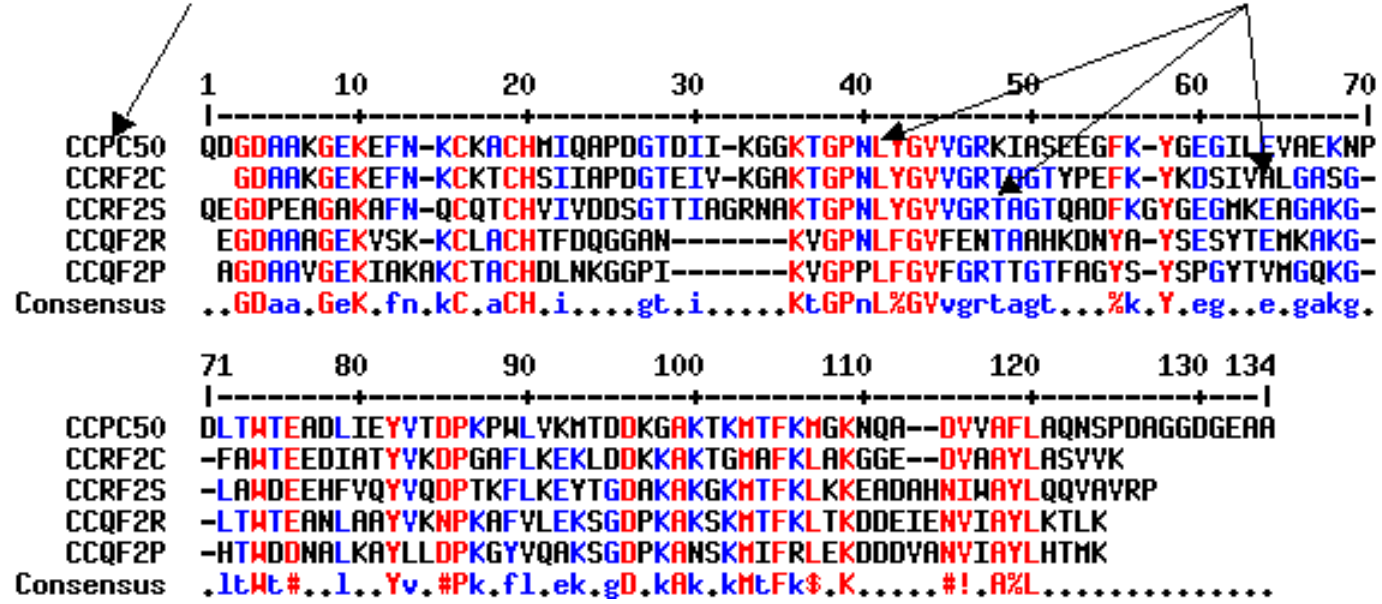
Alignements multiples

- identifications de motifs conservés dans un ensemble de séquences
- Relation évolutives entre séq / organismes
- Localiser des structures 2D ou 3D
- Caractériser une famille de protéines

  => Identification de nouveaux membres



Application des alignements multiples :

- Identification et caractérisation des motifs (*patterns,* signature) et des domaines protéiques (*profile,* matrice)

GHEGVGKVVKLGAGA
GHEKKGYFEDRGPSA
GHEGYGGRSRGGGYS
GHEFEGPKGCGALYI
GHELRGTTFMPALEC

GHE--G--------- Consensus à 100%
GHE--G-----G--- Consensus à 60%

GHE-x(2)-G-x(5)-[GA] *Pattern* PROSITE ou signature

Matrice PSSM ou profile

GHEG-GKVVKLGAGA

GHE--GYFEDRGPSA

GHEGYGGRSRGGGYS

GHEF-GPKG-GALYI

GHELRGTTFMPALEC

GHE-x(0,2)-G-x(4,5)-[GA]

<A-x-[ST](2)-x(0,1)-[APTL]-x(4,10)-C-{V}

<A	**en N term**
x	**= n'importe quel AA**
[ST](2)	**=Ser ou Thr 2 fois**
x(0,1)	**1 AA ou aucun**
x(4,10)	**entre 4 et 10 AA quelconques**
{V}	**tout sauf une Val**

NiceSite View of PROSITE: PS00191

General information about the entry	
Entry name	CYTOCHROME_B5_1
Accession number	PS00191
Entry type	PATTERN
Date	APR-1990 (CREATED); DEC-2004 (DATA UPDATE); SEP-2005 (INFO UPDATE).
PROSITE documentation	PDOC00170

Pattern

Name and characterization of the entry	
Description	Cytochrome b5 family, heme-binding domain signature.
Pattern	[FY]-[LIVMK]-{I}-{Q}-H-P-[GA]-G.

[FY]-[LIVMK]-{I}-{Q}-H-P-[GA]-G

Numerical results

- UniProtKB/Swiss-Prot release number: **48.1**, total number of sequence entries in that release: **195058**.
- Total number of hits in UniProtKB/Swiss-Prot: **86 hits in 86 different sequences**
- Number of hits on proteins that are known to belong to the set under consideration: **80 hits in 80 different sequences**
- Number of hits on proteins that could potentially belong to the set under consideration: **0 hits in 0 different sequences**
- Number of false hits (on unrelated proteins): **6 hits in 6 different sequences**
- Number of known missed hits: **4**
- Number of partial sequences which belong to the set under consideration, but which are not hit by the pattern or profile because they are partial (fragment) sequences: **2**
- Precision (true hits / (true hits + false positives)): **93.02 %**
- Recall (true hits / (true hits + false negatives)): **95.24 %**

Comments

- Taxonomic range: **Eukaryotes, Prokaryotes (Bacteria), Eukaryotic viruses**
- Maximum known number of repetitions of the pattern in a single protein: **1**
- 'Interesting' site in the pattern: **5,heme_iron**
- VERSION: **1**

Cross-references

True positive hits:

ACO1_AJECA (Q12618), ACO1_YEAST (P21147), CYB2_HANAN (P09437),
CYB2_YEAST (P00175), CYB51_ARATH (Q42342), CYB52_ARATH (O48845),
CYB52_SCHPO (Q9USM6), CYB5L_MIMIV (Q5UR80), CYB5L_NEUCR (Q8X0J4),
CYB5M_HUMAN (O43169), CYB5M_MOUSE (Q9CQX2), CYB5M_PONPY (Q5RDJ5),
CYB5M_RAT (P04166), CYB5R_DROME (P18967), CYB5R_DROWI (B5D266),

Alignement

```
AATTGA
AGGTCC
AGGATG
AGGCGT
```

Matrice de position

	1	2	3	4	5	6
A	4	1	0	1	0	1
C	0	0	0	1	1	1
G	0	3	3	0	2	1
T	0	0	1	2	1	1

Matrice de fréquences

	1	2	3	4	5	6
A	1	.25	0	.25	0	.25
C	0	0	0	.25	.25	.25
G	0	.75	.75	0	.5	.25
T	0	0	.25	.5	.25	.25

Matrice de fréquences

$$\log\left[\frac{f_b}{p_b}\right]$$

Matrice de poids de position (Position Weight Matrix ou Position Specific Scoring Matrix= PSSM)

	1	2	3	4	5	6
A	1.2	0	-1.6	0	-1.6	0
C	-1.6	-1.6	-1.6	0	0	0
G	-1.6	0.96	0.96	-1.6	0.59	0
T	-1.6	-1.6	0	0.59	0	0

C G T A T G T A A G G T G T A C G T A G

	1	2	3	4	5	6
A	1.2	0	-1.6	0	-1.6	0
C	-1.6	-1.6	-1.6	0	0	0
G	-1.6	0.96	0.96	-1.6	0.59	0
T	-1.6	-1.6	0	0.59	0	0

C G T A T G T A A G G T G T A C G T A G

	1	2	3	4	5	6
A	1.2	0	-1.6	0	-1.6	0
C	-1.6	-1.6	-1.6	0	0	0
G	-1.6	0.96	0.96	-1.6	0.59	0
T	-1.6	-1.6	0	0.59	0	0

T A A G G T

Score = -4.21

A A G G T G

Score =0.56

A G G T G T

Score = 4.3

pfam00173: Cyt-b5 ?

Cytochrome b5-like Heme/Steroid binding domain. This family includes heme binding domains from a diverse range of proteins. This family also includes proteins that bind to steroids. The family includes progesterone receptors. Many members of this subfamily are membrane anchored by an N-terminal transmembrane alpha helix. This family also includes a domain in some chitin synthases. There is no known ligand for this domain in the chitin synthases.

- Links ?
- Statistics ?
- Structure ?

