Recherche de similarités au moyen de BLAST

Guy Perrière

Laboratoire de Biométrie et Biologie Évolutive
UMR CNRS 5558
Université Claude Bernard – Lyon 1

http://pbil.univ-lyon1.fr/members/perriere/cours/M1/

Objectifs poursuivis

- Comparaison de séquences biologiques :
  - Identification d'homologues.
  - Recherche de contraintes fonctionnelles.
  - Prédiction de structure (ARN, protéine).
  - Prédiction de fonction.
  - Reconstitution des relations évolutives entre séquences (phylogénie).
  - Assemblage de lectures (séquençage).

Homologie ou similarité ?

- La phylogénie moléculaire se base sur l'utilisation de gènes homologues :
  - Deux séquences sont dites homologues si elles possèdent un ancêtre commun.
  - L'existence d'un ancêtre commun est inférée à partir de la similarité.
  - Seuil pour les protéines :
    - 30 % d'identité sur une longueur de 100 AA $\Rightarrow$ homologie entre les séquences.

Similarité sans homologie

- La similarité n'est pas toujours due à de l'homologie :
  - Convergence ou simple hasard pour de courtes séquences (quelques résidus).
  - Existence de régions de faible complexité (*e.g.*, cas de la fibroïne – AAF76983.1) :
    - Présentes dans 40 % des protéines.
    - Peuvent représenter jusqu'à 15 % du total des résidus (Ala, Gly, Pro, Ser, Glu et Gln).

Homologie sans similarité

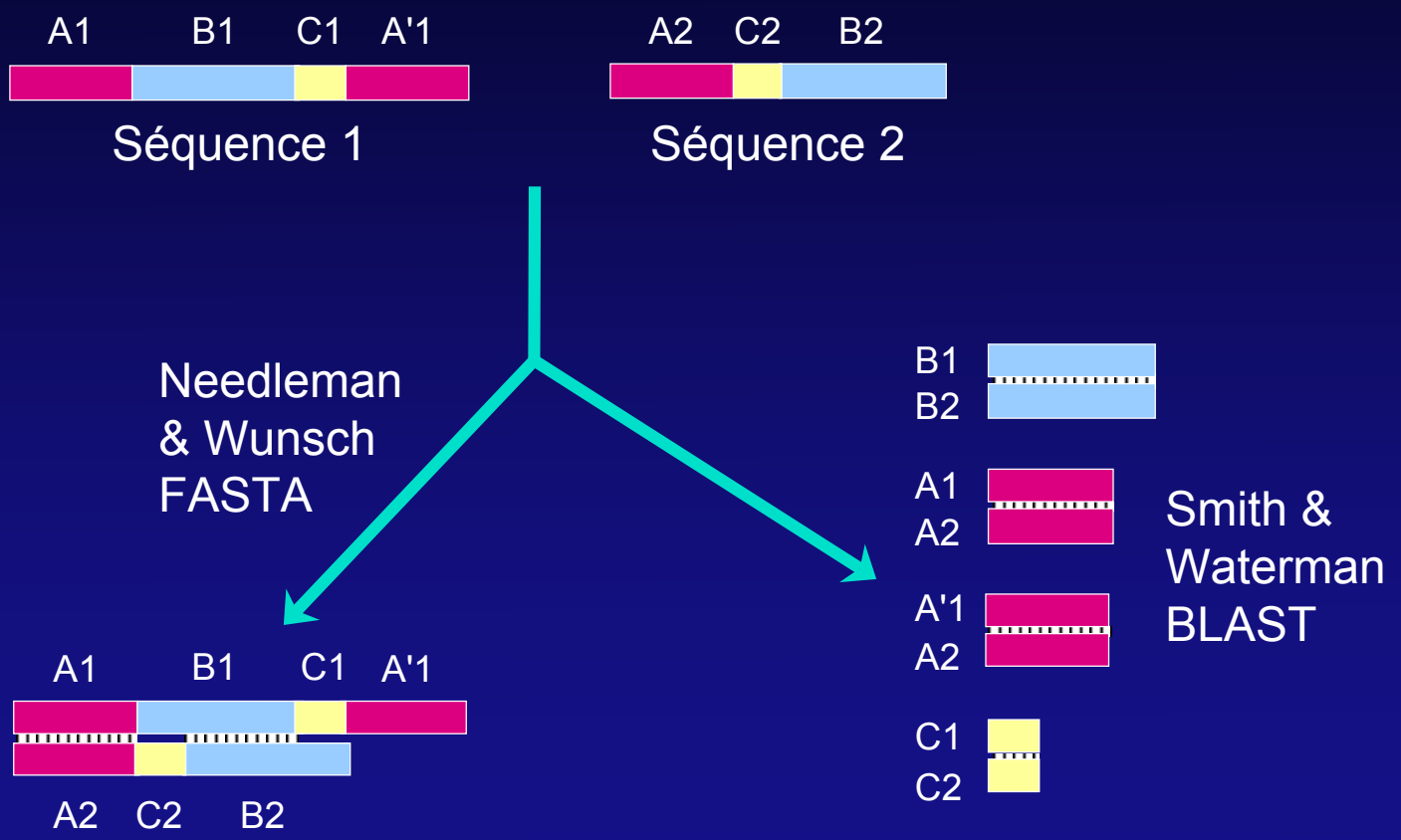
- Deux séquences peuvent être homologues sans que leur similarité soit forte :

```
ACP_KLEAE   ---MEMKIDALAGTLESSDVMVRIGPAAQPGIQLEIDSIVKQEFGAAIQQVVRETLAQLG
ACP_ECOLI   STIEERVKKIIGEQLGVKQEEVTDN--ASFVEDLGADSLDTVELVMALEEEFDTEIPDEE
                 *   :  :  *   :  *     *    :*  **:   *   *:::  :  :::

ACP_KLEAE   VKECDNVQLARVQAAALRWQQ
ACP_ECOLI   AEKITTVQAAIDYINGHQA--
             ::   ** *     : :
```

La similarité entre ces protéines est faible mais les données fonctionnelles et biochimiques montrent qu'elles sont homologues.

Alignement global et local



Représentation

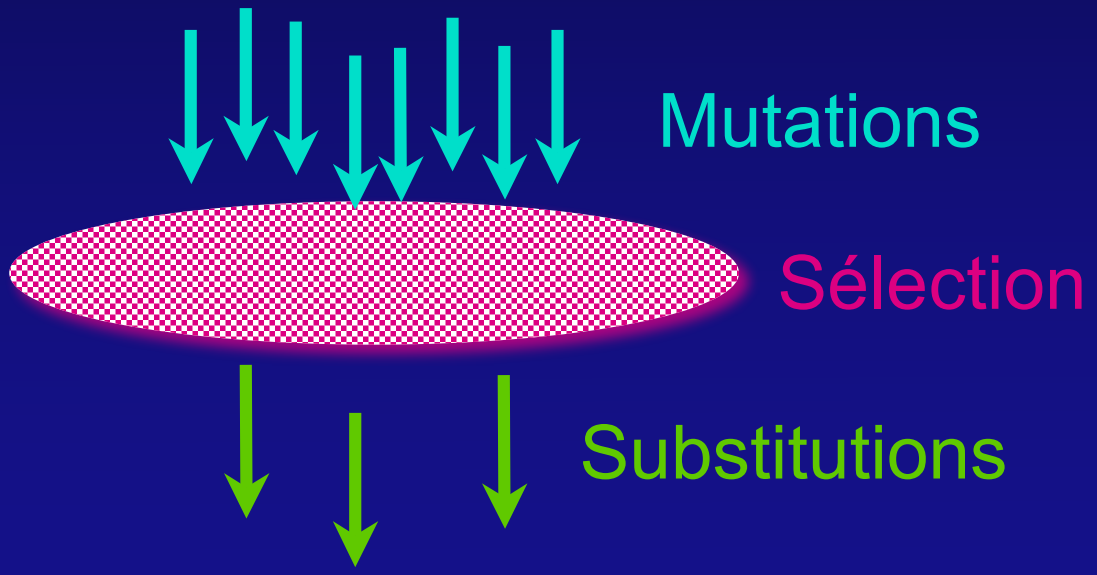
- Les résidus (nucléotides, acides aminés) sont superposés de façon à maximiser les identités entre les séquences :

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

- Il existe deux sortes de différences :
  - Substitutions (*mismatches*).
  - Insertions et délétions (*indels* ou *gaps*).

