www.facebook.com/DomaineSNV/

Domaine SNV : Biologie,Agronomie,Science Alimentaire,Ecologie

Recherche de séquences par similarité

Jacques van Helden

Jacques.van-Helden@univ-amu.fr
Aix-Marseille Université (AMU), France
Lab. Technological Advances for Genomics and Clinics
(TAGC, INSERM Unit U1090)
http://jacques.van-helden.p.luminy.univ-amu.fr/

Rappel : alignement de séquences avec gap

Exercice

- On dispose des deux séquences suivantes
  - `Pos1  12345678901234567890123456 7`
  - **`Seq1  TTTGCGTTAAATCGTGTAGCAATTTAA`**
  - **`Seq2  AAGAATGGCGTTTTTAATAGCAATAT`**
  - `Pos2  123456789012345678901234 56`
- Questions
  1. En décalant progressivement les séquences, identifiez le(s) décalage(s) qui révèlent des régions de similarité.
  2. A chaque position de décalage, identifiez les segments parfaitement conservés (successions ininterrompue de résidus identiques).
  3. Au vu du résultat, pensez-vous que l'insertion d'un gap permettrait d'augmenter le score d'alignement?

Solution de l'exercice

- Séquences
  - Pos1 12345678901234567890123456 7
  - **Seq1 TTTGCGTTAAATCGTGTAGCAATTTAA**
  - **Seq2 AAGAATGGCGTTTTTAATAGCAATAT**
  - Pos2 12345678901234567890123456
- Décalage -4: la seconde séquence est décalée de 4 nucléotides vers la gauche
  - Pos1 -4 123456789
  - Seq1 ----TTT**GCGTT**AAATCGTGTAGCAATTTAA
  - | ||||| | | |
  - Seq2 AAGAATG**GCGTT**TTTAATAGCAATAT
  - Pos2 12345678901234567890123456
- Décalage -1
  - Pos1 -123456789
  - Seq1 TTTGCGTTAAATCGTG**TAGCAATT**TAA
  - | | ||||||| |
  - Seq2 AAGAATGGCGTTTTTAA**TAGCAAT**AT
  - Pos2 12345678901234567890123456

Alignement avec « gaps » (brèches)

- Les alignements sans gaps sont rarement pertinents, car les divergences entre séquences incluent souvent des insertions et délétions.

- Les gaps permettent de mettre en évidence les régions de similarités multiples.

```
----TTTGCGTT--AAATCGTGTAGCAATTTAA     s=substitution; |=identité
1111s|s|||||11s||22222|||||||s|22     1=gap dans la 1ère séquence
AAGAATGGCGTTTTTAA-----TAGCAATAT--     2=gap dans la 2de séquence
```

- Gaps, insertions et délétions
  - Les "***gaps***" (***brèches***) reflètent soit une insertion dans l'une des séquences, soit une délétion dans l'autre.
  - Sur seule base de l'alignement d'une paire de séquences, on ne peut pas déterminer si un gap correspond à une délétion ou une insertion.
  - On utilise le terme ***indel*** pour désigner cet événement évolutif de nature indéterminée (insertion ou délétion) qui a donné lieu à un gap observé dans un alignement.

Algorithmes de recherche de séquences

Comparaison d'une séquence avec une base de données

- Exemples d'utilisation
  - Nous avons obtenu la séquence d'une protéine de fonction inconnue, et nous désirons la comparer à chacune des séquences d'une base de données de référence (Uniprot) pour émettre des hypothèses concernant sa fonction (prédiction de fonction par similarité).
- Approche: nous alignons successivement notre séquence à chaque entrée d'Uniprot.
- Problème de taille: Uniprot contient ~55 millions d'entrées (avril 2014).
- La programmation dynamique pourrait s'appliquer, mais elle demanderait un temps de calcul important.

Algorithmes rapides pour la recherche de similarités

- En résumé
  - Ces algorithmes, basés sur l'indexation de tous les oligomères ("mots") d'une base de données de séquences, sont ~50 fois plus rapides que celui de Smith-Waterman (1980).
  - Ils se basent cependant sur des approches heuristiques, qui ne peuvent pas garantir de trouver l'alignement optimal.
  - Une comparaison avec les résultats de programmation dynamique a cependant montré que les alignements obtenus sont généralement proches de l'optimum.
- FastA (Lipman & Pearson, 1988)
  - Algorithme de recherche rapide basé sur un index de mots (k-mères)
- BLAST (Basic Local Alignment Search Tool)
  - Version 1990 (Altschul et al., 1990)
    - Version sans gap
    - Apport statistique: calcul de la E-valeur
  - Version 1997 (Altschul et al., 1997)
    - Version avec gap (BLAST)
    - Version itérative (PSI-BLAST) basée sur des matrices de profil

- Lipman and Pearson. Rapid and sensitive protein similarity searches. Science (1985) vol. 227 (4693) pp. 1435-41
- Altschul et al. Basic local alignment search tool. J. Mol. Biol; 1990
- Altschul et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res, 1997, 25:3389-402