PubMed References ?

- Purification and partial sequencing of high-affinity progesterone-binding site(s) from porcine liver membranes. *Eur. J. Biochem. 1996 Aug 1; 239(3):726-731*
- Cloning and tissue expression of two putative steroid membrane receptors. *Biol. Chem. 1998 Jul; 379(7):907-911*
- Membrane-bound progesterone receptors contain a cytochrome b5-like ligand-binding domain. *Genome Biol. 2002; 3(12):RESEARCH0068*

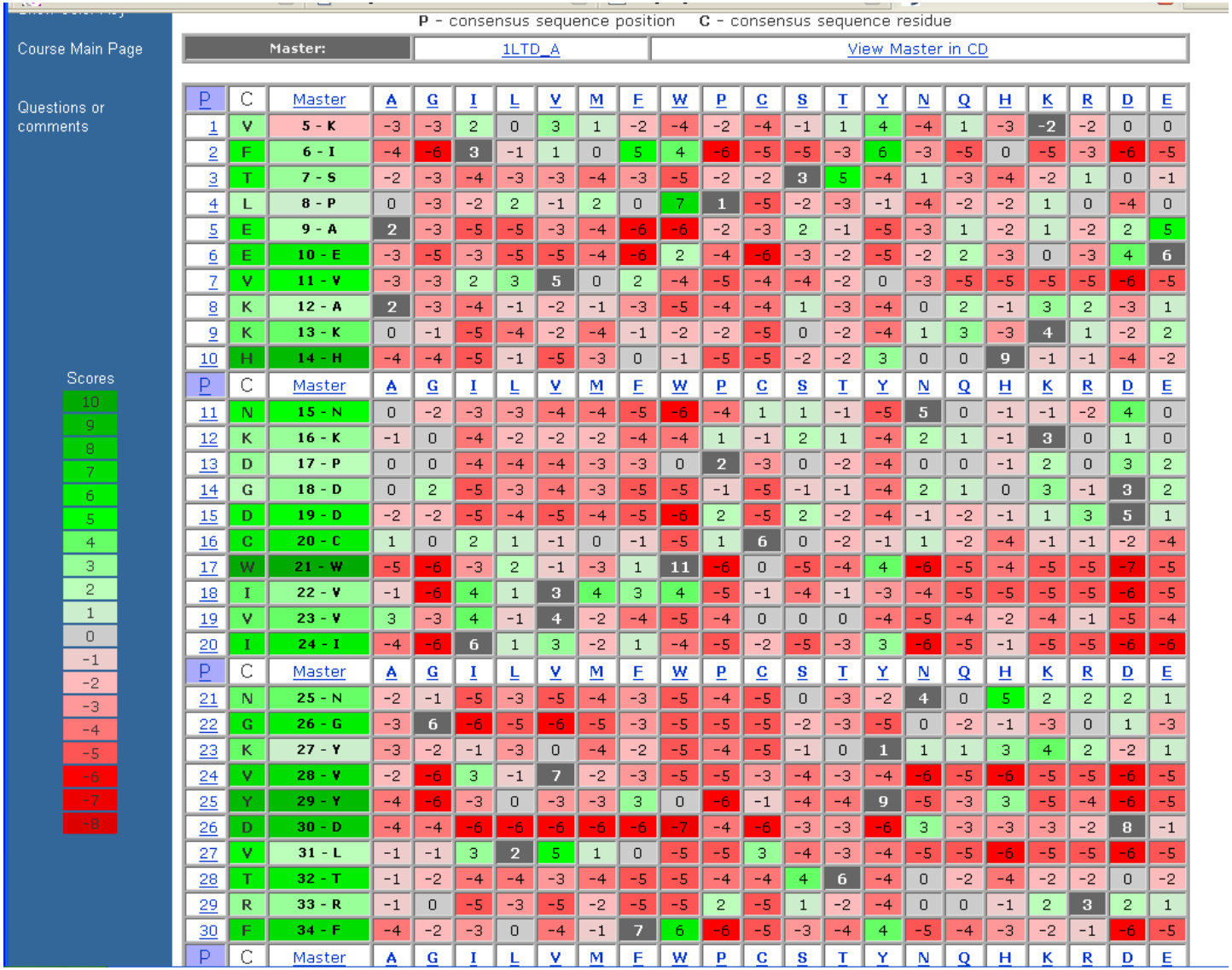
Sequence Alignment ?

Reformat | Format: Hypertext | Row Display: up to 10 | Color Bits: 2.0 bit | Type Selection: the most diverse members

```
                          10        20        30        40        50        60        70        80
                    ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....|
1LTD_A          5 KISPAEVAKHN--KPDDCWVVINGYVYDLTR-FLPNHPGG--------------QDVIKFNAGKDVTAIF---------- 57
gi 82111907    38 DFTLADLKPYDglQDPRILMAVNGKVFDVTRgKKFYGPEG----------------PYGVFAGRDASRGL---------- 91
gi 46577676    72 DFTPAELRRFDgvQDPRILMAINGKVFDVTKgRKFYGPEG----------------PYGVFAGRDASRGLatfcldkeal 135
gi 75024827    64 DMTVEELRKYDgvKNEHILFGLNGTIYDVTR-GKGFYGPG----------------KAYGTLAGHDATRALgtmdqnavss 127
gi 91206848  1290 YVRRADMENLL--LDGSRCIILAGYVCDLSG-YNCESETL--------------RSVLDSGLGKDLTAEMs--------- 1343
gi 74739702  1209 LIRKADLENHN--KDGGFWTVIDGKVYDIKD-FQTQSLTG--------------NSILAQFAGEDPVVAL---------- 1261
gi 74582634   303 YYNWTDI--HE---PGTSLMVFNGNVLDLSR-LRYLTPNIplpiq-----iaqiVGPGSAFIGRDATYWLs--------- 362
gi 148887356  407 YFTWADIRNNS-----RNLFVYSGNVLDLDL-LFWFNRDQvniprrfeelrdknNAANRAIRGRDATRTF---------- 470
gi 44889038   372 YFTWDDIKNSS-----RNLVVYSGHVLDLDL-LHWFNDTQvtyparfkelrdknTAGNQAIRGRDITHAF---------- 435
gi 122065155  402 QVSLQWNNVTD---PARNLAVYRGSVLDLNR-LNNLTTGLsypel-----ydtlKRRNDSWAGRDVTSAV---------- 462

                          90       100       110
                    ....*....|....*....|....*....|....*..
1LTD_A         58 --------------EPLHApNVIDKYIAPeKKLGPLQ 80
gi 82111907    92 --------------ATFCLeKDALKDEHD-----DLS 109
```