Mutations et substitutions

- Les différences observées dans un alignement correspondent aux substitutions :
  - Mutations ayant passé le filtre de la sélection :
    - Mutations neutres (*i.e.*, sans effet sur le phénotype) ou avantageuses du point de vue sélectif.



Quel est le bon alignement ?

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

ou

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

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

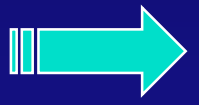
- Pour le biologiste, le bon alignement est celui qui représente le scénario évolutif le plus probable.
- Autres choix possibles (*e.g.*, assemblage de lectures).

Fonction de score de similarité

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

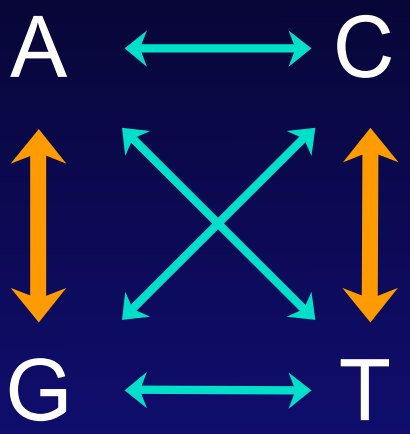
$$\text{Score} = \sum \text{identités} - \sum \text{différences}$$

Identité = +1
Substitution = 0
Gap = –1



Score = 10 – 4 = 6

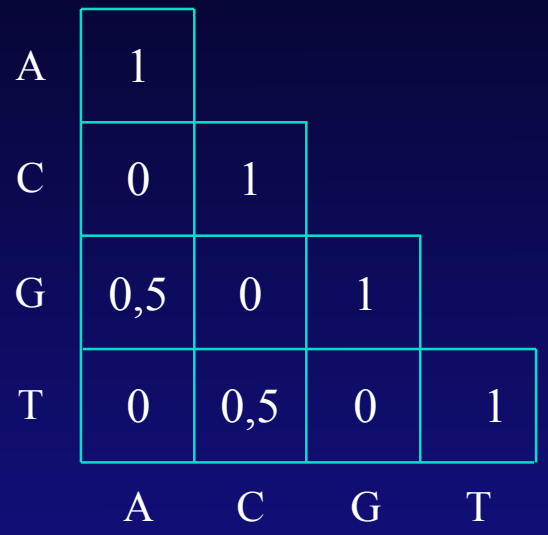
Modèle d'évolution



$$\mathbb{P}(\text{transition}) > \mathbb{P}(\text{transversion})$$

```
G T T A C G A          G T T A C G A
G T T G - G A    >     G T T - G G A
* * * :   * *          * * *     * *
```

Matrice de substitution



$s(A, A) = 1.0$
$s(A, G) = 0.5$
$s(A, –) = –1$

```
G  T  T  A  C  G  A      G  T  T  A  C  G  A
G  T  T  -  G  G  A  <   G  T  T  G  -  G  A
1  1  1 -1  0  1  1      1  1  1 .5 -1  1  1
```

Score = 4

Score = 4,5

Le cas des acides aminés

- Plus difficile à modéliser que celui des séquences nucléotidiques :
  - Un acide aminé peut être remplacé par un autre de différentes façons (code génétique) :
    - La probabilité des substitutions au niveau nucléotidique diffère suivant les codons :

$$\mathbb{P}(\mathrm{AAU}_{\mathrm{Asn}} \rightarrow \mathrm{GAU}_{\mathrm{Asp}}) > \mathbb{P}(\mathrm{AAU}_{\mathrm{Asn}} \rightarrow \mathrm{CAU}_{\mathrm{His}})$$

  - Certaines substitutions peuvent avoir plus ou moins d'effet sur la fonction des protéines.
  - Utilisation de nombreux critères :
    - Code génétique, propriétés physico-chimiques, hypothèses sur les processus évolutifs.

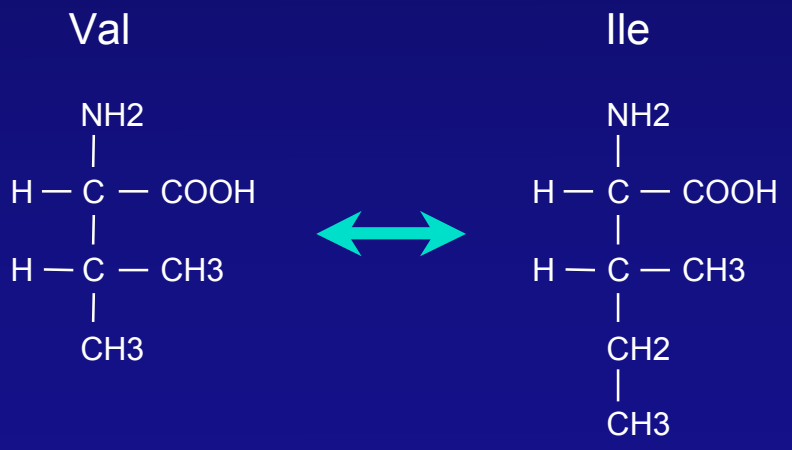
Premiers modèles utilisés

- Code génétique (Fitch, 1966) :

```
Asp (GAC, GAU) → Tyr (UAC, UAU)   1 mutation
Asp (GAC, GAU) → Cys (UGC, UGU)   2 mutations
Asp (GAC, GAU) → Trp (UGG)        3 mutations
```

- Propriétés physico-chimiques des acides aminés :
  - Acidité, polarité, hydrophobicité, etc.

Substitutions conservatrices



Modèles empiriques

- Matrices fondées sur des alignements de domaines conservés :
  - BLOSUM (Henikoff et Henikoff, 1992).
- Matrices fondées sur des arbres construits par maximum de parcimonie :
  - PAM (Dayhoff *et al*., 1978).
  - JTT (Jones *et al*., 1992).
- Matrices fondées sur des arbres construits par maximum de vraisemblance :
  - WAG (Whelan et Goldman, 2001).
  - LG (Le et Gascuel, 2008)

Matrices BLOSUM

- BLOSUM (*Blocks Substitution Matrices*) :
  - Utilisation de ~2000 domaines conservés provenant de ~500 familles de protéines.
  - Alignement de ces domaines :
    - Utilisation de la matrice identité pour effectuer l'alignement.
    - Alignements sans *gaps*.
  - Ensemble de matrices créées à partir de domaines comprenant des séquences ± divergentes :
    - Toutes les paires ayant servi à construire une matrice BLOSUM*k* ont une identité ≥ à *k* %.
    - Matrices bien adaptées pour des protéines distantes du point de vue évolutif.

