Analyse de séquences

Partie II-2 : Matrices de scores, BLAST, FASTA

Sèverine Bérard

AMAP - Université Montpellier 2

Similarité et homologie

- Tendance à les utiliser de manière interchangeable alors qu'ils ont des sens différents et impliquent des relations biologiques différentes
- Similarité
  - → Mesure quantitative de comment 2 séquences sont reliées
  - → La similarité est toujours basée sur de l'observable, en général sur l'alignement de ces deux séquences
  - → Ex : pourcentage d'identité, score d'alignement

⇒ Haut degrés de similarité *peut impliquer* une origine évolutive commune

- Homologie
  - → Conclusion supposée basée sur l'examen d'un alignement optimal entre 2 séquences et de leur similarité
  - → Les gènes (ou les protéines corr.) sont ou ne sont pas homologues, l'homologie ne se mesure pas en degrés
  - → Le concept d'homologie implique des relations évolutives et peut s'employer pour 2 types de rel. : orthologie et paralogie

⇒ Voir la partie phylogénie

Alignement global contre alignement local

- Ce sont les deux grands types d'alignements

- Ils permettent d'estimer la similarité entre séquences

→ GLOBAL : 2 séquences, comparaison sur leur totalité, trouve le meilleur alignement sur toute leur longueur
Utilisé pour des séq. très similaires et approx. de même lg
[Needleman & Wunsh, 1970]

→ LOCAL : chercher les régions les plus similaires dans les 2 séq.
Permet de trouver des sous-séq. ayant des relations biologiques
[Smith & Waterman, 1981]

Les alignement locaux sont meilleurs pour des séquences avec un faible degrés de similarité ou de tailles différentes

- Donner un score aux alignements
  - → Les matrices de scores, généralités
  - → Les matrices PAM
  - → Les matrices BLOSUM
  - → Les matrices nucléiques
  - → Pénalités de gaps
- BLAST
- FASTA
- Comparaison BLAST/FASTA
- La suite ...

Les matrices de scores

- Complément naturel des méthodes “numériques” produisant un score (comme les alignements)
- Shémas de score empirique
- Utilisées dans la comparaison de deux ou plusieurs séquences
- Comment sont-elles construites ? Comment les choisir ?

Choix important influençant fortement les résultats

[Henikoff & Henikoff, 2000]

Facteurs biologiques importants

- Conservation

  Conservation absolue et substitutions conservatives

→ *Quels résidus peuvent être substitués par d'autres sans affecter la fonction de la protéine ?*

- Fréquence

  Les résidus rares ont un poids plus fort que les résidus communs

→ *Les matrices doivent refléter le nombre de fois qu'un résidu apparaît dans l'ensemble des protéines*

- Évolution

  Les matrices représentent implicitement les schémas évolutifs

→ *Le choix des matrices à utiliser dépend de la distance évolutive*

Exemple : BLOSUM62

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

Exemple : BLOSUM62

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

Exemple : BLOSUM62

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

Exemple : BLOSUM62

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

Exemple : BLOSUM62 (2)

- Score = logarithme d'un "rapport de chances" (*odds ratio*) : prend en compte le nb de fois où un résidu particulier

→ est remplacé par un autre dans la nature

→ serait remplacé par un autre si ça se produisait par chance

Les substitutions observées fréquemment ont un score positif et celles peu fréquentes, un score négatif

- Log odds ratio : $S_{i,j} = \log[\frac{q_{i,j}}{p_i p_j}]$

  p_i prob. d'occurence de i parmis toutes les protéines

  $q_{i,j}$ fréquence de substitution i par j (à partir d'alignements)

- Le log odds ratio est le rapport de l'observé sur l'aléatoire

Les matrices PAM

- Les premières matrices utiles pour l'analyse de séquences [Dayoff et al., 1978]

- Basées sur l'examen des motifs de substitution d'un groupe de protéines partageant plus de 85% d'identité
  $\Rightarrow$ Table contenant la fréquence de substitution d'un a.a. par un autre à une position donnée

- L'unité de mesure résultant de cette analyse est le **P***oint* **A***ccepted* **M***utation* ou unité **PAM**

- Une unité PAM correspond à 1 changement d'a.a. pour 100 résidus

Les matrices PAM : les hypothèses

- Le remplacement d'un a.a. est indépendant des mutations précédentes à la même position
  $\Rightarrow$ la matrice originale a été extrapolée pour prédire les fréquences de substitution pour des distances évolutives plus élevées

  Ex : PAM1 multipliée par elle-même 100 fois donnerait PAM100

- Tous les sites sont mutables de manière équivalente

- Les remplacements sont indépendants des résidus voisins

- Pas de considération de bloc ou de motif

- Autres biais dûs par exemple aux nb et caractéristiques des protéines connues en 1978, ...

Les matrices BLOSUM

- Approche différente par [Henikoff & Henikoff, 1992]

- Identifier les motifs conservés à l'intérieur des familles de protéines
  $\Rightarrow$ Création de la base de données BLOCKS

$\rightarrow$ Motif : suite conservée d'a.a. qui confère une fonction ou une structure spécifique à une protéine

$\rightarrow$ Blocs : généralisation des motifs

- À l'aide de ces blocs, ils ont regardé les schémas de substitution dans les régions les plus conservées des protéines

  $\Rightarrow$ Génération des **BL***ocks* **SU***bstitution* **M***atrices* ou matrices **BLOSUM**

Les matrices BLOSUM (2)

- Plus de protéines disponibles en 1992 qu'en 1978, donc un jeu de données plus robuste pour dériver des matrices

- Les matrices BLOSUM sont directement calculées à différentes échelles évolutives, elles ne sont pas extrapolées

- BLOSUMn : n représente le niveau de conservation des séquences utilisées pour construire la matrice

  Ex : BLOSUM62 est calculée à partir de séq. qui ne partagent pas plus de 62 % d'id., les autres sont regroupées et comptent pour un

- Réduire la valeur de n permet d'attraper des séquences plus divergentes

Quelle matrice ? Quand ?

- Le choix par défaut des logiciels n'est pas forcément approprié

Matrices	Meilleur usage	Similarité (%)
PAM40	Alignements courts qui sont très similaires	70-90
PAM160	Détecter les membres d'une famille de protéines	50-60
PAM250	Ali. plus longs de séq. plus divergentes	~ 30
BLOSUM90	Alignements courts qui sont très similaires	70-90
BLOSUM80	Détecter les membres d'une famille de protéines	50-60
BLOSUM62	La matrice "tout-terrain"	30-40
BLOSUM30	Ali. plus longs de séq. plus divergentes	~ 30

- Rmq : Les numéros des 2 familles de matrices vont en sens opposé
- PAM250$\sim$BLOSUM45 ; PAM160$\sim$BLOSUM62 ; PAM120$\sim$BLOSUM80
- Il existe des matrices spécialisées : espèces spécifiques, classes de protéines (transmembranaire par ex), ...

- 4 nucléotides au lieu de 20 a.a.
- Hypothèse simple de fréquence identique (25% chacun), avec éventuellement un poids différents pour les transitions (plus fréquentes que les transversions)

	A	C	G	T
A	1			
C	0	1		
G	0	0	1	
T	0	0	0	1

	A	C	G	T
A	3			
C	0	3		
G	1	0	3	
T	0	1	0	3

- Plus d'études sur les séquences protéiques que sur les séquences nucléiques $\Rightarrow$ matrices nucléiques moins perfectionnées

$\Rightarrow$ Chercher sur les séq. protéiques est toujours plus efficace

Gaps et pénalités de gap

- Les gaps servent à améliorer les alignements

- Mais pour ne pas donner lieu à des scénarios biologiquement invraisemblables, il faut les garder en nombre raisonnable :
  $\rightarrow$ 1 indel pour 20 a.a.

- La méthode la plus utilisée pour pénaliser les gaps est la "pénalité de gap affine"

  $o + e \times (l - 1)$ où $o > e$

  o pénalité d'ouverture de gap
  e pénalité d'extension de gap
  l longueur du gap

- Autres types de pénalités possibles : linéaires, logarithmiques, ...

- Donner un score aux alignements
- BLAST
  - → L'algorithme
  - → Un exemple d'utilisation
- FASTA
- Comparaison BLAST/FASTA
- La suite ...

BLAST [Altschul *et al.*, 1990]

Basic Local Alignment Search Tool

- But : trouver un score d'alignement local élevé entre une séquence requête et une base de donnée cible
- Capable de détecter précisément et rapidement des similarités entre des séquences nucléiques ou protéiques, sans sacrifier la sensibilité
- Idée : les séquences similaires ont des segments communs de taille k quasiment identiques, ces petits segments vont servir de graînes (ADN $k = 11$, protéines $k = 3$)

Les différentes versions de BLAST

- BLASTN : séquences nucléiques
- BLASTP : séquences protéiques
- BLASTX : une séquence nucléique comparée à une BD protéique (traduction suivant les 6 cadres de lecture)
- TBLASTX : une séquence nucléique comparée à une BD nucléique, chacune suivant tous les cadres de lecture (~ 36 BLASTP)
- Psi-BLAST : itération de BLAST (*Position-Specific-Iterated Blast*)
  $\Rightarrow$ augmente la sensibilité
- Phi-BLAST : motifs (*Pattern-Hit Initiated BLAST*)
  $\Rightarrow$ augmente la sélectivité
- MEGABLAST : pour séq. longues ou très similaires (> 95 %)
- BLAST2SEQUENCES : trouve les alignement locaux entre 2 séq.
- BLAT : optimisé pour les génomes

- On découpe S la séquence requête, en mots de longueur fixée (petite) : les mots requêtes (*query words*)

  ex S=TLSHAWRLSNETDKRPFIETALRDQHKKDYPEYKYQPRRRKNGKP

  query words : TLS, LSH, SHA, ...

- Pour chaque mot requête, on détermine son voisinage

  ex pour RDQ

RDQ 16	QDQ 12	EDQ 11	RDN 11	RDB 11	BDQ 10	RDP 10
RBQ 14	REQ 12	HDQ 11	RDD 11	ADQ 10	XDQ 10	RDT 10
RDZ 14	RDR 12	ZDQ 11	RDH 11	MDQ 10	RQQ 10	RDY 10
KDQ 13	RDQ 12	RNQ 11	RDM 11	SDQ 10	RSQ 10	RDX 10
RDE 13	NDQ 11	RZQ 11	RDS 11	TDQ 10	RDA 10	DDQ 9 ...