Pfam : PSSM



P - consensus sequence position C - consensus sequence residue

Course Main Page

Questions or comments

Master: 1LTD_A View Master in CD

P	C	Master	A	G	I	L	V	M	F	W	P	C	S	T	Y	N	Q	H	K	R	D	E
1	V	5 - K	-3	-3	2	0	3	1	-2	-4	-2	-4	-1	1	4	-4	1	-3	-2	-2	0	0
2	F	6 - I	-4	-6	3	-1	1	0	5	4	-6	-5	-5	-3	6	-3	-5	0	-5	-3	-6	-5
3	T	7 - S	-2	-3	-4	-3	-3	-4	-3	-5	-2	-2	3	5	-4	1	-3	-4	-2	1	0	-1
4	L	8 - P	0	-3	-2	2	-1	2	0	7	1	-5	-2	-3	-1	-4	-2	-2	1	0	-4	0
5	E	9 - A	2	-3	-5	-5	-3	-4	-6	-6	-2	-3	2	-1	-5	-3	1	-2	1	-2	2	5
6	E	10 - E	-3	-5	-3	-5	-5	-4	-6	2	-4	-6	-3	-2	-5	-2	2	-3	0	-3	4	6
7	V	11 - V	-3	-3	2	3	5	0	2	-4	-5	-4	-4	-2	0	-3	-5	-5	-5	-5	-6	-5
8	K	12 - A	2	-3	-4	-1	-2	-1	-3	-5	-4	-4	1	-3	-4	0	2	-1	3	2	-3	1
9	K	13 - K	0	-1	-5	-4	-2	-4	-1	-2	-2	-5	0	-2	-4	1	3	-3	4	1	-2	2
10	H	14 - H	-4	-4	-5	-1	-5	-3	0	-1	-5	-5	-2	-2	3	0	0	9	-1	-1	-4	-2
P	C	Master	A	G	I	L	V	M	F	W	P	C	S	T	Y	N	Q	H	K	R	D	E
11	N	15 - N	0	-2	-3	-3	-4	-4	-5	-6	-4	1	1	-1	-5	5	0	-1	-1	-2	4	0
12	K	16 - K	-1	0	-4	-2	-2	-2	-4	-4	1	-1	2	1	-4	2	1	-1	3	0	1	0
13	D	17 - P	0	0	-4	-4	-4	-3	-3	0	2	-3	0	-2	-4	0	0	-1	2	0	3	2
14	G	18 - D	0	2	-5	-3	-4	-3	-5	-5	-1	-5	-1	-1	-4	2	1	0	3	-1	3	2
15	D	19 - D	-2	-2	-5	-4	-5	-4	-5	-6	2	-5	2	-2	-4	-1	-2	-1	1	3	5	1
16	C	20 - C	1	0	2	1	-1	0	-1	-5	1	6	0	-2	-1	1	-2	-4	-1	-1	-2	-4
17	W	21 - W	-5	-6	-3	2	-1	-3	1	11	-6	0	-5	-4	4	-6	-5	-4	-5	-5	-7	-5
18	I	22 - V	-1	-6	4	1	3	4	3	4	-5	-1	-4	-1	-3	-4	-5	-5	-5	-5	-6	-5
19	V	23 - V	3	-3	4	-1	4	-2	-4	-5	-4	0	0	0	-4	-5	-4	-2	-4	-1	-5	-4
20	I	24 - I	-4	-6	6	1	3	-2	1	-4	-5	-2	-5	-3	3	-6	-5	-1	-5	-5	-6	-6
P	C	Master	A	G	I	L	V	M	F	W	P	C	S	T	Y	N	Q	H	K	R	D	E
21	N	25 - N	-2	-1	-5	-3	-5	-4	-3	-5	-4	-5	0	-3	-2	4	0	5	2	2	2	1
22	G	26 - G	-3	6	-6	-5	-6	-5	-3	-5	-5	-5	-2	-3	-5	0	-2	-1	-3	0	1	-3
23	K	27 - Y	-3	-2	-1	-3	0	-4	-2	-5	-4	-5	-1	0	1	1	1	3	4	2	-2	1
24	V	28 - V	-2	-6	3	-1	7	-2	-3	-5	-5	-3	-4	-3	-4	-6	-5	-6	-5	-5	-6	-5
25	Y	29 - Y	-4	-6	-3	0	-3	-3	3	0	-6	-1	-4	-4	9	-5	-3	3	-5	-4	-6	-5
26	D	30 - D	-4	-4	-6	-6	-6	-6	-6	-7	-4	-6	-3	-3	-6	3	-3	-3	-3	-2	8	-1
27	V	31 - L	-1	-1	3	2	5	1	0	-5	-5	3	-4	-3	-4	-5	-5	-6	-5	-5	-6	-5
28	T	32 - T	-1	-2	-4	-4	-3	-4	-5	-5	-4	-4	4	6	-4	0	-2	-4	-2	-2	0	-2
29	R	33 - R	-1	0	-5	-3	-5	-2	-5	-5	2	-5	1	-2	-4	0	0	-1	2	3	2	1
30	F	34 - F	-4	-2	-3	0	-4	-1	7	6	-6	-5	-3	-4	4	-5	-4	-3	-2	-1	-6	-5
P	C	Master	A	G	I	L	V	M	F	W	P	C	S	T	Y	N	Q	H	K	R	D	E

Scores

10
9
8
7
6
5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-7
-8

wellcome trust sanger institute

Pfam : Logo

Pfam keyword search Go

Family: *Cyt-b5* (PF00173)

Loading page components (1 remaining)...

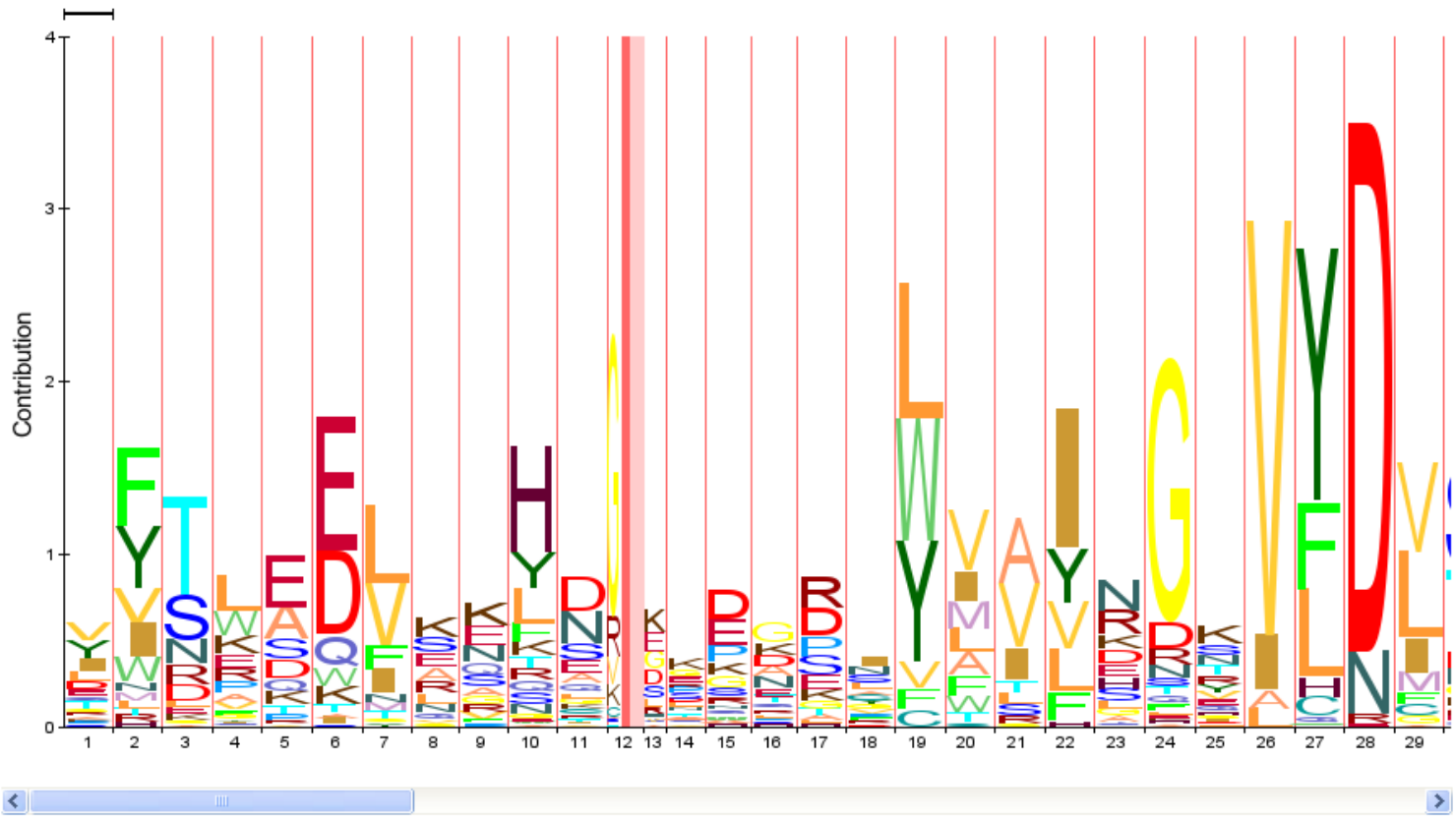
60 architectures 1547 sequences 2 interactions 316 species 63 structures

- Summary
- Domain organisation
- Alignments
- HMM logo
- Trees
- Curation & models
- Species
- Interactions
- Structures

Jump to... enter ID/acc Go

HMM logo

HMM logos are one way of visualising profile HMMs. They provide a quick overview of the properties of an HMM in a graphical form. You can see a more detailed description of HMM logos and find out how you can interpret them here. More...



Pfam : sequence search

wellcome trust sanger institute

HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

Pfam keyword search Go

Sequence search results

Show the detailed description of this results page.

We found **1** Pfam-A match to your search sequence (**all** significant). You did not choose to search for Pfam-B matches.

Show the search options and sequence that you submitted.

Return to the search form to look for Pfam domains on a new sequence.

Significant Pfam-A Matches

Show or hide all alignments.