Matrice BLOSUM62

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	9																			
S	-1	4																		
T	-1	1	5																	
P	-3	-1	-1	7																
A	0	1	0	-1	4															
G	-3	0	-2	-2	0	6														
N	-3	1	0	-2	-2	0	6													
D	-3	0	-1	-1	-2	-1	1	6												
E	-4	0	-1	-1	-1	-2	0	2	5											
Q	-3	0	-1	-1	-1	-2	0	0	2	5										
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8									
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5								
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5							
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5						
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4					
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4				
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4			
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6		
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7	
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11

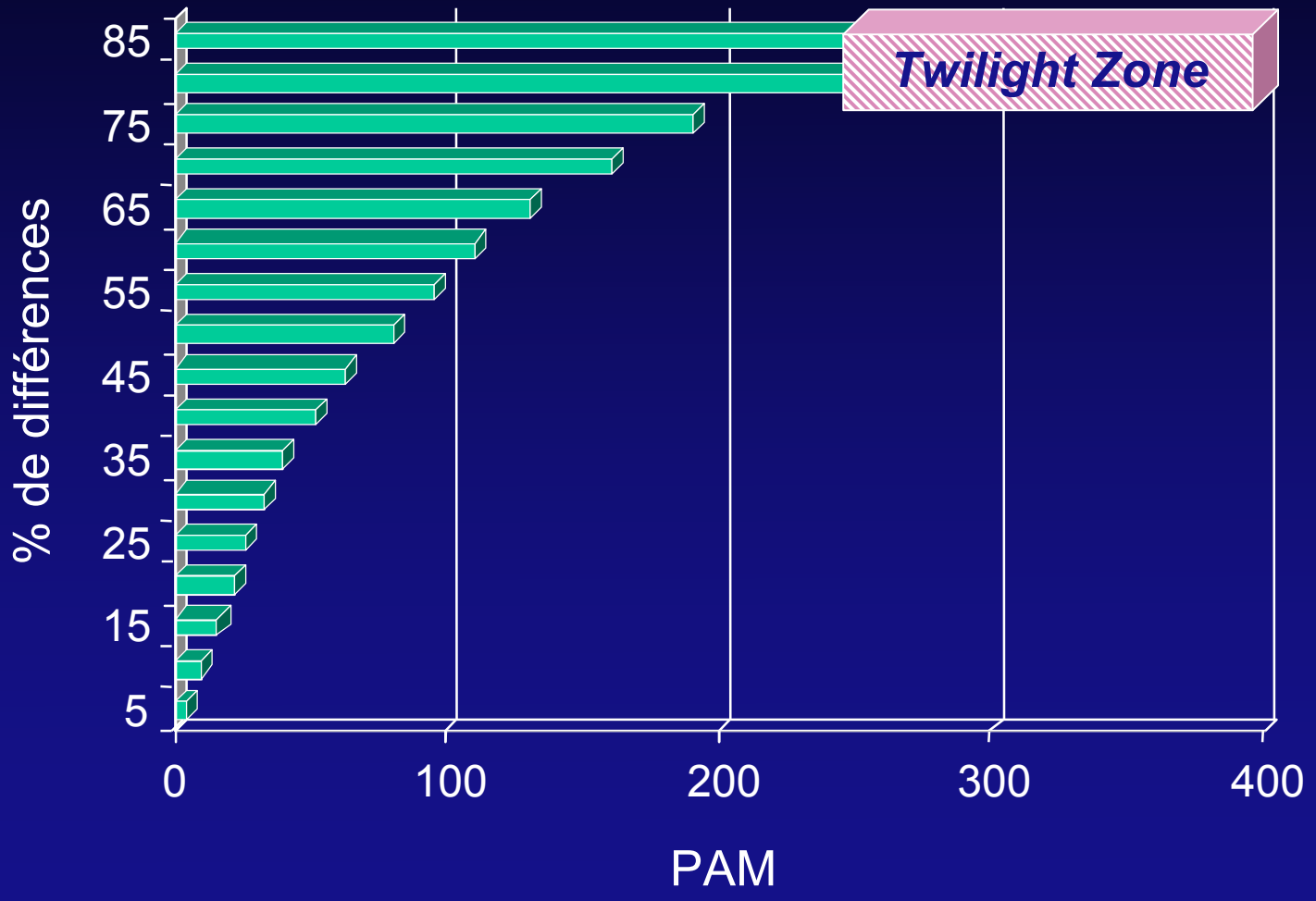
C	sulfhydryle
S	
T	
P	petit hydrophile
A	
G	
N	
D	acide, acide-amide
E	et hydrophile
Q	
H	
R	basique
K	
M	
I	petit hydrophobe
L	
V	
F	
Y	aromatique
W	

Les critères évolutifs recoupent les critères physico-chimiques

Matrices PAM et JTT

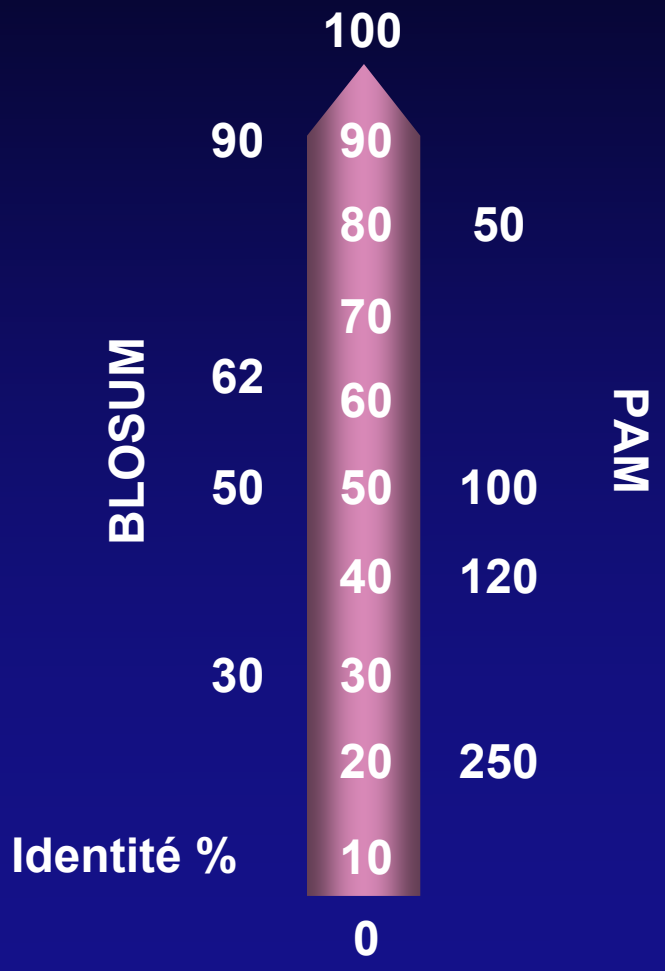
- Alignements globaux :
  - Utilisation de la matrice identité.
  - Pas de prise en compte des *gaps* dans les alignements.
- PAM (*Point Accepted Mutation*) :
  - 71 familles de gènes correspondant à 1300 séquences pour un total de 1572 substitutions.
- JTT (*Jones, Taylor and Thornton*) :
  - 16 300 protéines pour un total de 59 190 substitutions.

Seuil pour les matrices PAM

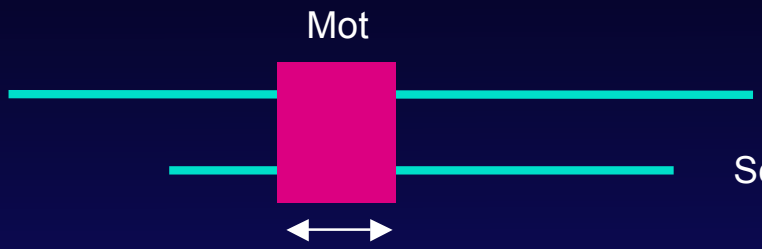


Choix d'une matrice

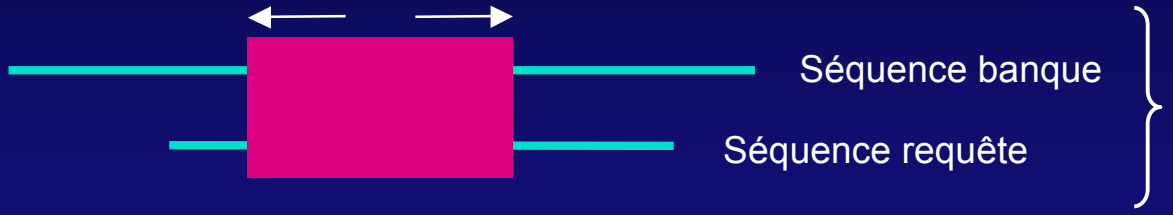
- Pas de matrice idéale.
- Meilleurs résultats avec les matrices construites avec un plus grand nombre de séquences.
- Structure des protéines (globulaires ou membranaires).
- Degré de similarité des séquences.
- Il est recommandé d'expérimenter !