- Score seuil T, par exemple $T = 11$, seuls les mots voisins dont le score d'ali. avec le *query word* est $\geq T$ passent à l'étape suivante

- On découpe S la séquence requête, en mots de longueur fixée (petite) : les mots requêtes (*query words*)

  ex S=TLSHAWRLSNETDKRPFIETALRDQHKKDYPEYKYQPRRRKNGKP

  query words : TLS, LSH, SHA, ...

- Pour chaque mot requête, on détermine son voisinage

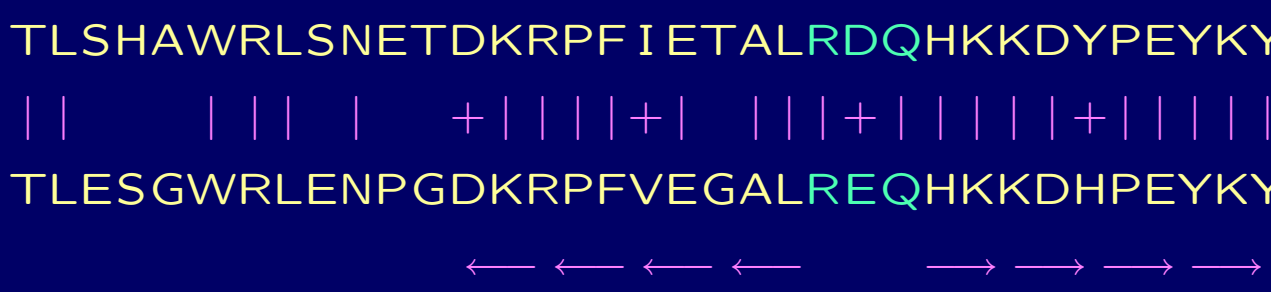
  ex pour RDQ

RDQ 16	QDQ 12	EDQ 11	RDN 11	RDB 11	BDQ 10	RDP 10
RBQ 14	REQ 12	HDQ 11	RDD 11	ADQ 10	XDQ 10	RDT 10
RDZ 14	RDR 12	ZDQ 11	RDH 11	MDQ 10	RQQ 10	RDY 10
KDQ 13	RDQ 12	RNQ 11	RDM 11	SDQ 10	RSQ 10	RDX 10
RDE 13	NDQ 11	RZQ 11	RDS 11	TDQ 10	RDA 10	DDQ 9 ...

- Score seuil T, par exemple $T = 11$, seuls les mots voisins dont le score d'ali. avec le *query word* est $\geq T$ passent à l'étape suivante

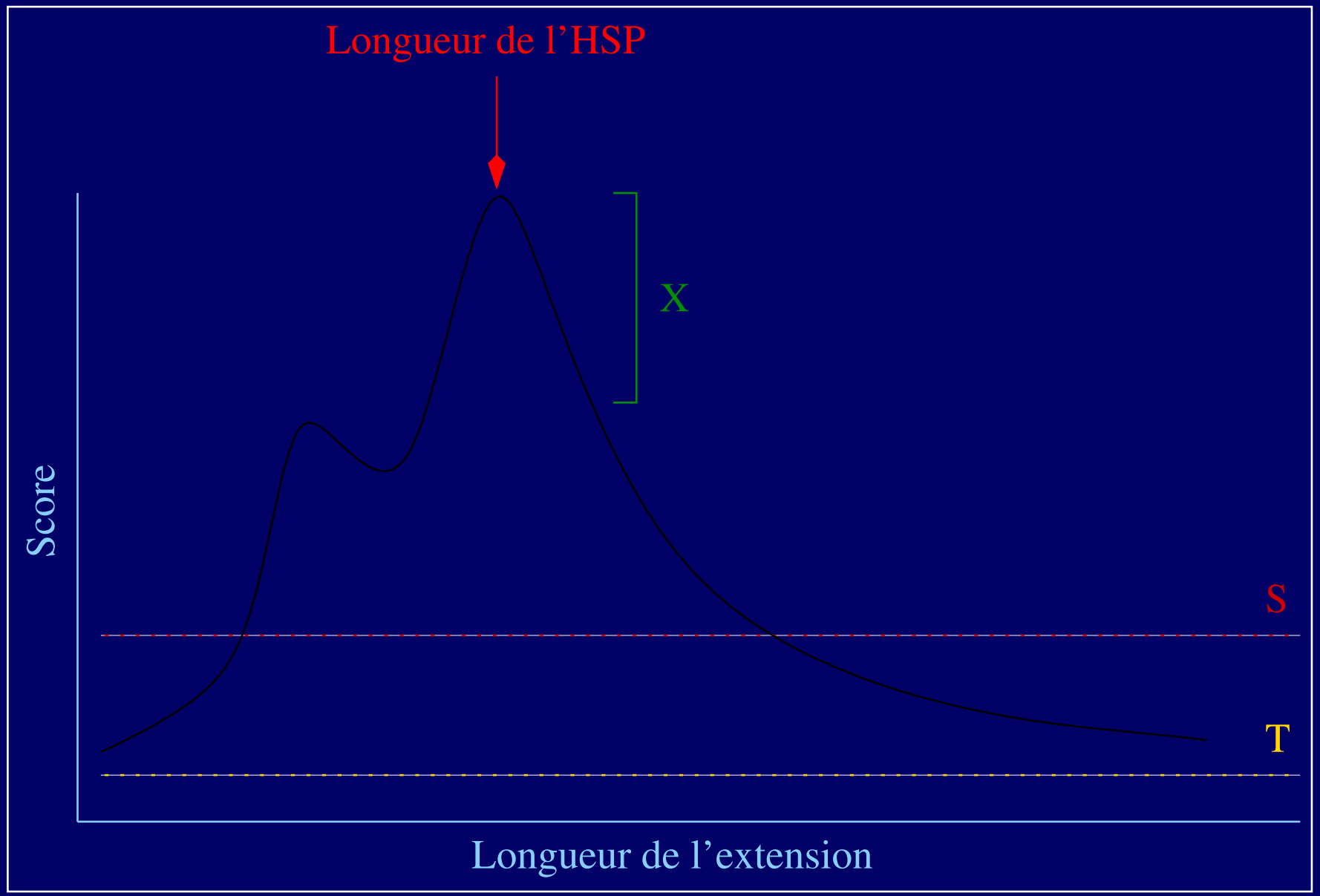
BLAST : L'algorithme (2)

- Aligner le mot requête avec un mot de son voisinage supérieur à T



- Extension jusqu'à la construction d'un alignement local de longueur maximale

- L'alignement résultant est l'Hight-scoring Segment Pair ou HSP (plusieurs HSP possibles par paire de séquences)


Longueur de l'HSP
X
Score
S
T
Longueur de l'extension

BLAST : Mesure de significativité

- Une fois qu'un HSP est identifié il est important de déterminer s'il est vraiment significatif

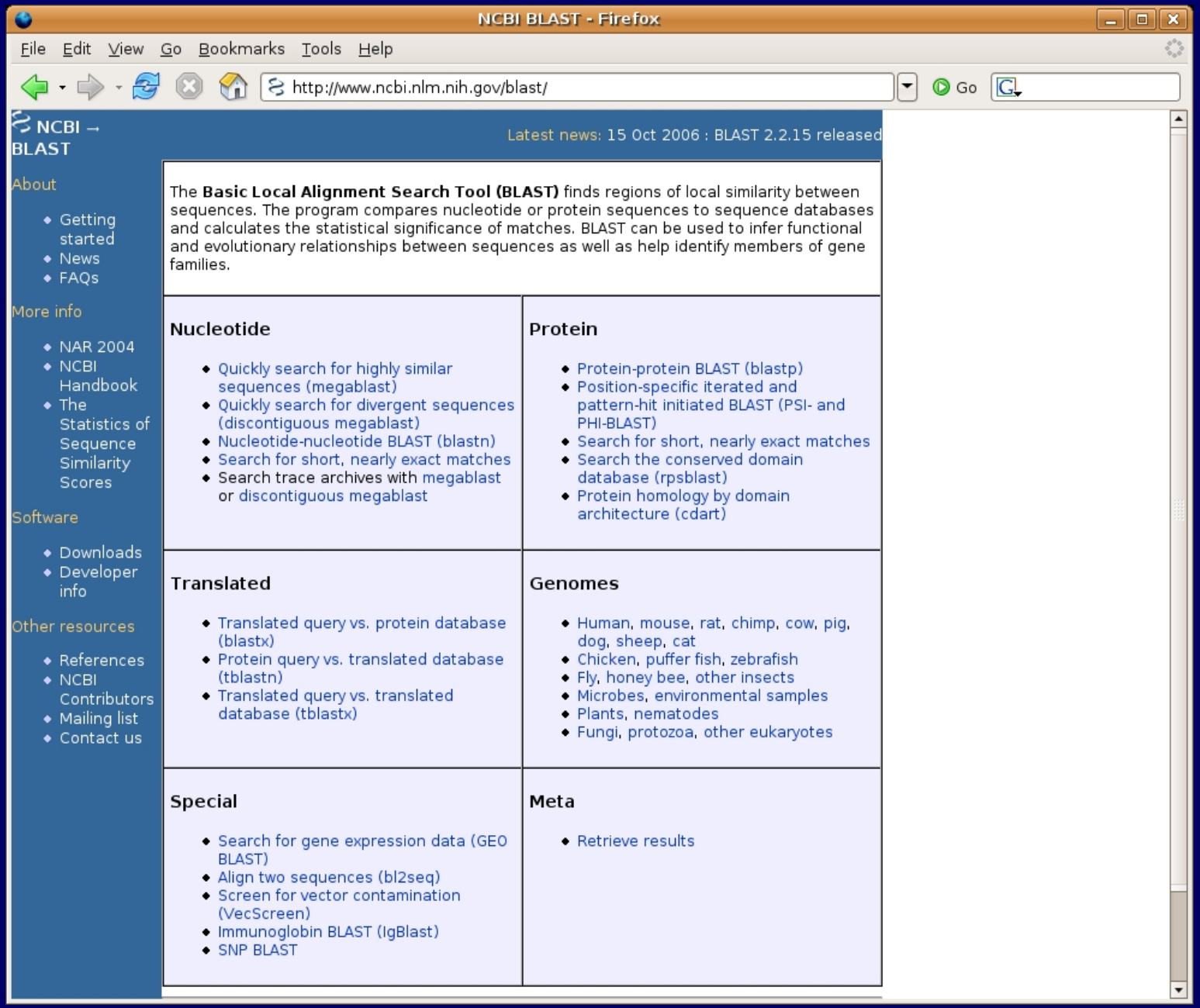
- BLAST calcule une valeur E en utilisant le score de l'alignement et d'autres paramètres : $E = kmNe^{-\lambda S}$