Family	Description	Entry type	Clan	Envelope		Alignment		HMM		Bit score	E-value	Predicted active sites	Show/hide alignment
				Start	End	Start	End	From	To				
Cyt-b5	Cytochrome b5-like Heme/Steroid binding domain	Domain	n/a	16	90	16	89	1	75	82.9	8.2e-24	n/a	Hide

```
#HMM      vftleevskhnkekdswvvirgkVYDvskflkdhpggasailafaGkDaTeafenashseaaeklllekyyrvGrl
#MATCH    ++ leev+khn++++ w+++++++YD++kfl++hpgg++++++ aG DaTe fe+  hs++a+ l e+ + +G+l
#PP       799*********************************************************.****98
#SEQ      YYRLEEVQKHNNSQSTWIIVHHRIYDITKFLDEHPGGEEVLREQAGGDATENFEDVGHSTDARALSET-FIIGEL
```

Ac. aspartique

```
Seq 1   QWESTATHNYDQP
        ||||  ||||:||
Seq 2   QWESR-THNYEQP
```

Ac. glutamique

	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	4															
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	-4	-4	-4	-6	-4	-5	-5	-5	-2	1	2	-5	0	9						
P	1	0	-1	-1	-3	0	-1	-1	0	-2	-3	-1	-2	-5	6					
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	3				
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-2	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

Matrice de substitution (PAM250)

$$\text{Score} = \sum_{se} - \sum_{p}$$

$$p = c + e * l$$

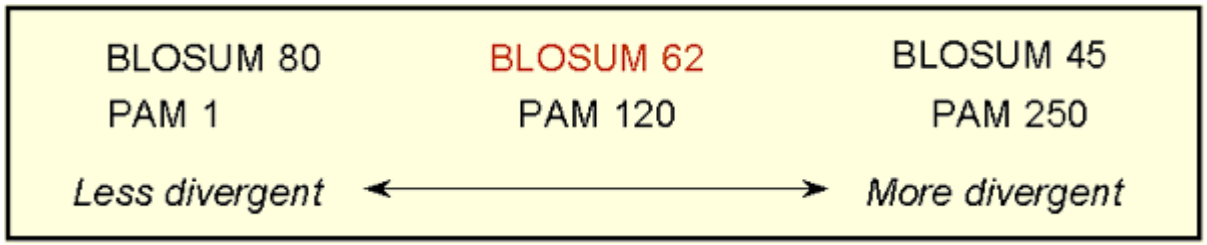
(10) (1)

$\Sigma se = 4+17+4+3-1+3+6+2+10+3+4+6 = 61$

$p = 10 + 1*0$

Score = 51

Matrices protéiques liées à l'évolution



$$se_{i,j} = \log_2 \left(\frac{Fobs_{i,j}}{Fthéo_{i,j}} \right)$$

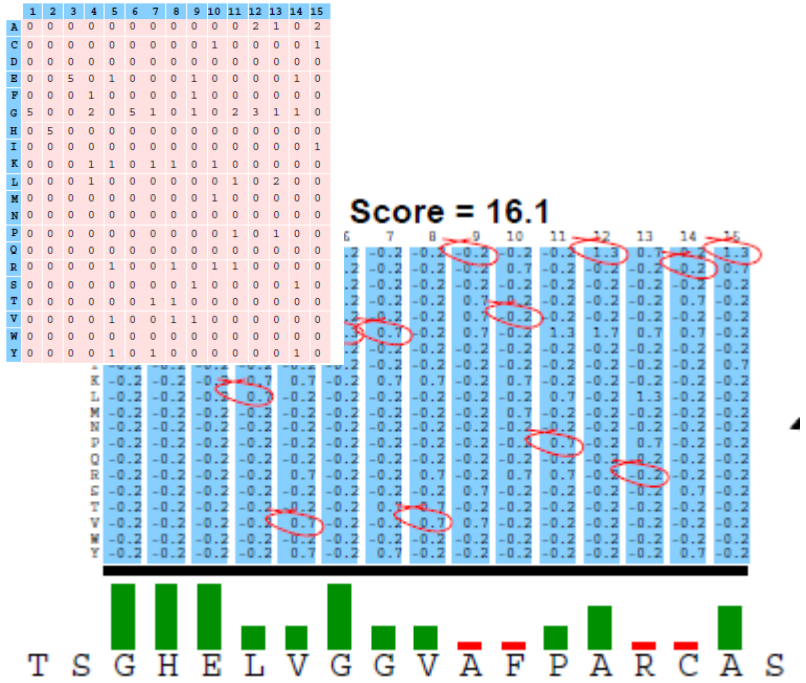
> 0 substitution favorisée au cours de l'evolution de l'AA i par l'AA j

< 0 substitution défavorisée

GHEGVGKVVKLGAGA
GHEKKGYFEDRGPSA
GHEGYGGRSRGGGYS
GHEFEGPKGCGALYI
GHELRGTTFMPALEC

GHE--G---------
GHE--G-----G---

GHE-x(2)-G-x(5)-[GA]



consensus

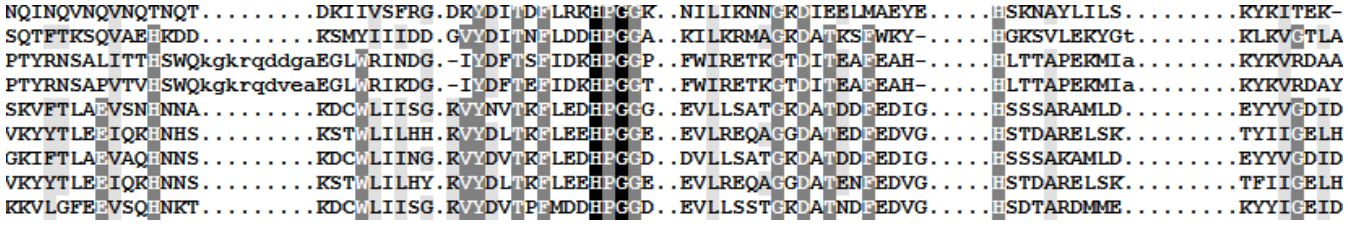
très restrictif (spécifique)
Pas de score

***Pattern* ou signature (PROSITE)**

\+ souple (+sensible)
Bcp de FP sur petit *pattern*
Gd *pattern* difficile à écrire
pas de # fréq des résidus
Pas de score

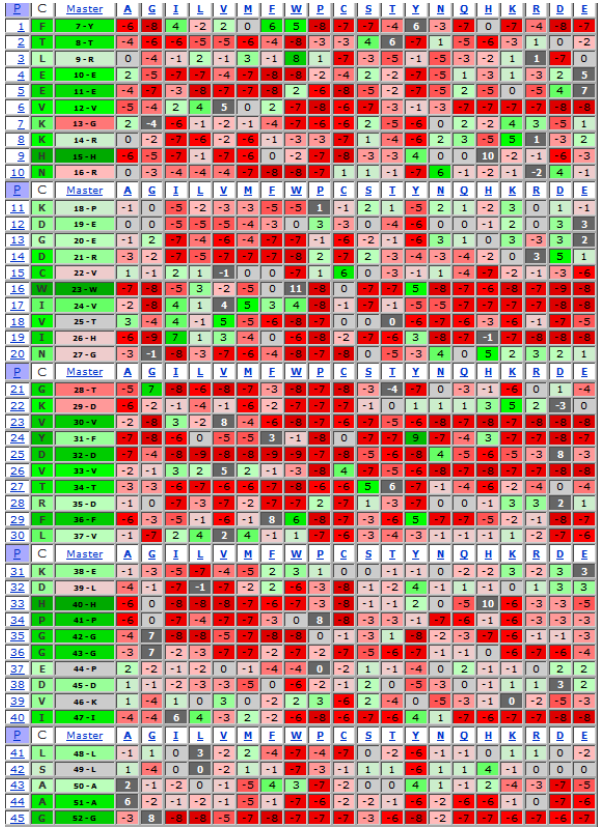
matrice de poids de position (PSSM) ou profile

++ sensible
Nécessité de bcp de séq
OK sur régions larges
Score



Domaine (cytB5, 74 aa)

motif



Matrice PSSM ou profile

PFAM, SMART, PROSITE

Banques de domaines

Pattern ou signature (PROSITE)

[FY]-[LIVMK]-{I}-{Q}-H-P-[GA]-G

Signatures PROSITE

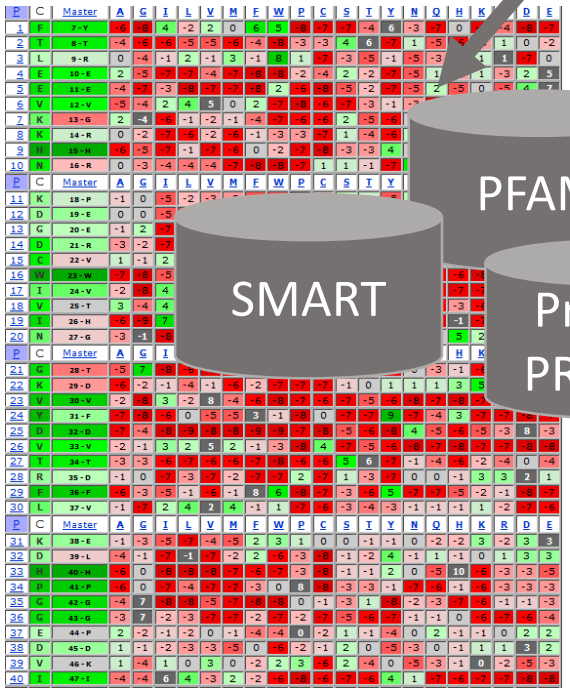
Ma sequence contient-elle un domaine ou motif connu ?