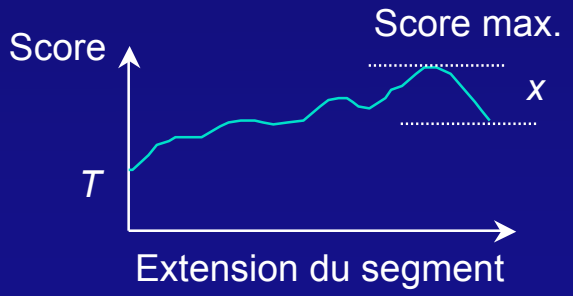
BLAST : principe général



Longueur du mot = *w*
Score ≥ *T*

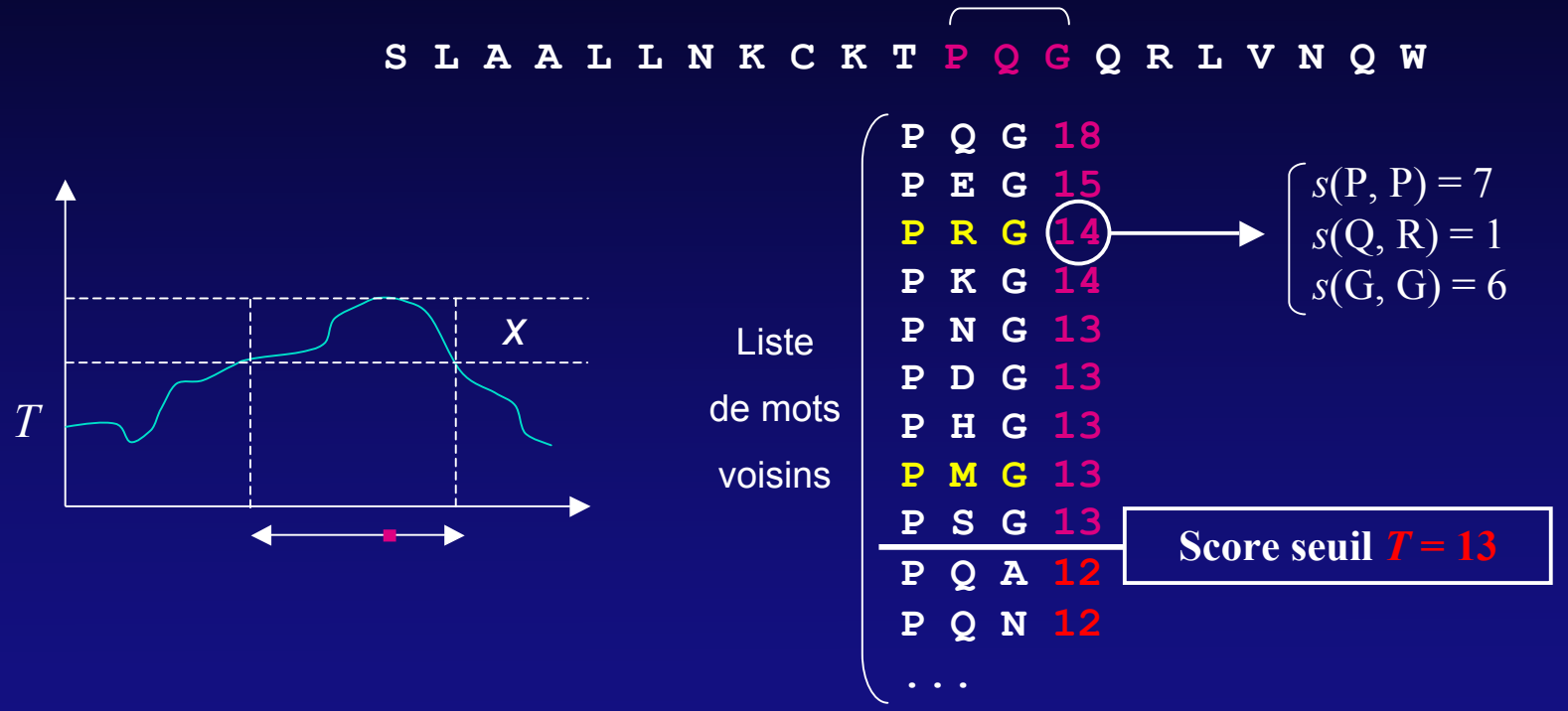


HSP : *High Scoring Pair*



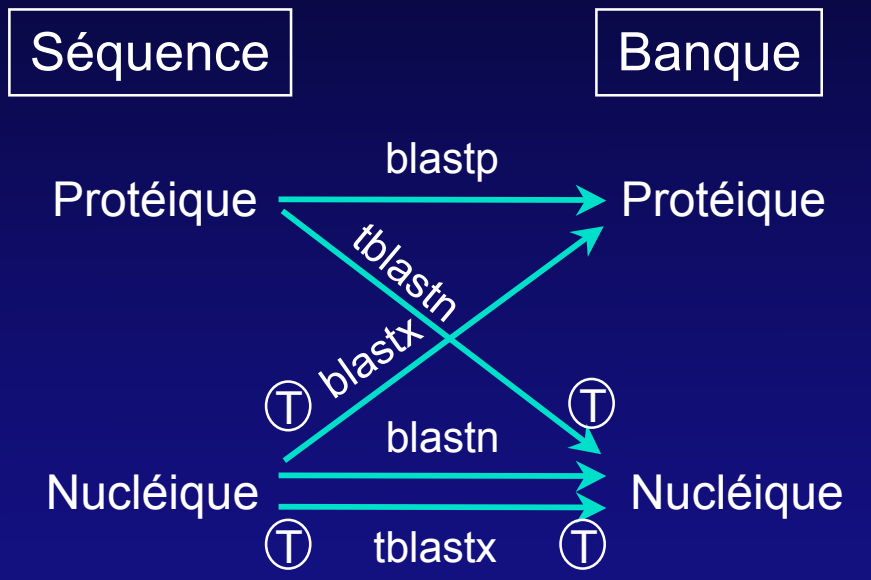
Extension stoppée quand :
- la fin d'une des deux séquences est atteinte
- score ≤ 0
- score ≤ score_max - *x*

Exemple


S L A A L L N K C K T P Q G Q R L V N Q W
T
X
Liste
de mots
voisins
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
Query : 325 S L A A L L N K C K T P Q G Q R L V N Q W 345
+ L A + + L + T P G R + + + W
Sbjct : 290 T L A S V L D C T V T P M G S R M L K R W 310