- Pour chaque résultat, E représente le nb d'HSP ayant un score de S ou plus, que BLAST aurait pu trouver par chance

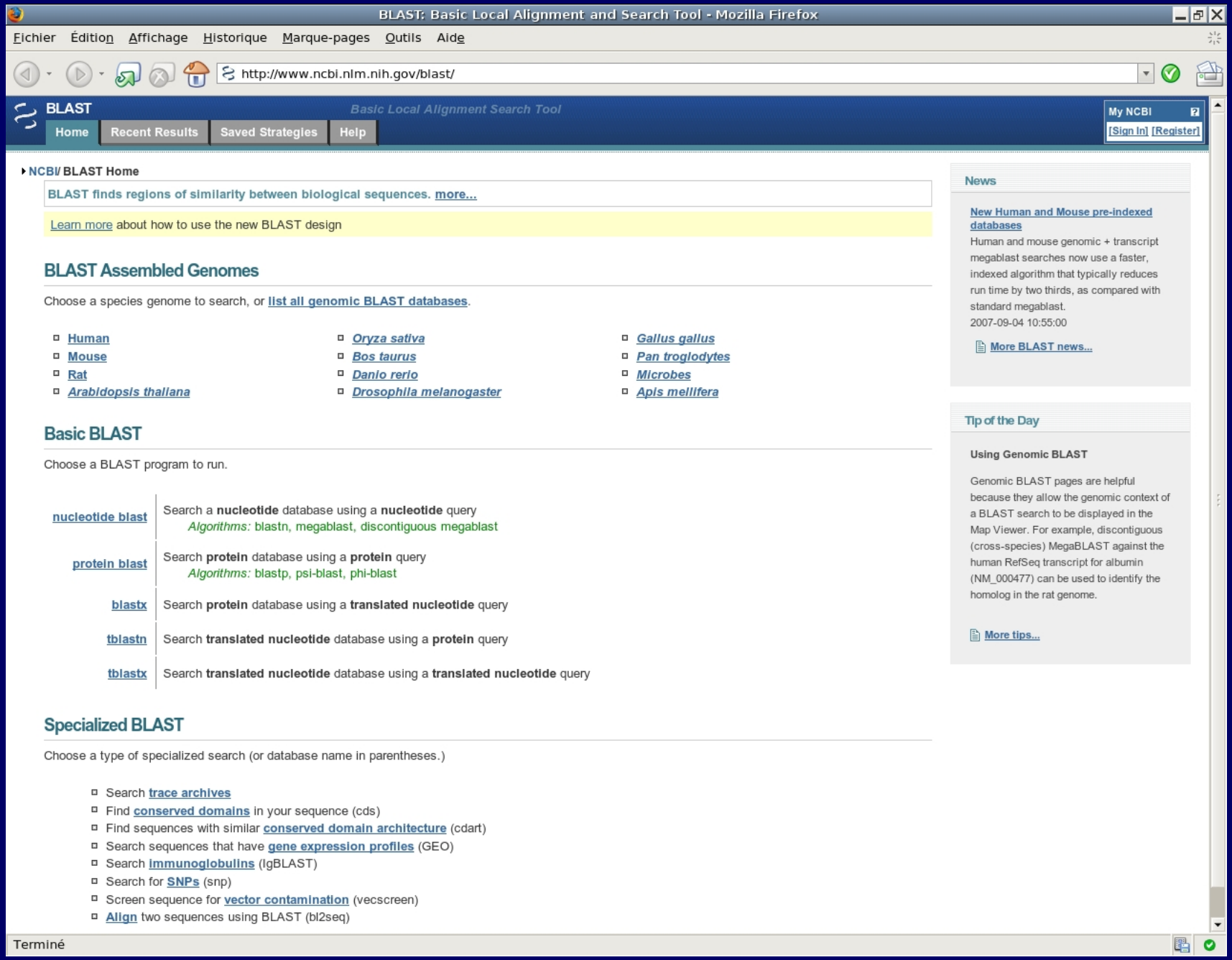
- La valeur de E fournit une mesure indiquant si l'HSP est un faux positif