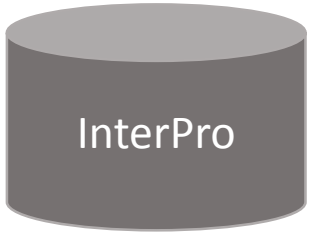
```
APSYPEYTREEVGRHRSPEERVWVTHGTDVFDVTDFVELHPGGPDKILLAAGGALEPFWALYAVHG
EPHVLELLQQYKVGELSPDEAPAAPDAQDPFAGDPPRHPGLRVNSQKPFNAEPPAELLAERFLTPN
ELFFTRNHLPVPAVEPSSYRLRVDGPGGGTLSLSLAELRSRFPKHEVTATLQCAGNRRSEMSRVRP
VKGLPWDIGAISTARWGGARLRDVLLHAGFPEELQGEWHVCFEGLDADPGGAPYGASIPYGRALSP
AADVLLAYEMNGTELPRDHGFPVRVVVPGVVGARSVKWLRRVAVSPDESPSHWQQNDYKGFSPCVD
WDTVDYRTAPAIQELPVQSAVTQPRPGAAVPPGELTVKGYAWSGGGREVVRVDVSLDGGRTWKVAR
LMGDKAPPGRAWAWALWELTVPVEAGTELEIVCKAVDSSYNVQPDSVAPIWNLRGVLSTAWHRVRV
```

ScanProsite,
Pfamsearch…

InterPro Scan

ScanProsite





Si Score >T
Domaine trouvé

```
APSYPEYTREEVGRHRSPEERVWVTHGTDVFDVTDFVELHPGGPDKILLAAGGALEPFWALYAVHG
```

36-43 FVELHPGG

[FY]-[LIVMK]-{I}-{Q}-H-P-[GA]-G

PFAM

Pssm-ID: 249651 Cd Length: 74 Bit Score: 74.54 E-value: 1.22e-16

```
                 10        20        30        40        50        60        70
                 ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*.
lcl|TempId       7 YTREEVGRHRSPEERVWVTHGTdVFDVTDFVELHPGGPDKILLAAGGALEPFWaLYAVHGEpHVLELLQQYKVGEL 82
Cdd:pfam00173    1 FTLEEVKKHNKDGDCWIVINGK-VYDVTKFLKDHPGGEDVILSAAGKDATEAF-EDAIHSE-AARKLLEKYRVGEL 73
```

PROSITE

4 - 83: score = 20.296

```
YPEYTREEVGRHRSPeERVWVTHGTDVFDVTDFVELHPGGPDKILLAAGGALEPFWALYA
vHGEPHVLELLqQYKVGELS
```

Predicted features:

DOMAIN	4	83	Cytochrome b5 heme-binding

Recherche de similarités dans les banques de séquences

- Séquence déjà connue ?
- Ce gène appartient-il à une famille ?
- Homologues chez d'autres organismes ?

Recherche de similarités

- Heuristique = méthode permettant d'obtenir rapidement **une bonne solution** sans être assuré que cette solution soit la meilleure

=> simplifications/restrictions du problème en faisant des hypothèses (+/- liées à information biologique)

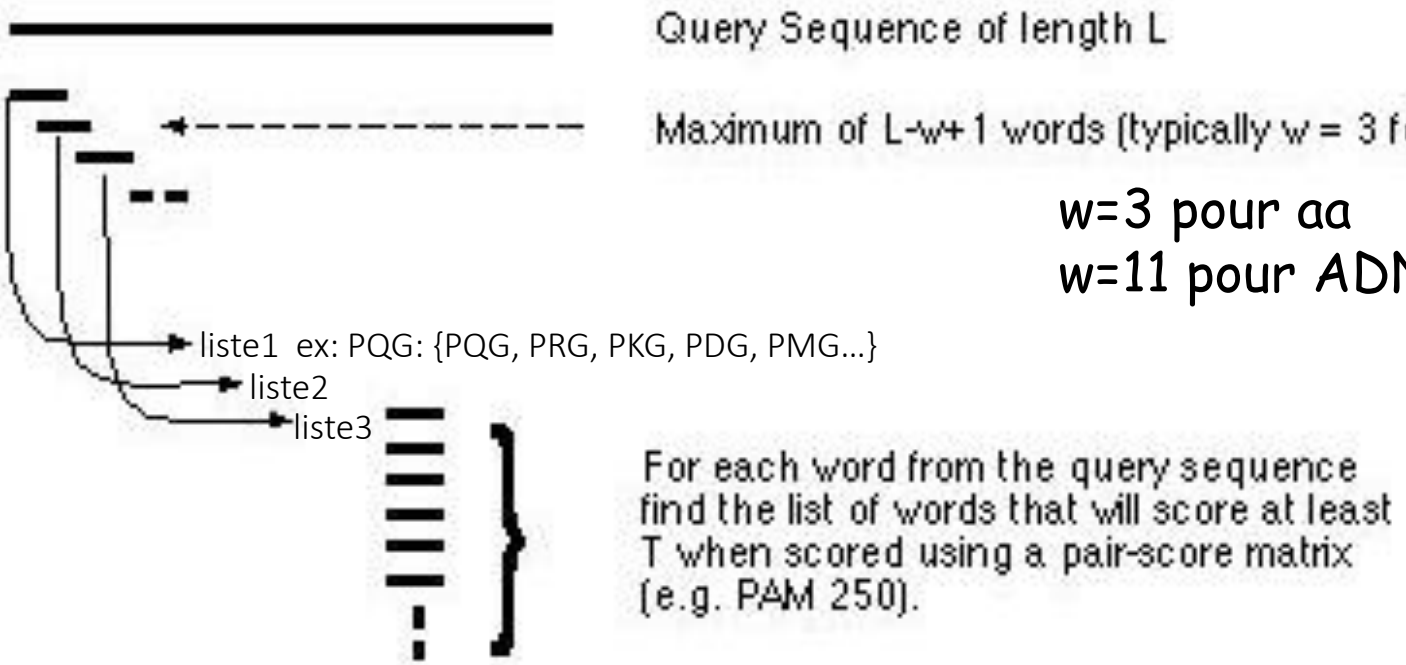
- Surtout pour traiter un grand nombre de séquences et d'alignements

- **programme BLAST**
  Basic **L**ocal **A**lignment **S**earch **T**ool

Algorithme de BLAST (1)

(Altschul *et al.*, 1990)

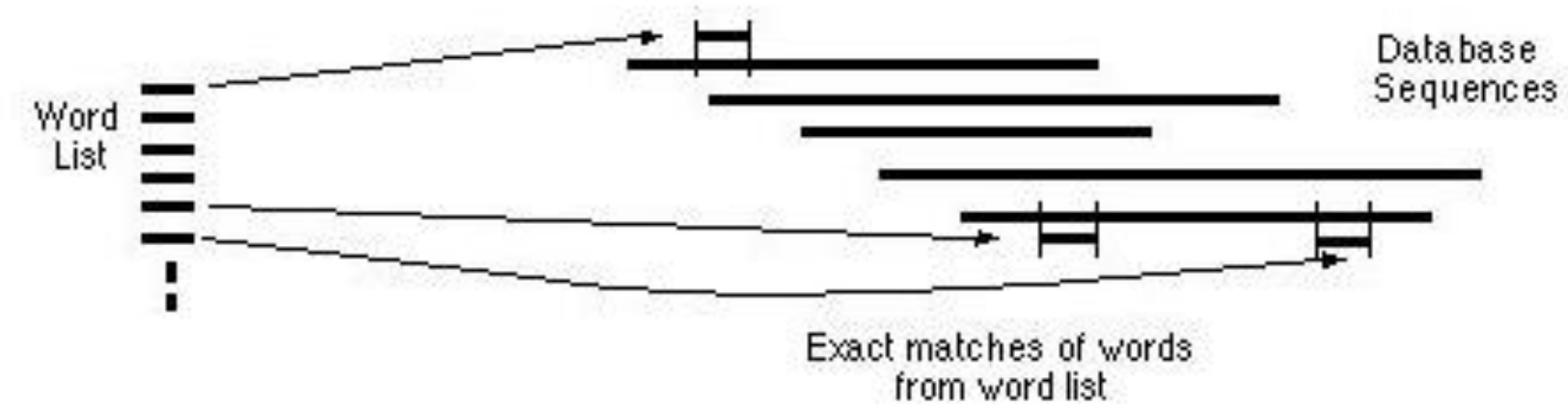
(1) For the query, find the list of high scoring words of length w



w=3 pour aa
w=11 pour ADN

Algorithme de BLAST (2)