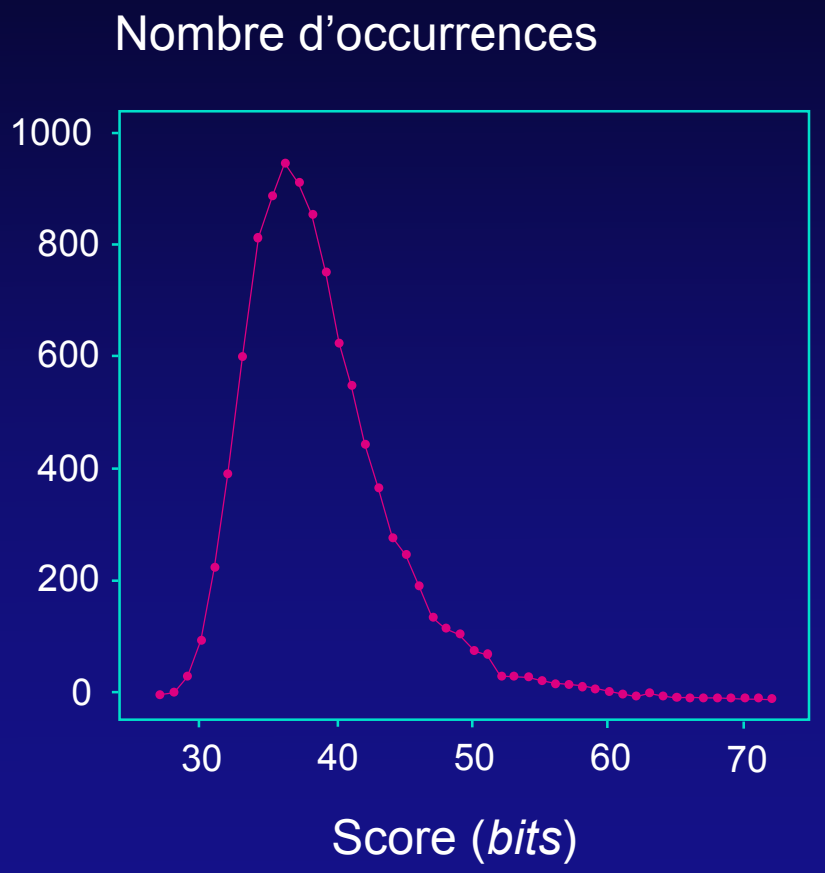
Versions de BLAST

- blastp : protéine *vs.* protéine.
- blastn : utile pour le non-codant.
- blastx : séquences co-dantes non identifiées.
- tblastn : homologues dans un génome non complètement annoté.



Évaluation statistique

- Similarités détectées :
  - Relations significatives.
  - Similarités dues au hasard.
- Fonction de score :
  - Mesure sous la forme :
    - D'une espérance mathématique (*E-value*).
    - Valeur en *bits*.
  - Basée sur une distribution calculée à partir séquences non homologues.
  - Les scores dépendent de la taille de la banque.



E-value, *bits* et similarité

- Soit E, l'espérance mathématique d'avoir une similarité $\geq$ au score S observé :

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

  Avec m et n les tailles *effectives* de la séquence requête et de la banque, et K et λ deux paramètres dérivés de la distribution précédente.

- Le score en *bits* S' est donné par :

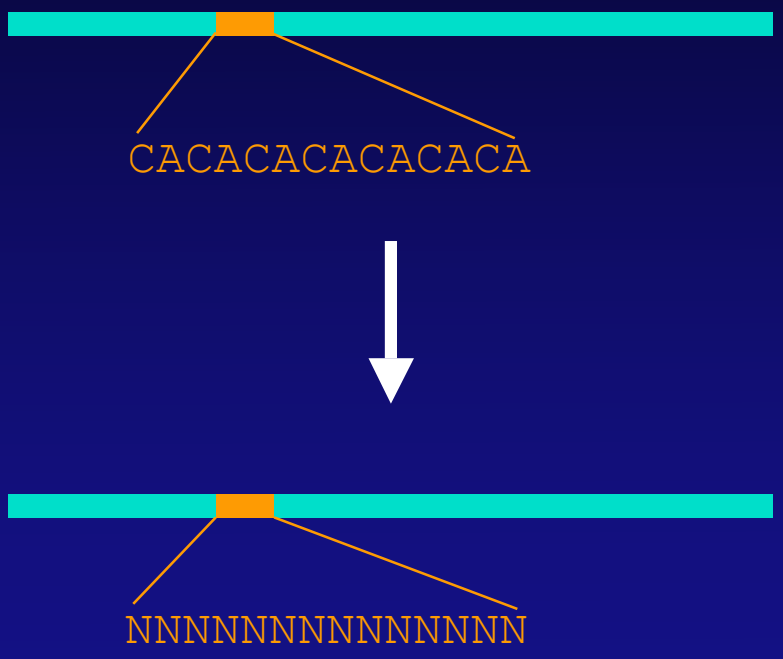
$$S' = [\lambda S - \ln(K)] / \ln(2)$$

- La relation entre E et S' est donc donnée par :

$$E = mn\, 2^{-S'}$$

Séquences abondantes

- Immunoglobulines :
  - > 42 000 séquences dans GenBank.
- Séquences répétées :
  - 10^6 Alu et 10^5 L1 dans le génome humain.
- Programmes de masquage :
  - DUST, SEG, Repeat Masker.



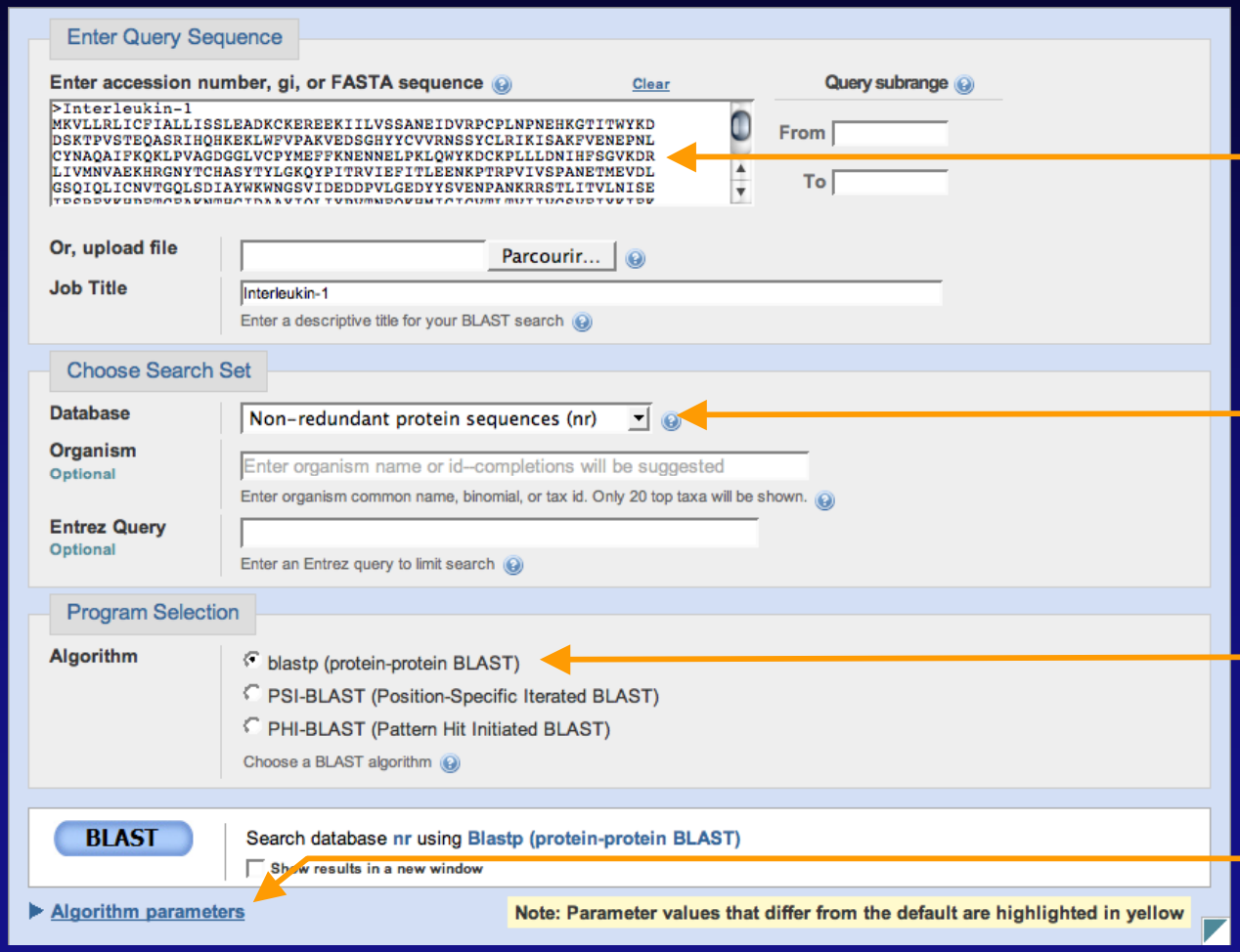
Serveurs BLAST

- Il existe un grand nombre de serveurs permettant d'effectuer des recherches BLAST mais…
  - Toutes les options ne sont pas toujours accessibles.
  - Peu sont exhaustifs du point de vue des banques.
  - Tous ne permettent pas d'accéder à des banques mises à jour quotidiennement.
  - Les possibilités de filtrage pré- ou post-recherche sont rares et limitées.
  - Généralement pas de lien directs avec d'autres applications (*e.g.*, alignements multiples).