Plus la valeur de E est petite, plus la similarité est significative

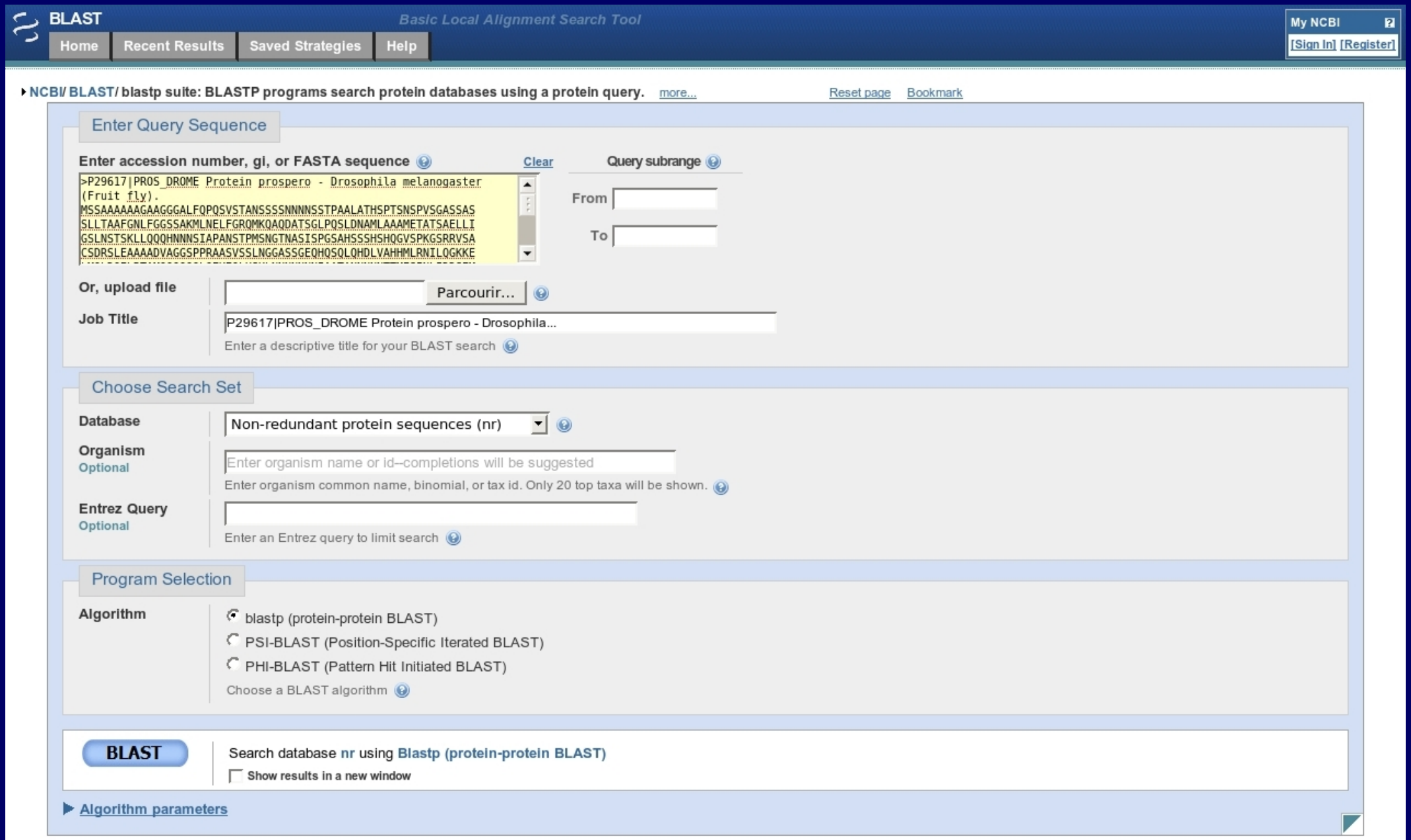
BLAST : Exemple d'utilisation



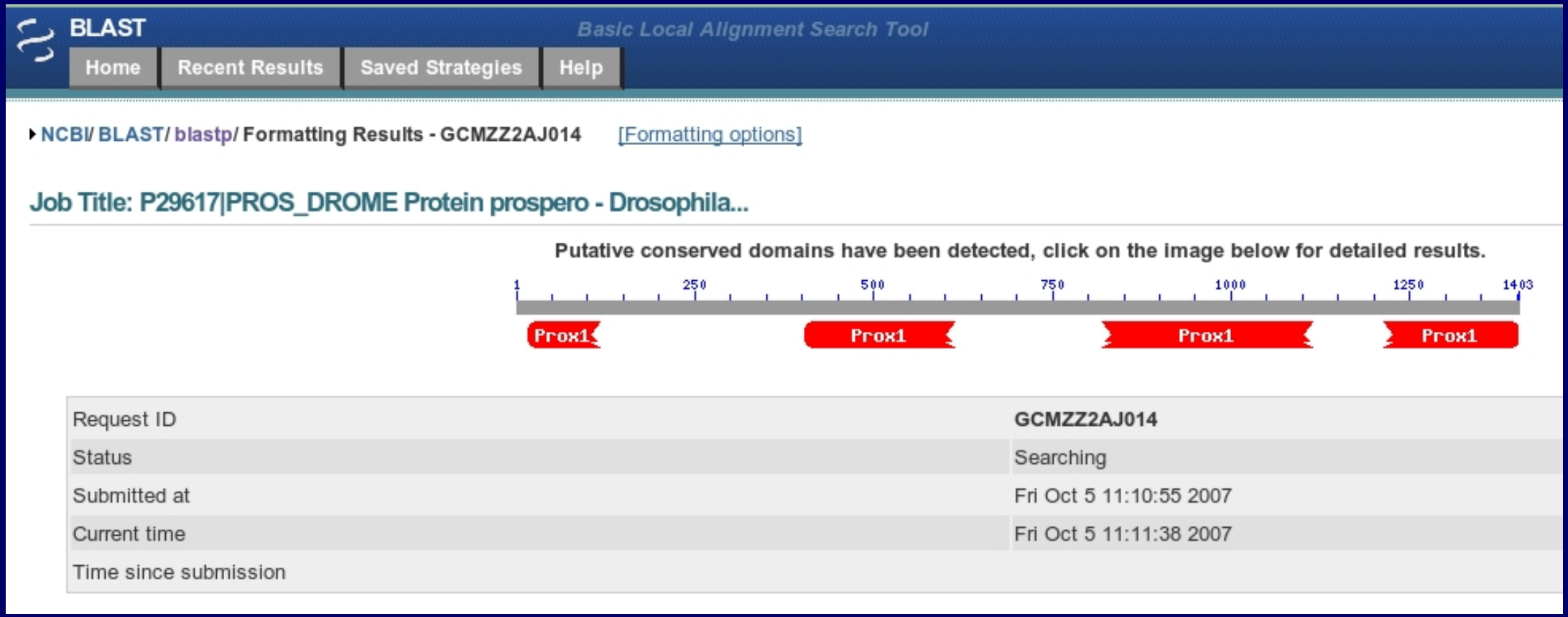
BLAST : nouvelle interface (2007)



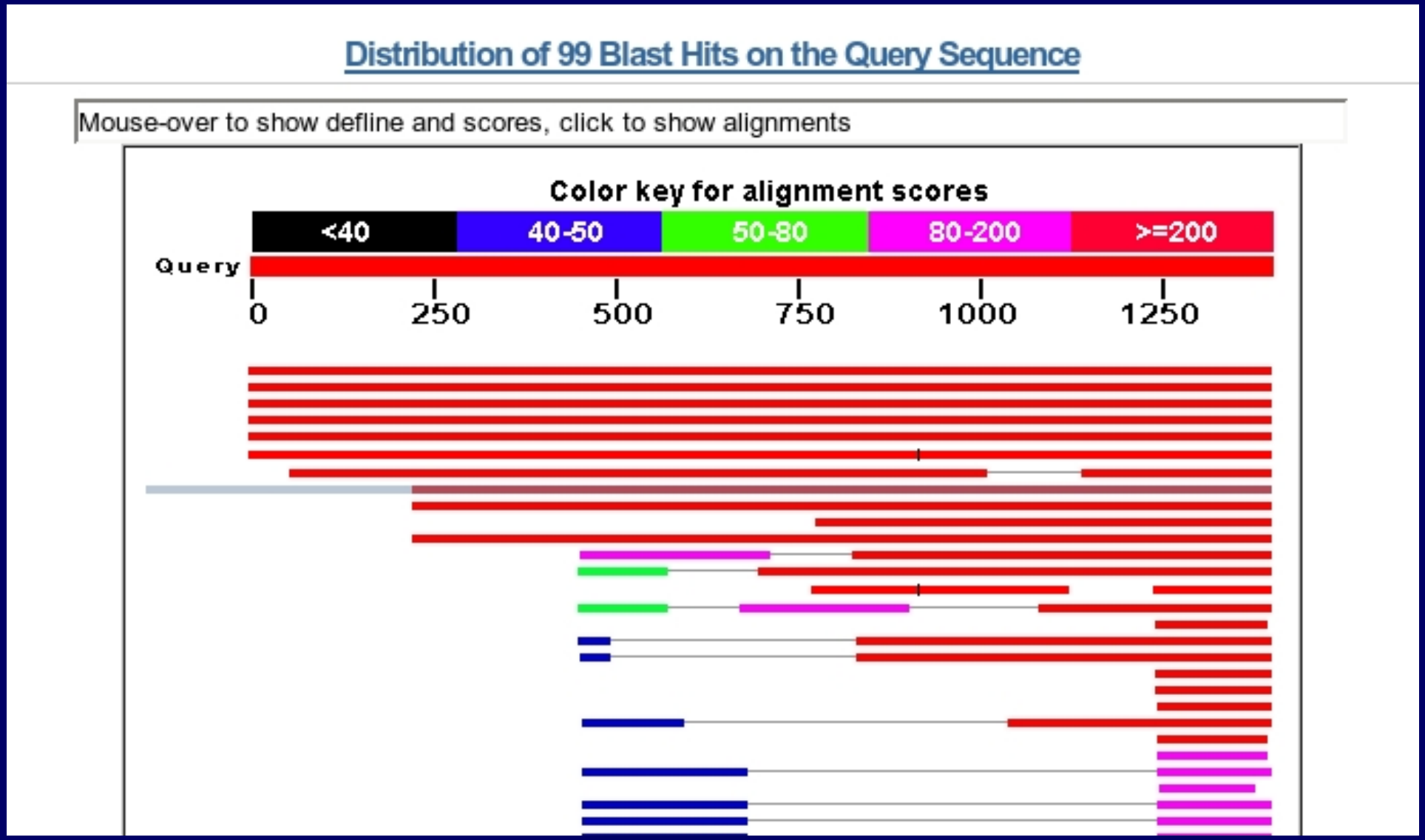
BLAST : Exemple d'utilisation (2)



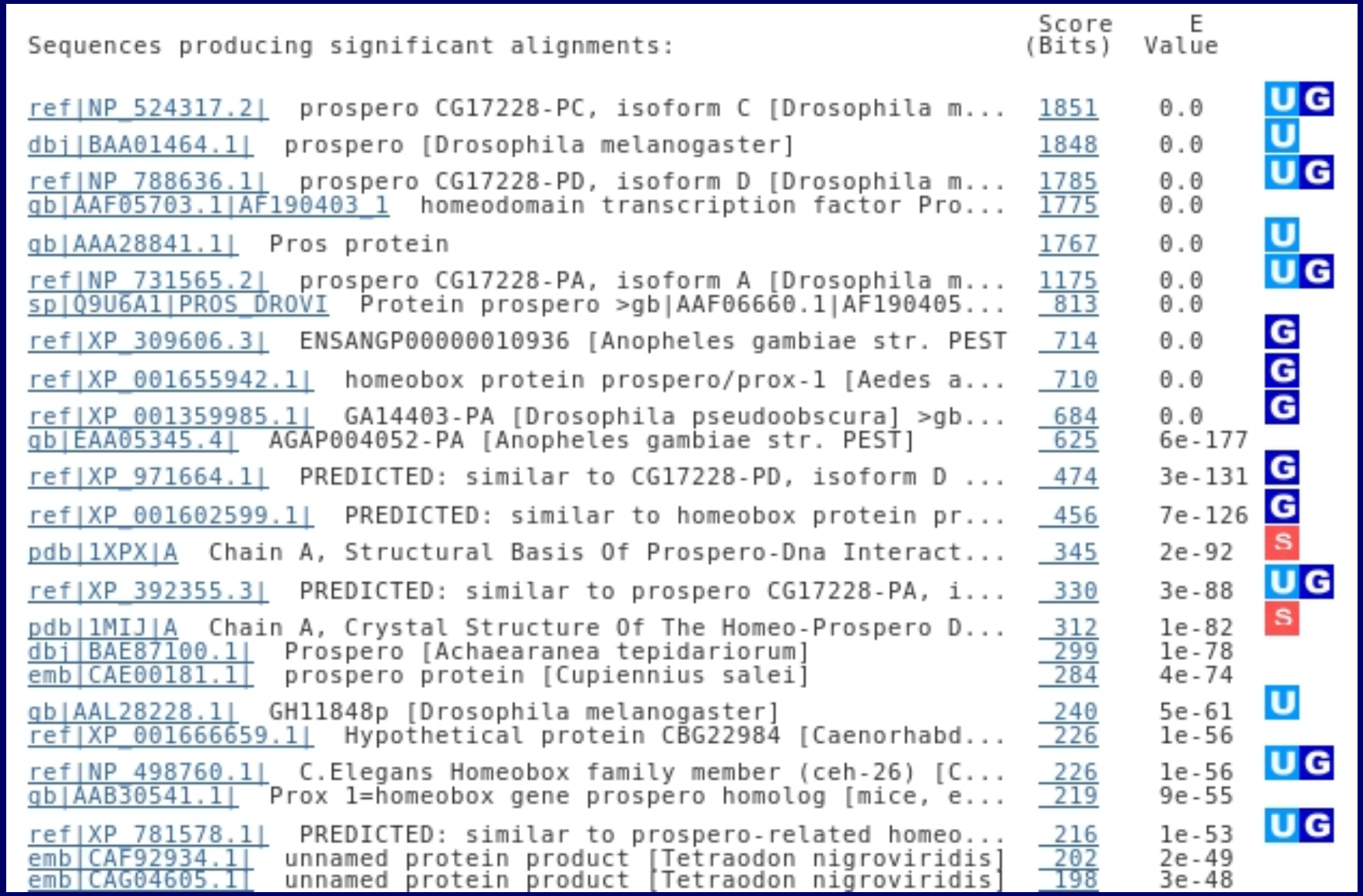
BLAST : Résultats grapiques (1)



BLAST : Résultats grapiques (2)