(2) Compare the word list to the database and identify exact matches



S L A A L L N K C K T P Q G Q R L V N Q W

Mots m (w=3)

20x20x20=8000 mots

Liste

de mots

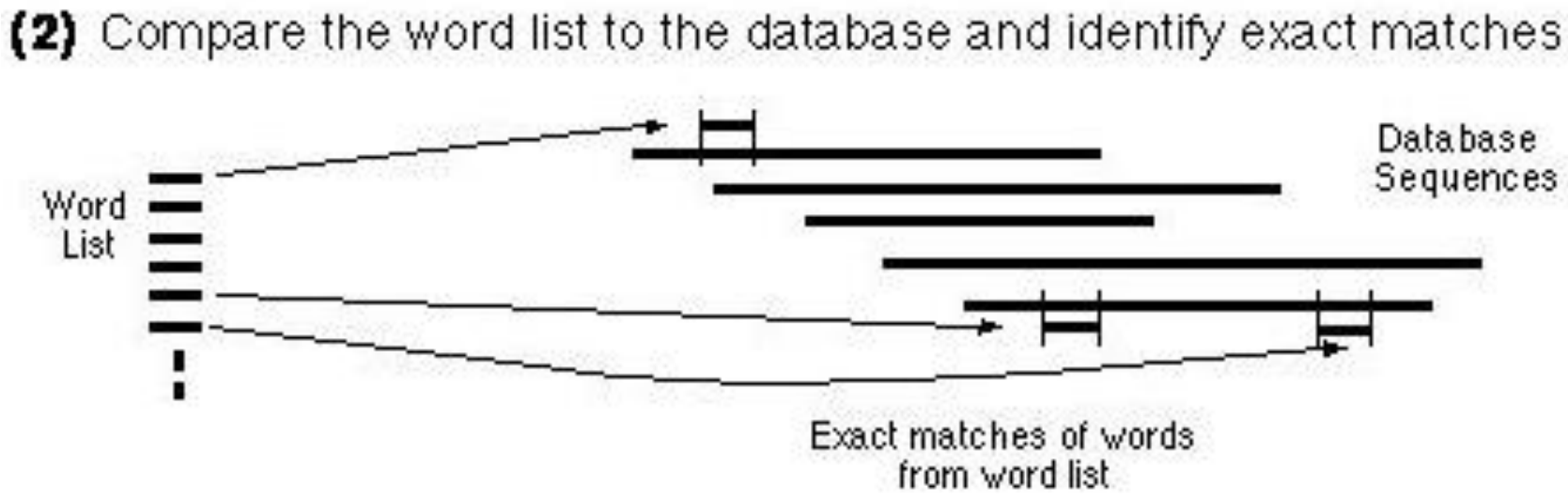
synonymes

P Q G	18
P E G	15
P R G	14
P K G	14
P N G	13
P D G	13
P H G	13
P M G	13
P S G	13
P Q A	12
P Q N	12
. . .	

S(**P**,**P**) = 7
S(**Q**,**R**) = 1
S(**G**,**G**) = 6

Score seuil *T* = 13

Algorithme de BLAST (3)

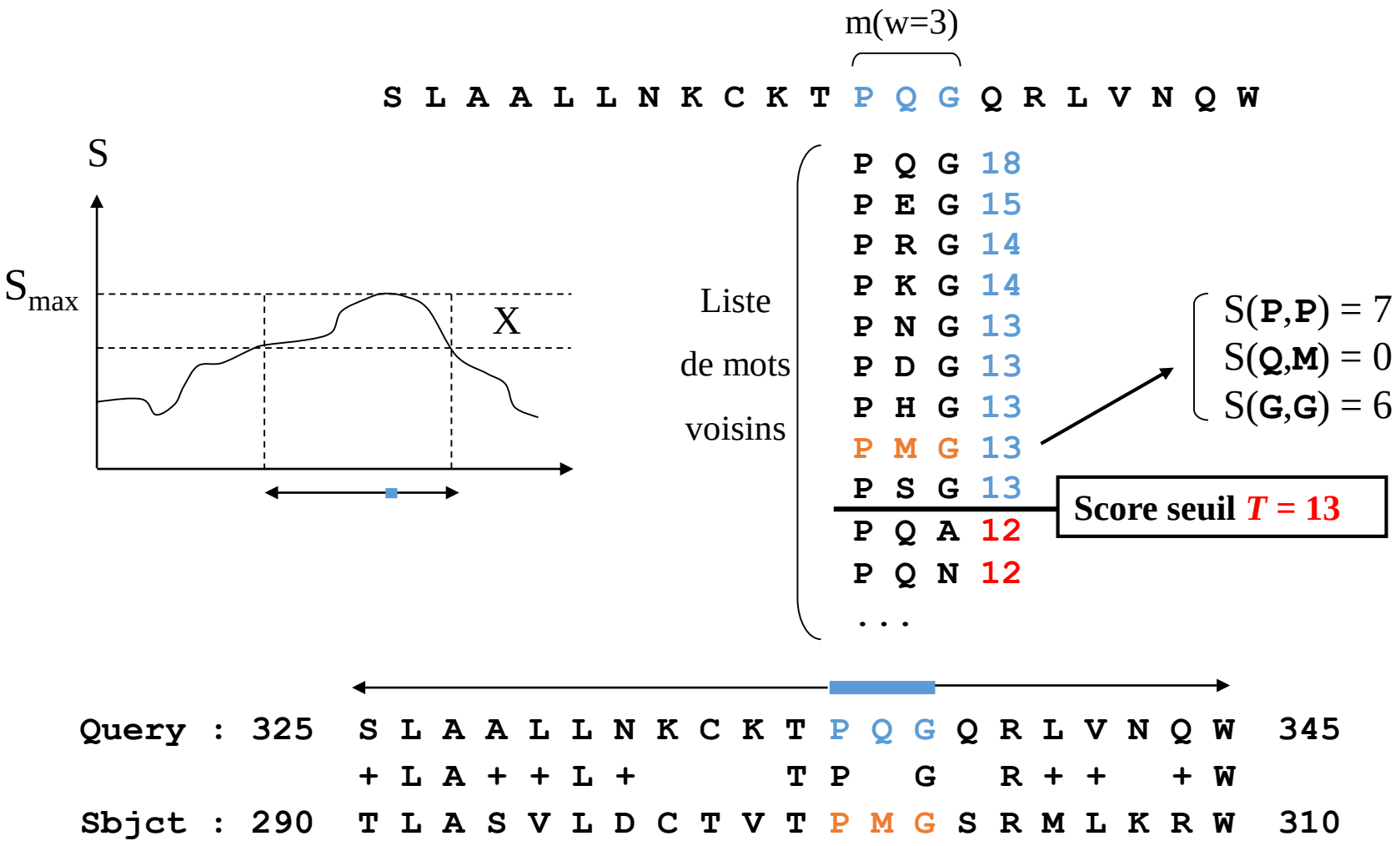



```
Query : 325  S L A A L L N K C K T P Q G Q R L V N Q W  345

Sbjct : 290  T L A S V L D C T V T P M G S R M L K R W  310
```

High Scoring Pairs (HSP)

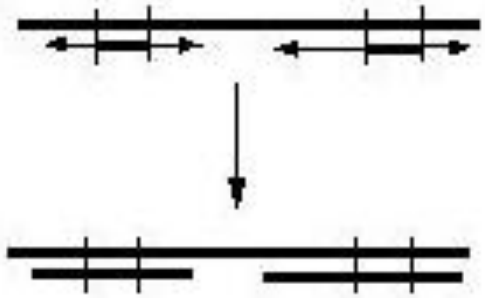
Algorithme de BLAST (4)



High Scoring Pairs (HSP)

Algorithme de BLAST (5)

(3) For each word match, extend the alignment in both directions to find alignments that score greater than a threshold of value S



HSP = High-scoring Segment Pairs

Figure from Barton, G.J. Protein Sequence Alignment and Database Scanning (University of Oxford, Laboratory of Molecular Biophysics)

Arrêt de l'extension si :

- score descend de x par rapport à la valeur maximale atteinte
- score < 0
- fin d'une des 2 séquences

Significativité des alignements

3 mesures : le score, la P-value, la E-value

- **Le Score**

le score S' est dérivé du score brut de l'alignement.