BLAST au NCBI

- Répond à (presque) toutes les questions précédentes.
- Est particulièrement rapide.
- Bénéficie d'une interface graphique de visualisation des résultats.
- Mais…
  - Est de ce fait très sollicité !
  - Toujours pas de tuyauterie vers l'étape suivante (construction d'un alignement multiple).

BLAST au NCBI (1)



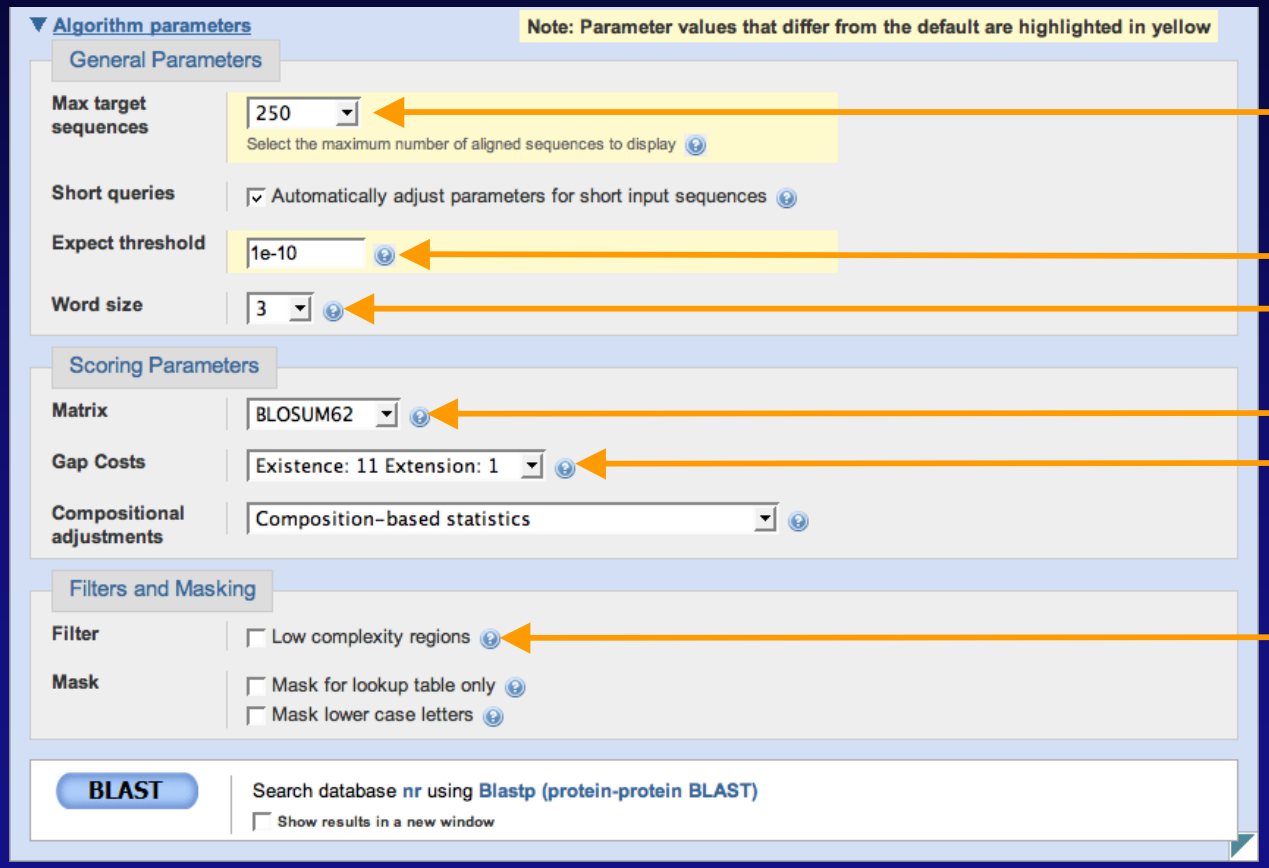
Séquence requête

Choix de la base de données

Choix de l'algorithme

Paramètres

BLAST au NCBI (2)



Nb. maximum de séquences à visualiser

E-value limite

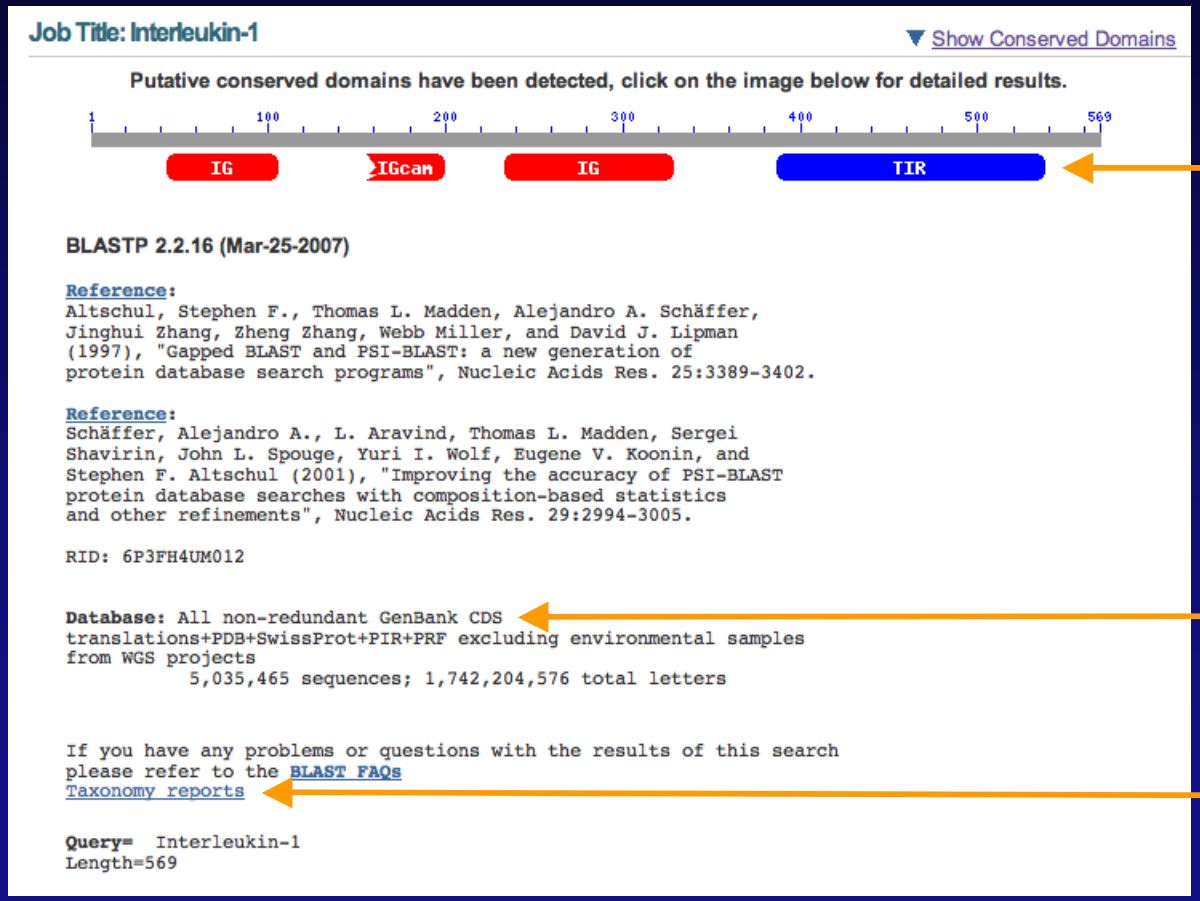
Taille du mot

Choix de la matrice

Pénalités des *gaps*

Filtrage (séquences de faible complexité)