BLAST : Liste des résultats



Sequences producing significant alignments:	Score (Bits)	E Value
ref|NP_524317.2| prospero CG17228-PC, isoform C [Drosophila m...	1851	0.0
dbj|BAA01464.1| prospero [Drosophila melanogaster]	1848	0.0
ref|NP_788636.1| prospero CG17228-PD, isoform D [Drosophila m...	1785	0.0
gb|AAF05703.1|AF190403_1 homeodomain transcription factor Pro...	1775	0.0
gb|AAA28841.1| Pros protein	1767	0.0
ref|NP_731565.2| prospero CG17228-PA, isoform A [Drosophila m...	1175	0.0
sp|Q9U6A1|PROS_DROVI Protein prospero >gb|AAF06660.1|AF190405...	813	0.0
ref|XP_309606.3| ENSANGP00000010936 [Anopheles gambiae str. PEST	714	0.0
ref|XP_001655942.1| homeobox protein prospero/prox-1 [Aedes a...	710	0.0
ref|XP_001359985.1| GA14403-PA [Drosophila pseudoobscura] >gb...	684	0.0
gb|EAA05345.4| AGAP004052-PA [Anopheles gambiae str. PEST]	625	6e-177
ref|XP_971664.1| PREDICTED: similar to CG17228-PD, isoform D ...	474	3e-131
ref|XP_001602599.1| PREDICTED: similar to homeobox protein pr...	456	7e-126
pdb|1XPX|A Chain A, Structural Basis Of Prospero-Dna Interact...	345	2e-92
ref|XP_392355.3| PREDICTED: similar to prospero CG17228-PA, i...	330	3e-88
pdb|1MIJ|A Chain A, Crystal Structure Of The Homeo-Prospero D...	312	1e-82
dbj|BAE87100.1| Prospero [Achaearanea tepidariorum]	299	1e-78
emb|CAE00181.1| prospero protein [Cupiennius salei]	284	4e-74
gb|AAL28228.1| GH11848p [Drosophila melanogaster]	240	5e-61
ref|XP_001666659.1| Hypothetical protein CBG22984 [Caenorhabd...	226	1e-56
ref|NP_498760.1| C.Elegans Homeobox family member (ceh-26) [C...	226	1e-56
gb|AAB30541.1| Prox 1=homeobox gene prospero homolog [mice, e...	219	9e-55
ref|XP_781578.1| PREDICTED: similar to prospero-related homeo...	216	1e-53
emb|CAF92934.1| unnamed protein product [Tetraodon nigroviridis]	202	2e-49
emb|CAG04605.1| unnamed protein product [Tetraodon nigroviridis]	198	3e-48

BLAST : Alignements


```
>ref|XP_392355.3|  PREDICTED: similar to prospero CG17228-PA, isoform A [Apis mellifera]
Length=1146

 Score =  330 bits (847),  Expect = 3e-88, Method: Composition-based stats.
 Identities = 208/320 (65%), Positives = 234/320 (73%), Gaps = 50/320 (15%)

Query  1083  MMPVSLPTSVAIPNPSLHESKVFSPYSPFFNPHAAAGQATAAQLHQHHQQHHPHHQSMQL  1142
             M+PVSLPTSVAIPNPSLHES+VFSPYSPFFNPHA              H    P      L
Sbjct  876   MLPVSLPTSVAIPNPSLHESQVFSPYSPFFNPHAG------------HPGQVPPPGPHHL  923

Query  1143  SSSPPGSLGALMDSRDSPPLPHPPSMLHPALLAAAHHGGSPDYKTCLRAVMDAQDRQSEC  1202
              +SPP   G  +D RDSP  P P   LHPALLAAA H GSPDY       MD+ +R ++C
Sbjct  924   PASPP---GGGVDVRDSP--PLPHMPLHPALLAAAQH-GSPDYG---HLRMDSNERPNDC  974

Query  1203  NSADMQFDGMAPTISFYKQMQLKTEHQESLMAKHCESLTPLHSSTLTPMHLRKAKLMFFW  1262
             NS D+ +DG+ PT                                SS LTP+HLRKAKLMFFW
Sbjct  975   NSDDISYDGIQPT-----------------------------SSMLTPIHLRKAKLMFFW  1005

Query  1263  VRYPSSAVLKMYFPDIKFNKNNTAQLVKWFSNFREFYYIQMEKYARQAVTEGIKTPDDLL  1322
             VRYPSS++LKMYFPDI+FNKNNTAQLVKWFSNFREFYYIQMEKYARQAV+EG+K  DDL
Sbjct  1006  VRYPSSSILKMYFPDIRFNKNNTAQLVKWFSNFREFYYIQMEKYARQAVSEGVKNADDLR  1065

Query  1323  IAGDSELYRVLNLHYNRNNHIEVPQNFRFVVESTLREFFRAIQGGKDTEQSWKKSIYKII  1382
             + GDSE+YRVLNLHYNRNNHIEVP NFR+VVE TL+EFF+AIQGGKDTEQSWKKSIYK+I
Sbjct  1066  VGGDSEIYRVLNLHYNRNNHIEVPSNFRYVVEQTLKEFFKAIQGGKDTEQSWKKSIYKVI  1125

Query  1383  SRMDDPVPEYFKSPNFLEQL  1402
             SR+DDPVPEYFK+PNFL+QL
Sbjct  1126  SRLDDPVPEYFKTPNFLQQL  1145


 Score = 85.1 bits (209),  Expect = 4e-14, Method: Composition-based stats.
 Identities = 87/236 (36%), Positives = 127/236 (53%), Gaps = 33/236 (13%)

Query  674  GHPALPQGFP----PLLQHMGDMSHAAAMYQQFFFEQEARMAKEAAEQQQQQQQQQQQQQ  729
             G PALP    P      + HMG          Q+ + E               QQQ   ++ +Q
Sbjct  488  GLPALPTEHPHAAAAAMYHMG---------QKLYLE-------------QQQAALERMKQ  525
```

PSI-BLAST

- PSI-BLAST : Position-Specific-Iterated BLAST
- Permet d'identifier des protéines très divergentes
- PSI-BLAST utilise des PSSM (Position-Specific Scoring Matrices)
- Une PSSM est une représentation numérique d'un alignement multiple :

```
BIRC6_HUMAN   RRLAQEAVTLST.....S.........LPLSSSSSVFVRCde.......eRLDIMKVLITGP...ADTPY
COP10_ARATH   KRIQREMAELNI.....D.........PPPDCSAGPKGD-..........NLYHWIATIIGP...SGTPY
FTS1_HUMAN    YSLLAEFTLVVK.....Q.........KLPGVYVQPSYRS.........ALMWFGVIFI--...RHGLY
...
consensus     RRLMKELVLLLT.....Q.........KLPGVYVQPSYRS.........ALVWFGVIFI--...RHGLY
```

PSSM associée :

```
      A   B   C   D   E   F   G   H   I   K   L   M   N   P   Q   R   S   T   V   W   Y   Z
'R'; -10, -6,-25, -9, -2,-13,-19, -7,-16, 16,-14, -6, -1,-17,  1, 18, -6, -4,-13,-20, -4, -2;
'R'; -15,-12,-26,-13, -4,-13,-20, -5,-21, 16,-10, -6, -4,-20,  3, 47, -9, -6,-14,-21, -8, -4;
'L';  -9,-28,-22,-30,-20,  4,-32,-22, 25,-26, 35, 17,-26,-26,-19,-20,-24, -9, 16,-21, -2,-20;
'M';  -7,-14,-20,-19, -8, -8,-23, -7,  0, -5,  6, 14,-13,-19,  5, -3,-13, -7, -4,-19, -3, -2;
...
```

PSI-BLAST : la méthode

- Commence d'abord par un BLASTP avec la protéine requête
- Les séquences obtenues avec une valeur E inférieure à un certain seuil et la protéine requête sont alignées et une PSSM est construite
- La PSSM sert alors de requête pour rechercher de nouvelles séquences
- Une PSSM est reconstruite en intégrant ces séquences et une nouvelle recherche est effectuée
  $\rightarrow$ Cette étape est itérée plusieurs fois
- Fin : quand la recherche converge ou que la limite du nombre d'itération est atteint

Conclusions sur BLAST

- Où couper dans la liste des résultats ?

  → Nucléotides : valeur E $< 10^{-6}$ et identité $> 70\ \%$

  → Protéines : valeur E $< 10^{-3}$ et identité $> 25\ \%$

  Ne pas utiliser ces limites sans réfléchir !
  Est-ce que la matrice utilisée est correcte ? Regarder l'alignement. Quel sens biologique ? . . .

- Les artefacts de la recherche :

  → Score élevé dû à des régions de faible complexité ou d'éléments répétés comme les séquences LINE, SINE ou Alu
  ⇒ Masquage de ces régions (par N ou X pour faible compl.)

  → Attention à la fiabilité des résultats qui peuvent être des protéines hypothétiques (issues de prédiction) ou des EST (moins fiables)

- Donner un score aux alignements
- BLAST
- FASTA
  - → L'algorithme
  - → Un exemple d'utilisation
- Comparaison BLAST/FASTA
- La suite ...

FASTA [Pearson et Lipman, 1988]

FAST Alignment

- Le premier programme largement utilisé pour la recherche de similarité dans les banques
- Plusieurs versions disponibles

Programme	Requête	Base de données	équiv.
FASTA	Nucléotide	Nucléotide	BLASTN
	Protéine	Protéine	BLASTP
FASTX/FASTY	ADN	Protéine	BLASTX
TFASTX/TFASTY	Protéine	ADN traduit	TBLASTN