Il a été normalisé et peut donc être utilisé pour comparer des scores provenant de recherches différentes.

$S' = (\lambda S - \ln K)/\ln 2$ “bit score”

K : paramètre lié à la composition (“bruit de fond”)

λ : paramètre lié au système de score

- **La E-value** (Expected) :

nombre d'alignements différents que l'on peut espérer trouver dans les banques avec un score supérieur ou égal à S

Nombre attendu de HSP dont score >= S est :

$$E = Kmne^{-\lambda S}$$

m : longueur seq 1
n : longueur seq 2

$$E=mn2^{-S'}$$

Contre une banque de séquences :
n => N (longueur totale de la base)

- Plus la E-value est faible, plus l'alignement est significatif.

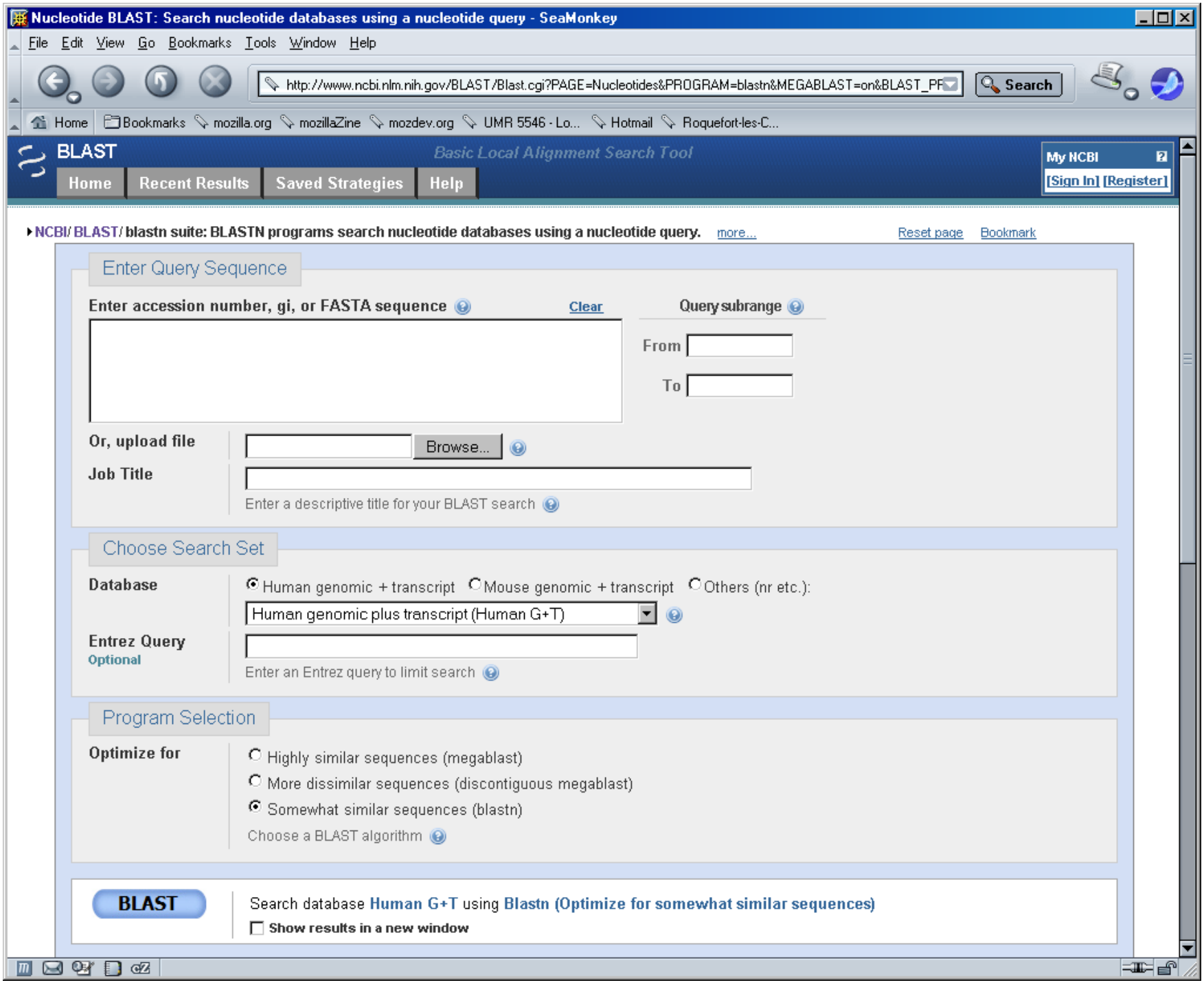
E-value	Interprétation
< e-100	**match exact**
e-100...e-50	**gènes quasiment identiques**
e-50...0.1	**relation plus lointaines**
> 0.1	**pas de relation**

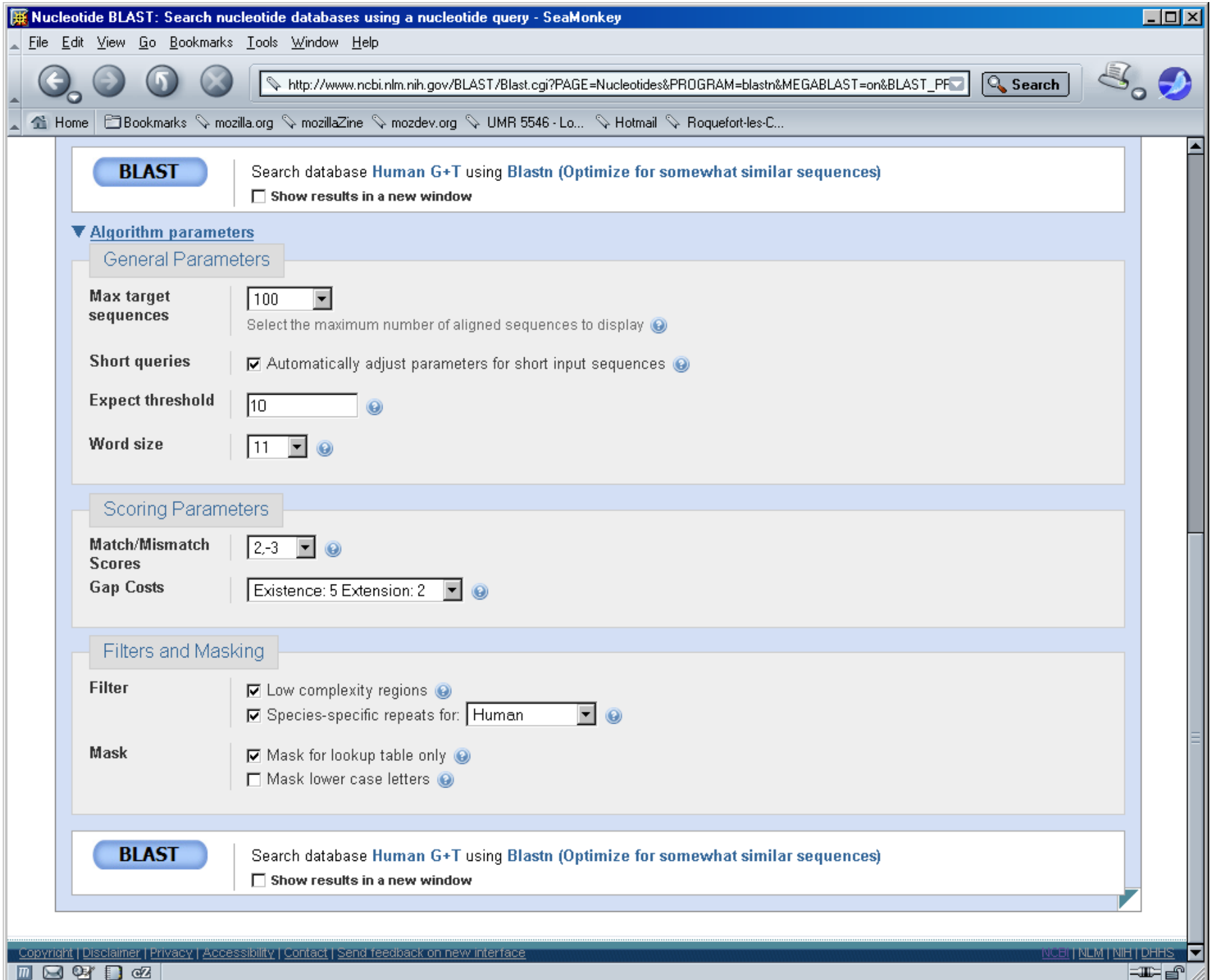
Ce n'est pas une règle, pas de seuils universels !

