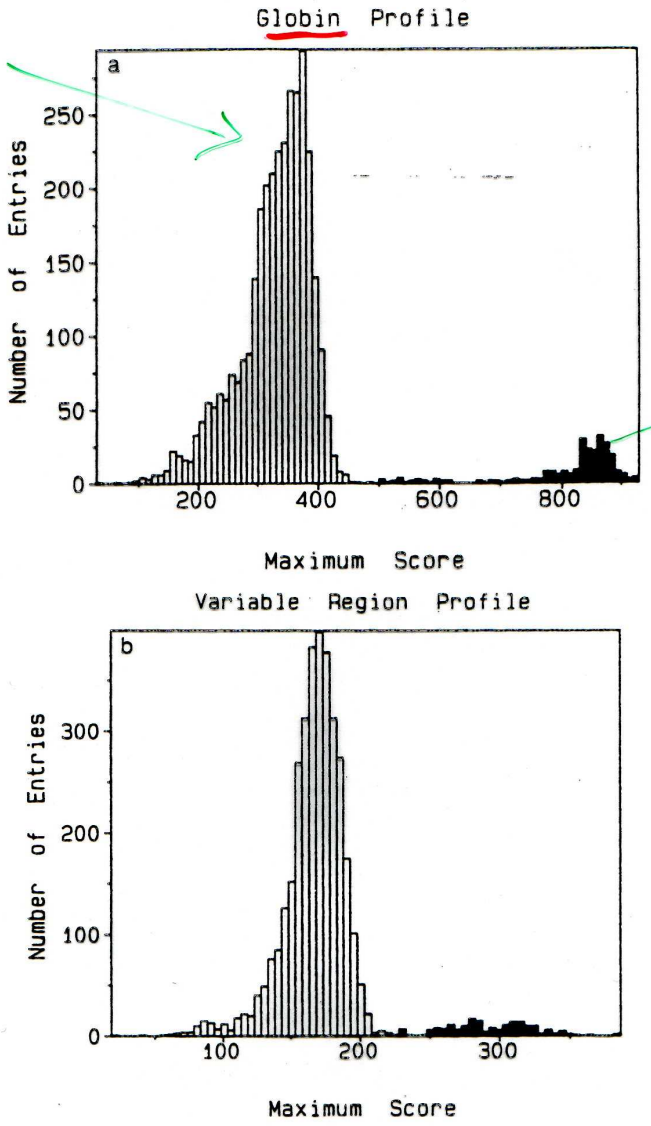
Iteration zero refers to the initial BLAST run, using the 215 C-terminal residues of BRCA1 (68) (SWISS-PROT accession no. P38398) as query. Subsequent PSI-BLAST iterations use derived position-specific score matrices in place of the query. The score matrix for iteration $i + 1$ is constructed from alignments achieving an E-value ≤ 0.01 for iteration i. For each protein, the E-value is that returned during the PSI-BLAST iteration indicated, and precedes the protein's use for score matrix construction. Only one representative is listed for families of closely related proteins. On its 16th iteration PSI-BLAST uncovered no new proteins with E-value ≤ 0.01, and therefore ceased iteration. At the end of the table are shown BRCT proteins returned by PSI-BLAST with E-value >0.01 but ≤ 10, listed for the iteration in which they achieved their lowest E-value.

[a]Recent additions to the database, first identified as BRCT proteins here.

[b]The *C.elegans* F26D2.b protein (74) while a recent addition to the databases, is a close homolog of the previously recognized (66,67) family of *C.elegans* BRCT proteins containing, for example, F37A4.4 (90).

[c]The trypanosome EST (70) and the *M.jannaschii* mutT protein (71) are the only likely false positives.

non globin sequences



globin sequence

FIG. 2. Profile analysis of globin and immunoglobulin sequences. (*a*) Globins: Distribution of scores for the comparison of a profile generated from human α-hemoglobin, rhesus monkey β-hemoglobin, human myoglobin, lamprey cyanohemoglobin, and soybean leghemoglobin, to the PIR database. The profile is 124 residues long, with an average of 5.6 gapped positions per sequence. Scores for nonglobin proteins ranged from 32.4 to 455.8, with a mean of 326.9 and SD of 60. Scores for globin sequences (shaded) ranged from 469.5 to 928.7. (*b*) Variable region immunoglobulin domain: Distribution of scores for the comparison of a profile generated from 49-residue (including

PSI-BLAST Assessment

Given known distant family members that are not identified by FASTA and BLAST, what fraction is identified by PSI-BLAST? ~25%-45%

Detection is not symmetric

Possible problems
after several iterations, accumulation of distant sequences might insert errors ("profile drift")

false homologs.

PSI-BLAST: the problem of corruption

Corruption is defined as the presence of at least one false positive alignment with an E value < 10^{-4} after five iterations.

Three approaches to stopping corruption:

[1] Apply filtering of biased composition regions

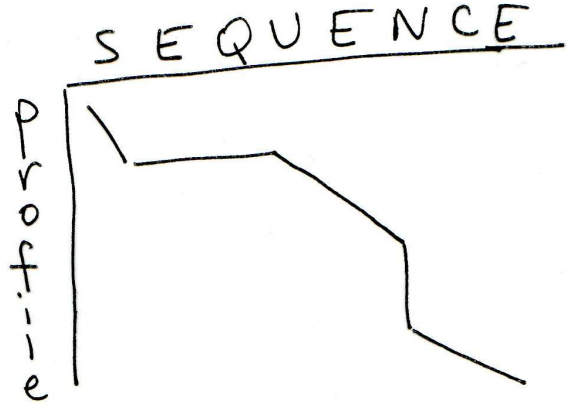
[2] Adjust E value from 0.001 (default) to a lower value such as E = 0.0001.

[3] Visually inspect the output from each iteration. Remove suspicious hits by unchecking the box.

Algorithmes de recherche de motifs ou d'homologie en général

- seq - seq alignment
- Profile - Seq alignment. Dynamic Programming

S E Q U E N - C E
p r - - o f i l e



- ~~Seq~~ Profile - Profile alignment