- Similaire à BLAST : méthode à graînes

FASTA : L'algorithme

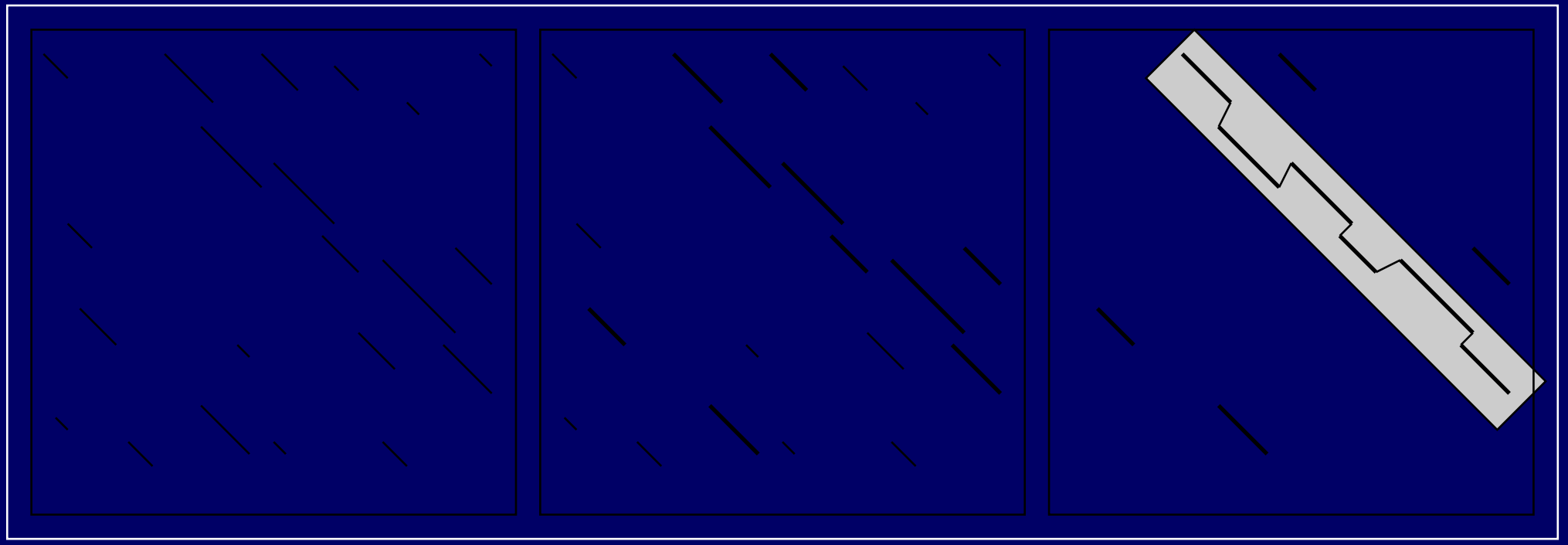
1. Localiser les mots de lg $ktup$ matchant exactement entre la séquence requête et les séquences de la BD
   (ADN $ktup = 4$ ou 6, protéines $ktup = 1$ ou 2)

2. Regarder les 10 diagonales de meilleurs scores pour chaque alignement 2 à 2
   Essayer de les joindre entre elles ($\sim$ hit extension de BLAST)

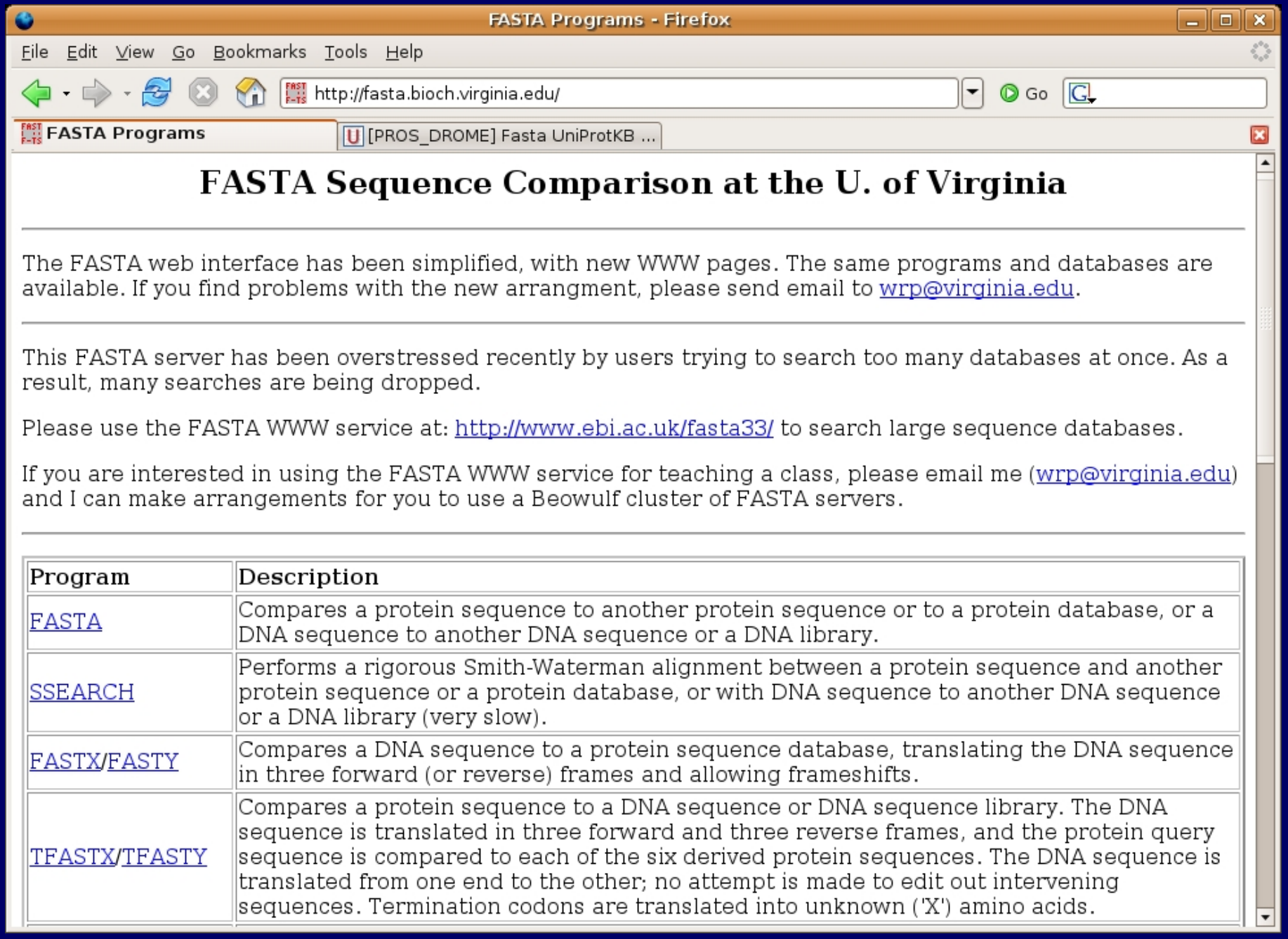
3. Sélectionner les candidats parmi tous les ali. 2 à 2 avec les meilleurs scores et réaligner leurs diagonales avec le principe de programmation dynamique (Smith & Waterman)

4. Calculer les significativités des résultats : calcul d'une valeur E représentant la probabilité que le résultat trouvé le soit par chance