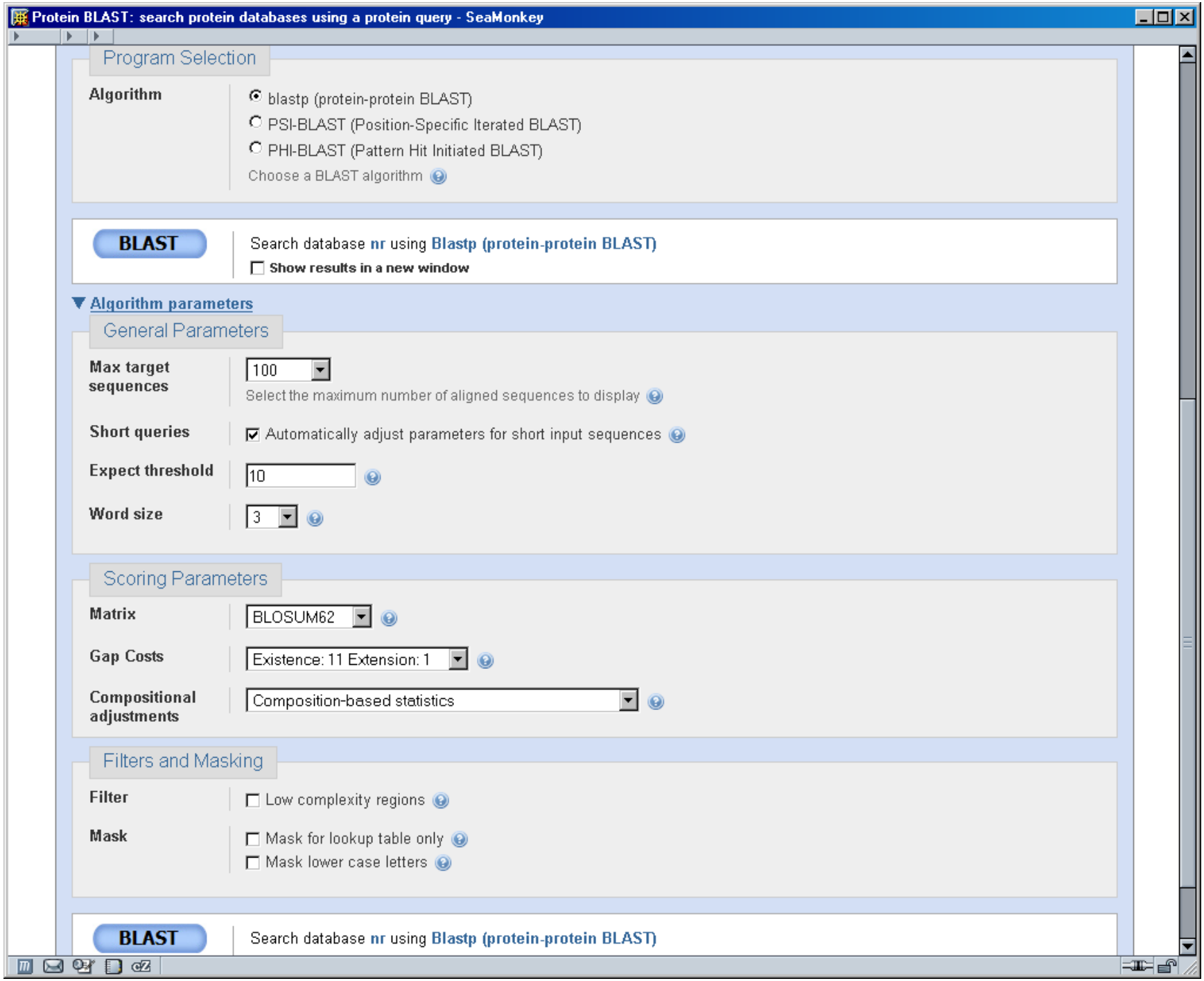
- Dépend de la taille de la banque de données utilisée !
  Valeurs non comparables entre deux banques

- Dépend de la longueur de l'alignement

- Alignement de 2 grandes régions de similarité modérée est + significatif que 2 petites régions à fort taux d'identité


Nucleotide BLAST: Search nucleotide databases using a nucleotide query - SeaMonkey
File Edit View Go Bookmarks Tools Window Help
http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi?PAGE=Nucleotides&PROGRAM=blastn&MEGABLAST=on&BLAST_PF
Search
Home Bookmarks mozilla.org mozillaZine mozdev.org UMR 5546 - Lo... Hotmail Roquefort-les-C...
BLAST
Basic Local Alignment Search Tool
Home Recent Results Saved Strategies Help
My NCBI
[Sign In] [Register]
NCBI/ BLAST/ blastn suite: BLASTN programs search nucleotide databases using a nucleotide query. more...
Reset page Bookmark
Enter Query Sequence
Enter accession number, gi, or FASTA sequence
Clear
Query subrange
From
To
Or, upload file
Browse...
Job Title
Enter a descriptive title for your BLAST search
Choose Search Set
Database
Human genomic + transcript
Mouse genomic + transcript
Others (nr etc.):
Human genomic plus transcript (Human G+T)
Entrez Query
Optional
Enter an Entrez query to limit search
Program Selection
Optimize for
Highly similar sequences (megablast)
More dissimilar sequences (discontiguous megablast)
Somewhat similar sequences (blastn)
Choose a BLAST algorithm
BLAST
Search database Human G+T using Blastn (Optimize for somewhat similar sequences)
Show results in a new window


Nucleotide BLAST: Search nucleotide databases using a nucleotide query - SeaMonkey
File Edit View Go Bookmarks Tools Window Help
http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi?PAGE=Nucleotides&PROGRAM=blastn&MEGABLAST=on&BLAST_PF
Search
Home Bookmarks mozilla.org mozillaZine mozdev.org UMR 5546 - Lo... Hotmail Roquefort-les-C...
BLAST
Search database Human G+T using Blastn (Optimize for somewhat similar sequences)
Show results in a new window
Algorithm parameters
General Parameters
Max target sequences
100
Select the maximum number of aligned sequences to display
Short queries
Automatically adjust parameters for short input sequences
Expect threshold
10
Word size
11
Scoring Parameters
Match/Mismatch Scores
2,-3
Gap Costs
Existence: 5 Extension: 2
Filters and Masking
Filter
Low complexity regions
Species-specific repeats for:
Human
Mask
Mask for lookup table only
Mask lower case letters
BLAST
Search database Human G+T using Blastn (Optimize for somewhat similar sequences)
Show results in a new window
Copyright | Disclaimer | Privacy | Accessibility | Contact | Send feedback on new interface
NCBI | NLM | NIH | DHHS


Protein BLAST: search protein databases using a protein query - SeaMonkey
Program Selection
Algorithm
blastp (protein-protein BLAST)
PSI-BLAST (Position-Specific Iterated BLAST)
PHI-BLAST (Pattern Hit Initiated BLAST)
Choose a BLAST algorithm
BLAST
Search database nr using Blastp (protein-protein BLAST)
Show results in a new window
Algorithm parameters
General Parameters
Max target sequences
100
Select the maximum number of aligned sequences to display
Short queries
Automatically adjust parameters for short input sequences
Expect threshold
10
Word size
3
Scoring Parameters
Matrix
BLOSUM62
Gap Costs
Existence: 11 Extension: 1
Compositional adjustments
Composition-based statistics
Filters and Masking
Filter
Low complexity regions
Mask
Mask for lookup table only
Mask lower case letters
BLAST
Search database nr using Blastp (protein-protein BLAST)

- **MegaBLAST**

Pour séquences nucléiques

Optimisé pour séquences peu différentes (erreurs séquençage)

W = 28 nt

BLAST par défaut quand on va sur le BLAST NCBI

- **Discontiguous MegaBLAST**

Pour séquences divergentes (organismes #)

Cherche des mots discontinus, dans un ensemble

Considère codant/ non-codant

- **CDS-BLAST (Conserved Domain Search)**

Recherche contre des collections de domaines (protéiques) conservés, via des PSSM