BLAST au NCBI (3)

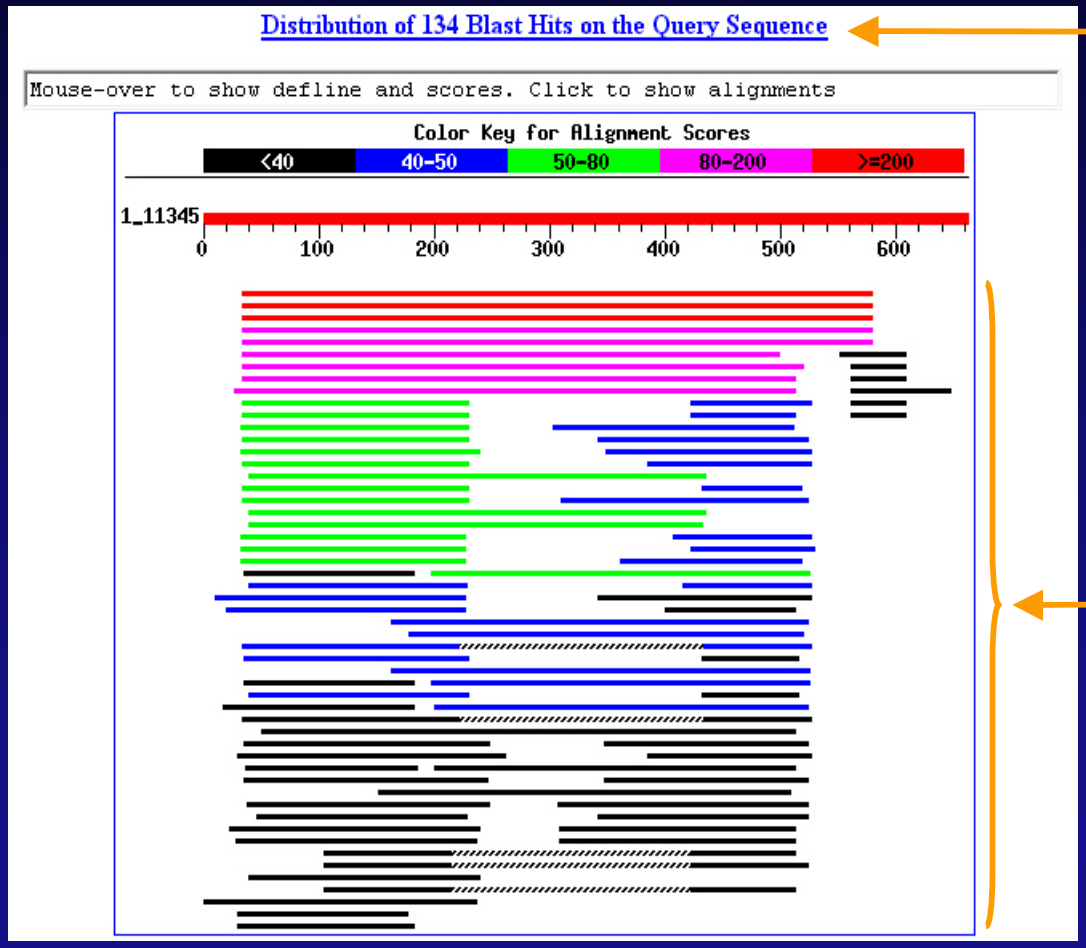


Domaines conservés

Banque de données

Résultats par taxon

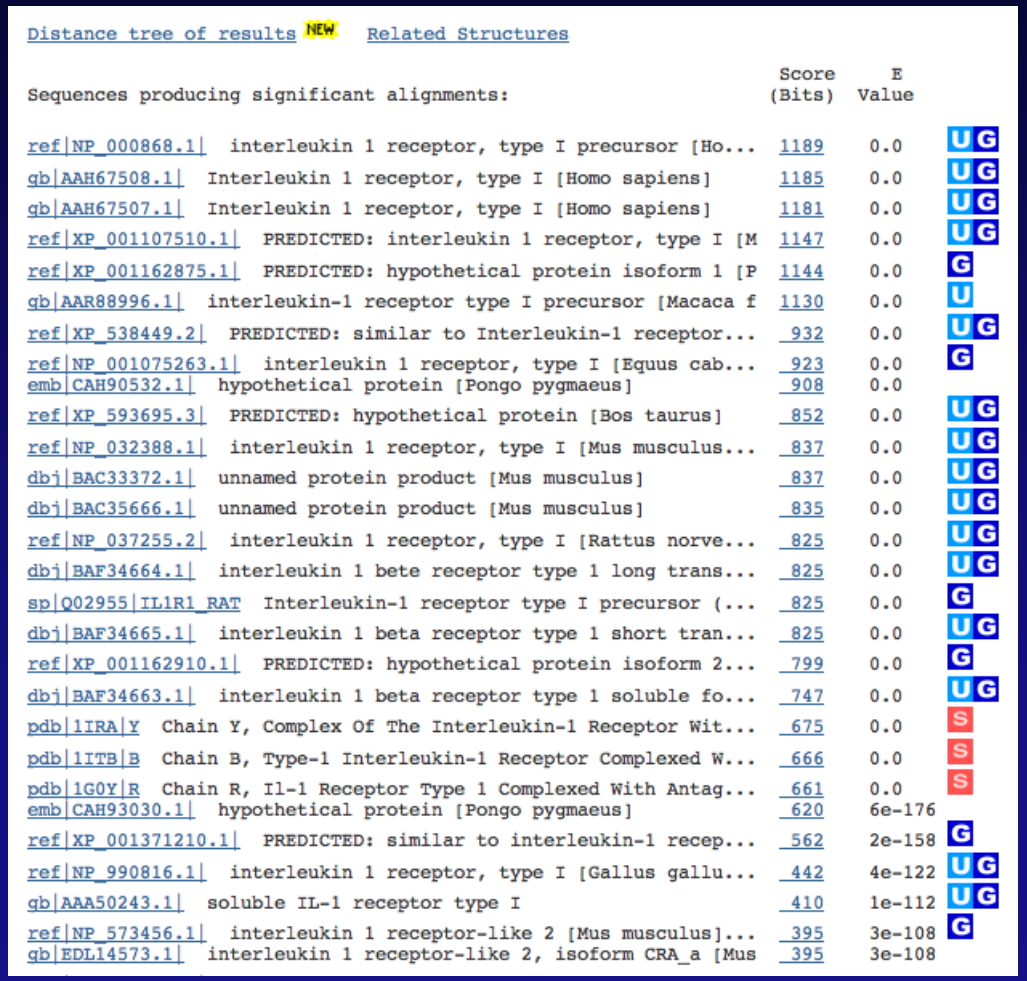
BLAST au NCBI (4)



Nombres de *hits*

Répartition des *hits* en fonction du score

BLAST au NCBI (5)