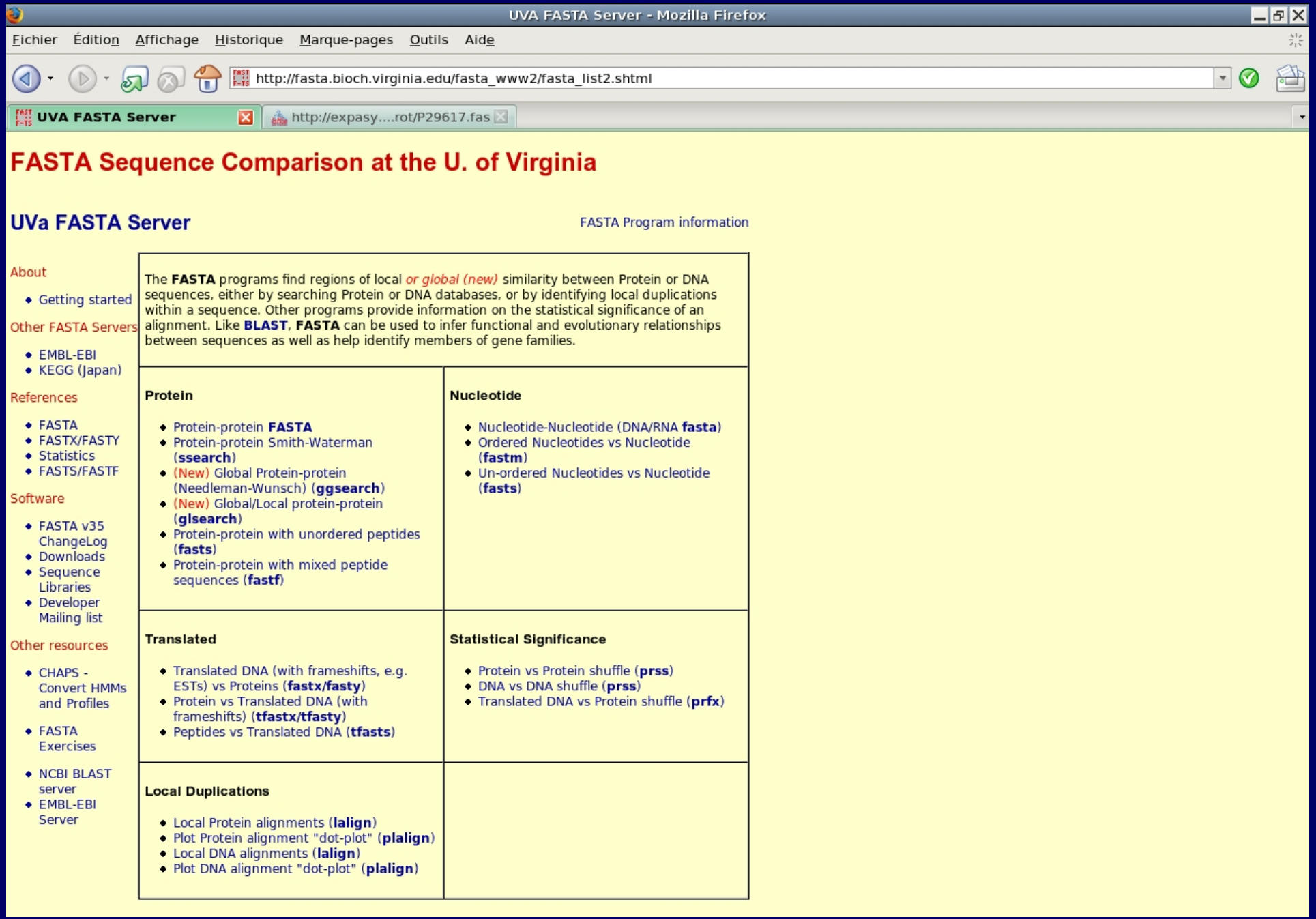
FASTA : Phases 1 & 2



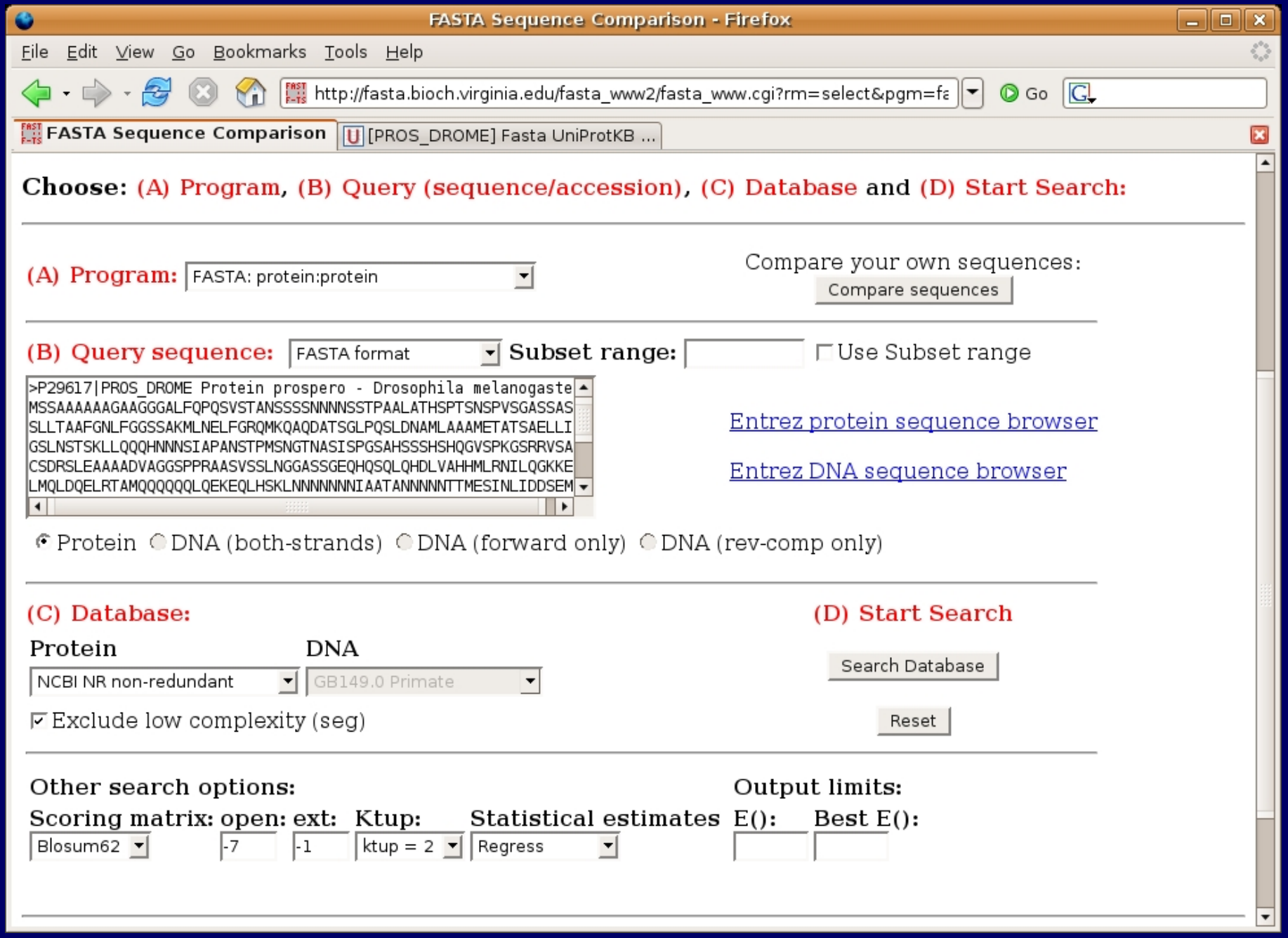
FASTA : Page d'accueil



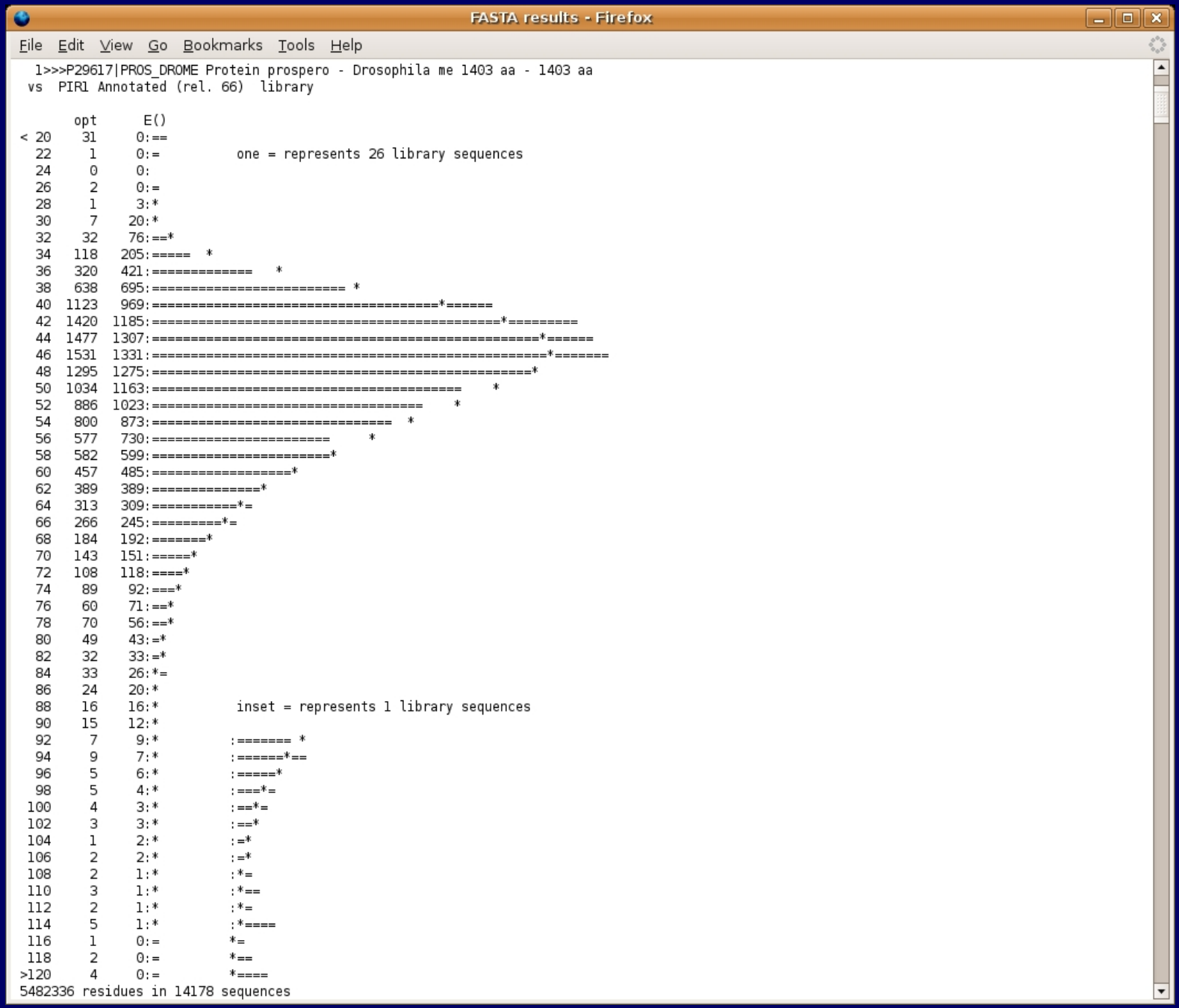
FASTA : nouvelle interface (2007)



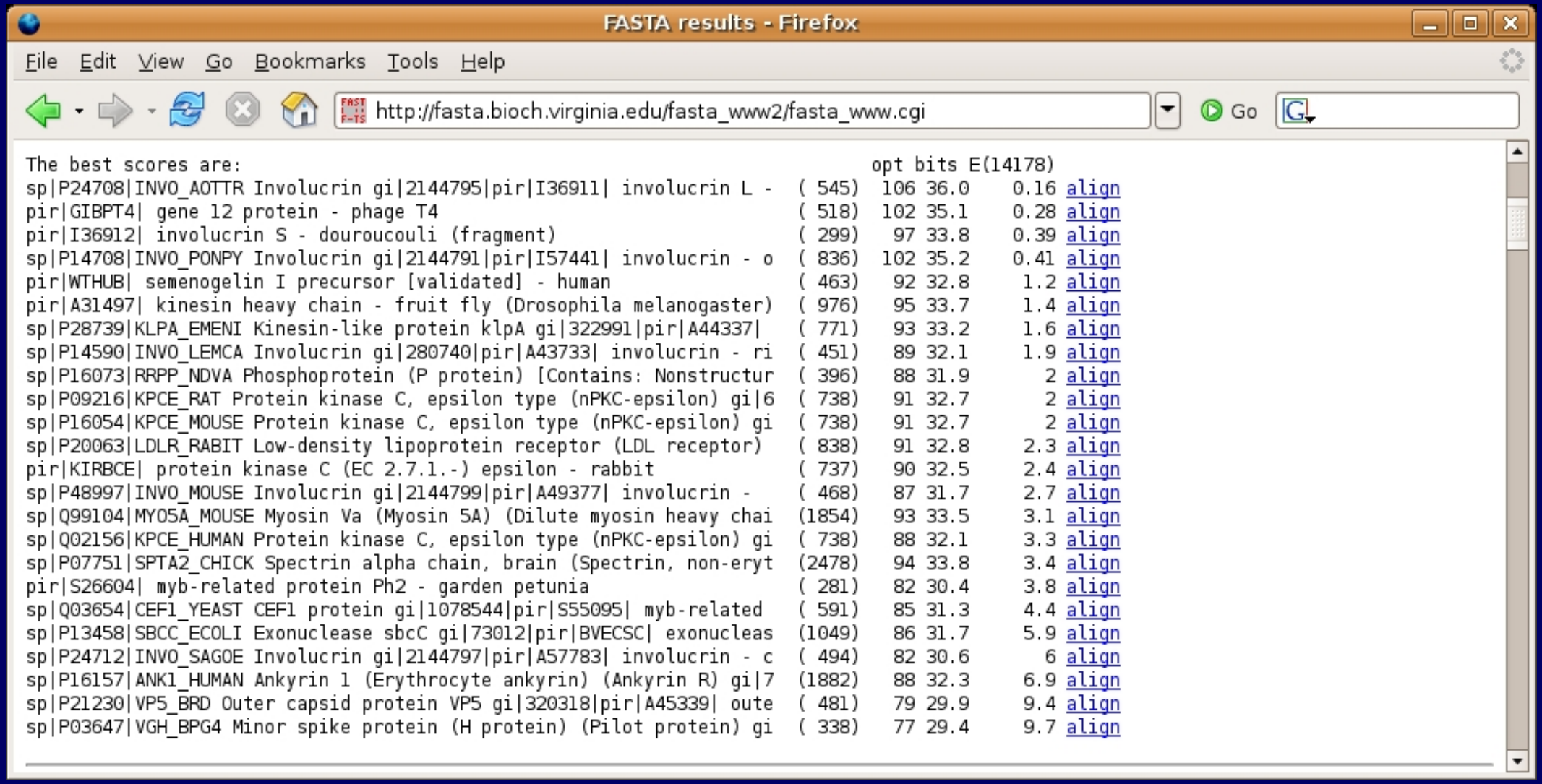
FASTA : Exemple d'utilisation



FASTA : Résumé de la recherche



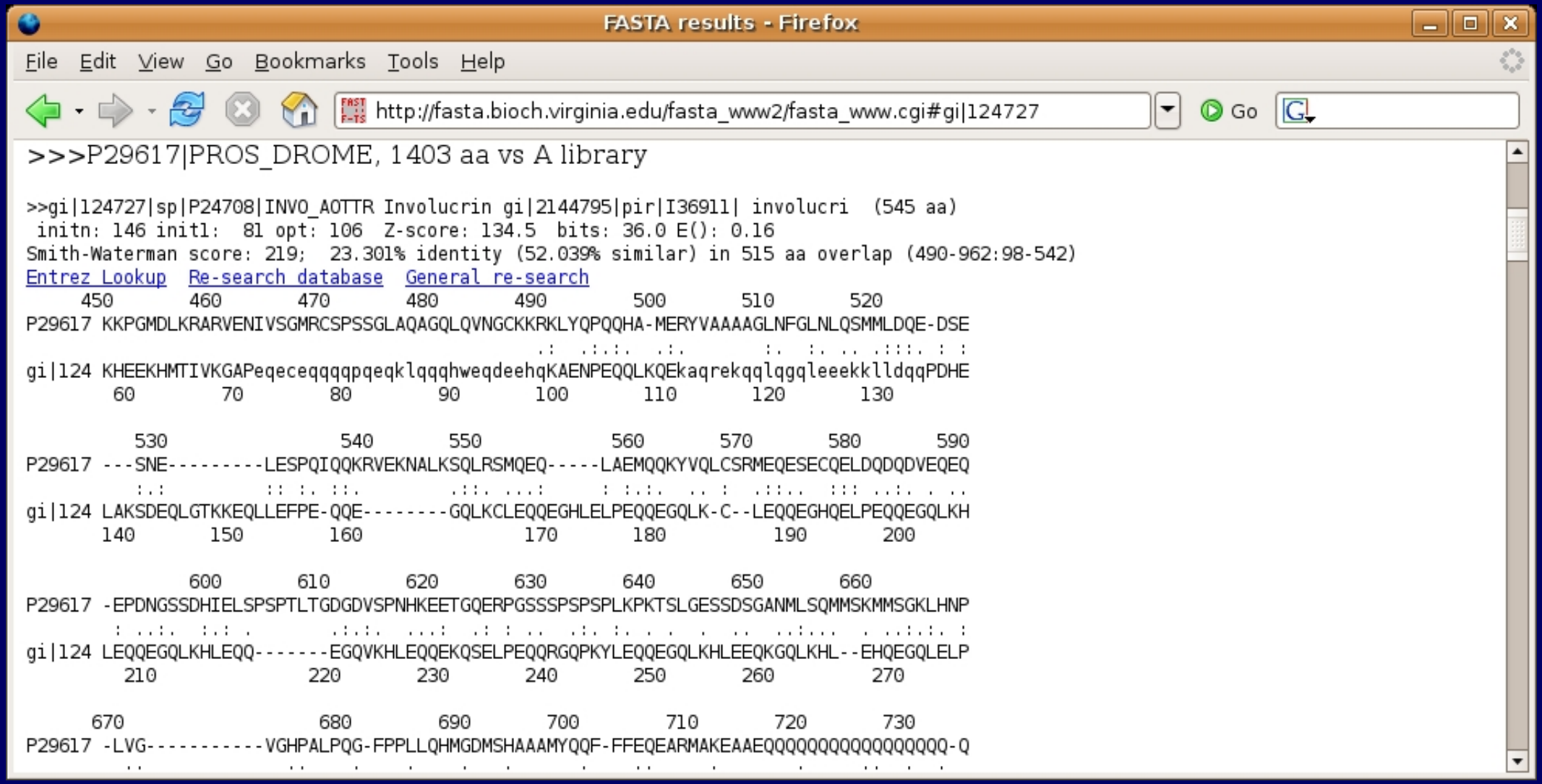
FASTA : Liste des résultats



The best scores are:

		opt	bits	E(14178)	
sp|P24708|INVO_AOTTR Involucrin gi|2144795|pir|I36911| involucrin L -	(545)	106	36.0	0.16	align
pir|GIBPT4| gene 12 protein - phage T4	(518)	102	35.1	0.28	align
pir|I36912| involucrin S - douroucouli (fragment)	(299)	97	33.8	0.39	align
sp|P14708|INVO_PONPY Involucrin gi|2144791|pir|I57441| involucrin - o	(836)	102	35.2	0.41	align
pir|WTHUB| semenogelin I precursor [validated] - human	(463)	92	32.8	1.2	align
pir|A31497| kinesin heavy chain - fruit fly (Drosophila melanogaster)	(976)	95	33.7	1.4	align
sp|P28739|KLPA_EMENI Kinesin-like protein klpA gi|322991|pir|A44337|	(771)	93	33.2	1.6	align
sp|P14590|INVO_LEMCA Involucrin gi|280740|pir|A43733| involucrin - ri	(451)	89	32.1	1.9	align
sp|P16073|RRPP_NDVA Phosphoprotein (P protein) [Contains: Nonstructur	(396)	88	31.9	2	align
sp|P09216|KPCE_RAT Protein kinase C, epsilon type (nPKC-epsilon) gi|6	(738)	91	32.7	2	align
sp|P16054|KPCE_MOUSE Protein kinase C, epsilon type (nPKC-epsilon) gi	(738)	91	32.7	2	align
sp|P20063|LDLR_RABIT Low-density lipoprotein receptor (LDL receptor)	(838)	91	32.8	2.3	align
pir|KIRBCE| protein kinase C (EC 2.7.1.-) epsilon - rabbit	(737)	90	32.5	2.4	align
sp|P48997|INVO_MOUSE Involucrin gi|2144799|pir|A49377| involucrin -	(468)	87	31.7	2.7	align
sp|Q99104|MYO5A_MOUSE Myosin Va (Myosin 5A) (Dilute myosin heavy chai	(1854)	93	33.5	3.1	align
sp|Q02156|KPCE_HUMAN Protein kinase C, epsilon type (nPKC-epsilon) gi	(738)	88	32.1	3.3	align
sp|P07751|SPTA2_CHICK Spectrin alpha chain, brain (Spectrin, non-eryt	(2478)	94	33.8	3.4	align
pir|S26604| myb-related protein Ph2 - garden petunia	(281)	82	30.4	3.8	align
sp|Q03654|CEF1_YEAST CEF1 protein gi|1078544|pir|S55095| myb-related	(591)	85	31.3	4.4	align
sp|P13458|SBCC_ECOLI Exonuclease sbcC gi|73012|pir|BVECSC| exonucleas	(1049)	86	31.7	5.9	align
sp|P24712|INVO_SAGOE Involucrin gi|2144797|pir|A57783| involucrin - c	(494)	82	30.6	6	align
sp|P16157|ANK1_HUMAN Ankyrin 1 (Erythrocyte ankyrin) (Ankyrin R) gi|7	(1882)	88	32.3	6.9	align
sp|P21230|VP5_BRD Outer capsid protein VP5 gi|320318|pir|A45339| oute	(481)	79	29.9	9.4	align
sp|P03647|VGH_BPG4 Minor spike protein (H protein) (Pilot protein) gi	(338)	77	29.4	9.7	align

FASTA : Exemple d'alignement



Comparaison BLAST/FASTA

- Quel est le meilleur ? pas de bonne réponse

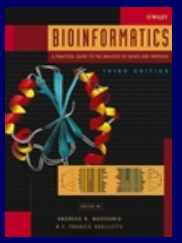
→ FASTA commence la recherche en cherchant des mots exacts alors que BLAST autorise les substitutions conservatives

→ FASTA retourne 1 seul alignement par paire requête/séq. rés., alors que BLAST autorise plrs HSP

→ FASTA utilise un alignement local plus rigoureux, donc ses alignements finaux sont plus fiables

→ BLAST est plus rapide que FASTA

- Chap 11 dans “Bioinformatics : A Practical Guide to the Analysis of Genes and Proteins, 3rd Edition” [Baxevanis & Ouellette,2005]



- Illustrations :

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

  → FASTA : http://fasta.bioch.virginia.edu/fasta_www2/fasta_list2.shtml

Plan

- Donner un score aux alignements
- BLAST
- FASTA
- Comparaison BLAST/FASTA
- La suite ...

Alignements multiples

- Et si l'on veut comparer plusieurs séquences à la fois ?
- Plusieurs séquences $\Rightarrow$ meilleure fiabilité
- Multiple Sequence Alignement ou MSA
- Comme pour les alignements de 2 séquences, il existe des MSA globaux et des MSA locaux