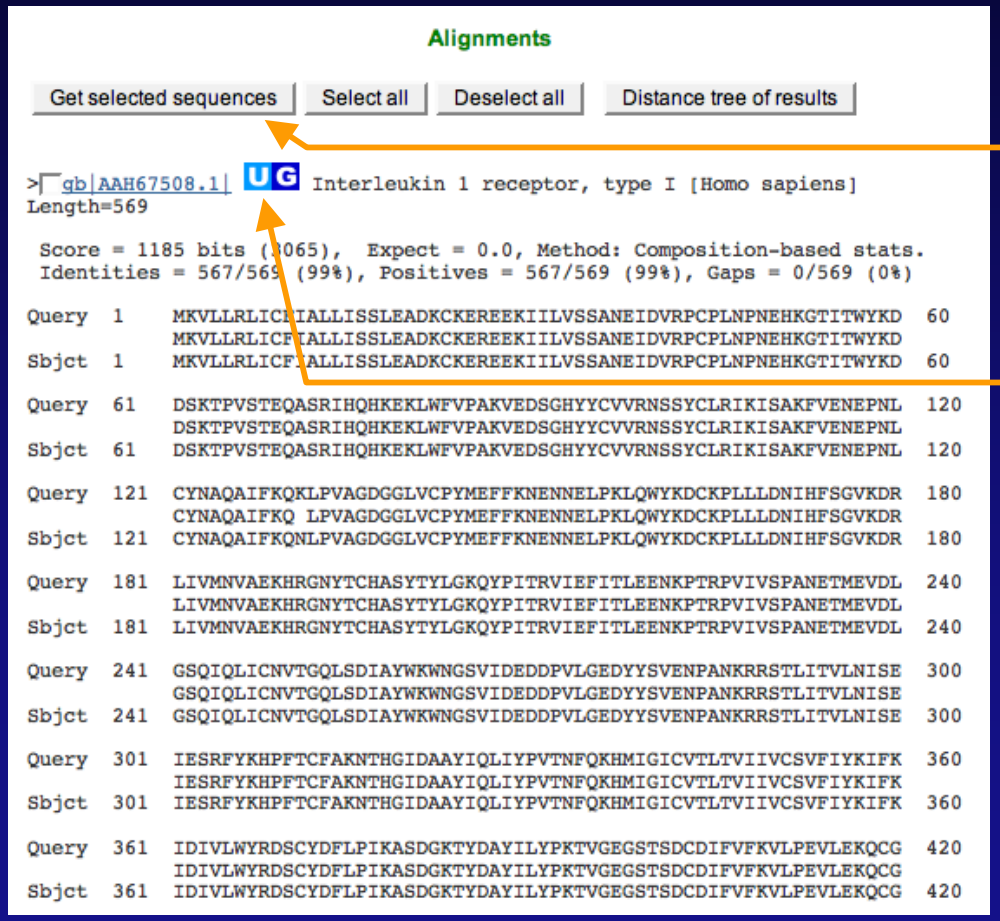


Distance tree of results NEW Related Structures

Sequences producing significant alignments:	Score (Bits)	E Value	
ref|NP_000868.1| interleukin 1 receptor, type I precursor [Ho...	1189	0.0	UG
qb|AAH67508.1| Interleukin 1 receptor, type I [Homo sapiens]	1185	0.0	UG
qb|AAH67507.1| Interleukin 1 receptor, type I [Homo sapiens]	1181	0.0	UG
ref|XP_001107510.1| PREDICTED: interleukin 1 receptor, type I [M	1147	0.0	UG
ref|XP_001162875.1| PREDICTED: hypothetical protein isoform 1 [P	1144	0.0	G
qb|AAR88996.1| interleukin-1 receptor type I precursor [Macaca f	1130	0.0	U
ref|XP_538449.2| PREDICTED: similar to Interleukin-1 receptor...	932	0.0	UG
ref|NP_001075263.1| interleukin 1 receptor, type I [Equus cab...	923	0.0	G
emb|CAH90532.1| hypothetical protein [Pongo pygmaeus]	908	0.0	
ref|XP_593695.3| PREDICTED: hypothetical protein [Bos taurus]	852	0.0	UG
ref|NP_032388.1| interleukin 1 receptor, type I [Mus musculus...	837	0.0	UG
dbj|BAC33372.1| unnamed protein product [Mus musculus]	837	0.0	UG
dbj|BAC35666.1| unnamed protein product [Mus musculus]	835	0.0	UG
ref|NP_037255.2| interleukin 1 receptor, type I [Rattus norve...	825	0.0	UG
dbj|BAF34664.1| interleukin 1 bete receptor type 1 long trans...	825	0.0	UG
sp|Q02955|IL1R1_RAT Interleukin-1 receptor type I precursor (...	825	0.0	G
dbj|BAF34665.1| interleukin 1 beta receptor type 1 short tran...	825	0.0	UG
ref|XP_001162910.1| PREDICTED: hypothetical protein isoform 2...	799	0.0	G
dbj|BAF34663.1| interleukin 1 beta receptor type 1 soluble fo...	747	0.0	UG
pdb|1IRA|Y Chain Y, Complex Of The Interleukin-1 Receptor Wit...	675	0.0	S
pdb|1ITB|B Chain B, Type-1 Interleukin-1 Receptor Complexed W...	666	0.0	S
pdb|1G0Y|R Chain R, Il-1 Receptor Type 1 Complexed With Antag...	661	0.0	S
emb|CAH93030.1| hypothetical protein [Pongo pygmaeus]	620	6e-176	
ref|XP_001371210.1| PREDICTED: similar to interleukin-1 recep...	562	2e-158	G
ref|NP_990816.1| interleukin 1 receptor, type I [Gallus gallu...	442	4e-122	UG
qb|AAA50243.1| soluble IL-1 receptor type I	410	1e-112	UG
ref|NP_573456.1| interleukin 1 receptor-like 2 [Mus musculus]...	395	3e-108	G
qb|EDL14573.1| interleukin 1 receptor-like 2, isoform CRA_a [Mus	395	3e-108	

Séquences similaires

BLAST au NCBI (6)



Récupération des séquences

Liens vers des banques NCBI

Quelle approche adopter ?

- Choix d'un algorithme.
- Choix de la matrice de substitution.
- Pondération des *gaps*.
- Stratégie de recherche (nucléique ou protéique).
- Traitement du bruit de fond.
- Banque sur laquelle effectuer la recherche.
- Répétition de la recherche.

Du bon usage de BLAST

- L'annotation par similarité peut conduire à certains abus…

```
MZEORFG     ILNSPDRACNLAKQAFDEAISELDSLGEESYKDSTLIMQLLXDNLTLWTSDTNEDGGDE
BOV1433P    IQNAPEQACLLAKQAFDDAIAELDTLNEDSYKDSTLIMQLLRDNLTLWTSDQQDEEAGE
            * * *:.** ********:** ***.* *:************ ********* .::   *

Score = 87.4 bits (213), Expect = 1e-17
Identities = 41/59 (69%), Positives = 50/59 (84%)

LOCUS       BOV1433P     1696 bp    mRNA            MAM       26-APR-1993
DEFINITION  Bovine brain-specific 14-3-3 protein eta chain mRNA, complete
            cds.

LOCUS       MZEORFG       187 bp    mRNA            PLN       31-MAY-1994
DEFINITION  Zea mays putative brain specific 14-3-3 protein, tau protein
            homolog  mRNA, partial cds.
```

Exercice d'application

- Récupérer la séquence P04118 au format Fasta sur le serveur du NCBI.
- Déterminer quels sont les homologues de cette séquence au moyen de BLAST :
  - Utilisation de la banque RefSeq.
  - Paramètre *Max target sequences* à 1000.
- Sachant que la colipase est d'origine pancréatique, que pensez-vous des résultats obtenus.