Stratégie de FastA – l'indexation des mots

- Préalablement à l'analyse, le programme indexe tous les mots (k-mères) d'une taille donnée présents dans les séquences de la base de données. Cette opération ne doit être faite qu'une fois au départ (formatage de la DB), on peut ensuite soumettre de nombreuses requêtes.
- Au moment de la requête, FastA construit un index avec les positions de tous les mots trouvés dans la séquence de requête.
- Le programme détecte des diagonales de mots alignés entre la requête et la base de données.
- Quant une diagonale significative est détectée, les deux séquences sont alignées par l'algorithme Smith-Waterman.
- La taille des mots (k) influence fortement le comportement du programme.
  - Quand k augmente, la recherche est accélérée mais on peut louper des similarités pertinentes.

```
                    1         2         3
pos  123456789012345678901234567890123456
seq  MVDFYYLPGSSMVDVFDFYAKAVGVELNKKLL
```

3-mères

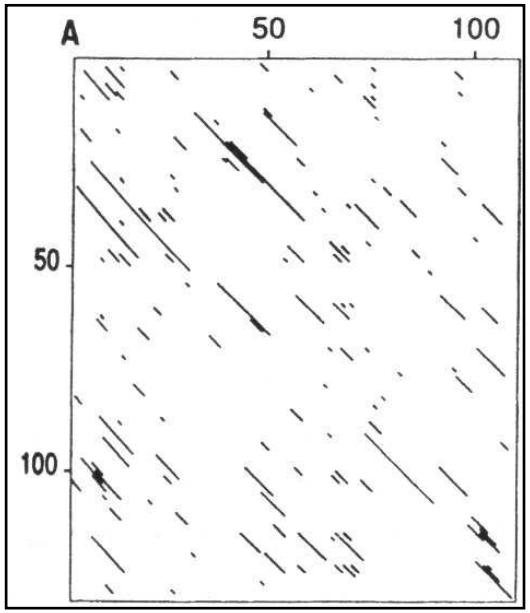
- MVD
- VDF
- DFY
- FYY
- YYL
- ...

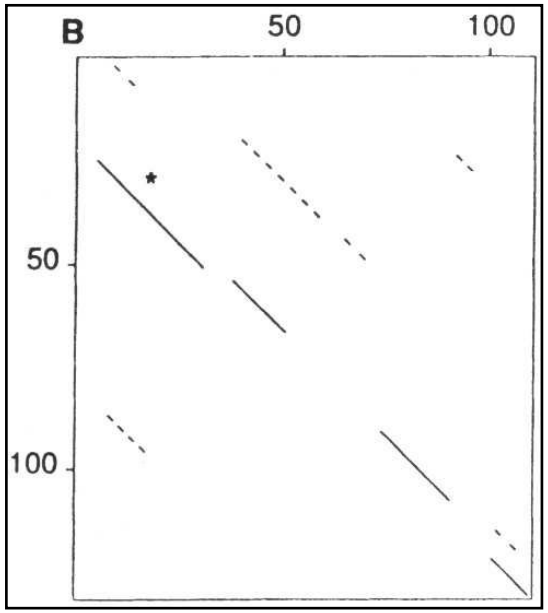
position index

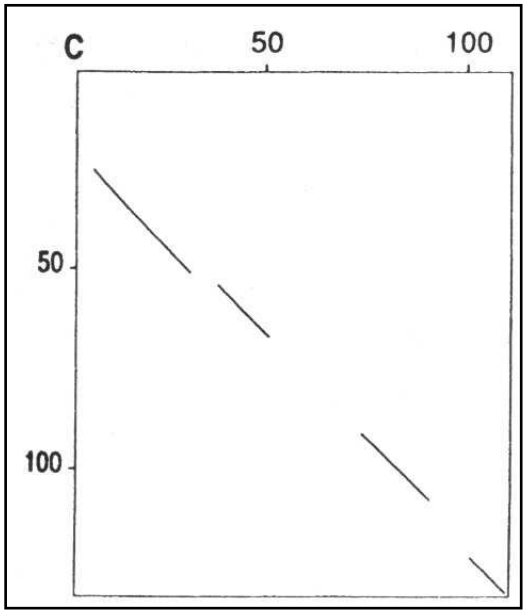
k-mère	positions
MVD	1, 12
VDF	2
DFY	3,17
...	...

Principe de la stratégie de FastA

- A gauche
  - Comparaisons de tous les « mots » (k-mères) entre les séquences de requête et de la base de données, et assignation de scores à chaque position possible d'alignement
- Au centre:
  - Les régions de forte densité (« régions d'initiation ») sont identifiées.
  - La meilleure région d'initiation est marquée d'une étoile.
  - Les régions associées à un score trop faibles sont marquées en pointillés, pour illustration.
- A droite
  - Les régions de faible score sont filtrées, et les régions restantes sont jointes pour former l'alignement.







Source: Mount (2000)

Stratégies de BLAST (Altschul et al., 1990; 1997)

- Version 1 (1990): BLAST sans gaps
  - ***Indexation*** préalable de tous les mots (k-mères) de la base de données (formatdb).
  - Au démarrage de la requête, construction d'un ***dictionnaire de mots*** trouvés dans la séquence requête.
  - Utilisation d'une ***matrice de substitution*** (par ex BLOSUM) pour calculer le score entre chaque mot de la séquence requête et tous les mots trouvés dans la base de données.
  - Sélection des mots avec un score suffisant (***seuil*** sur le score de paires de mots).
  - Chaque fois qu'un mot du dictionnaire passe le seuil (***hit***), ***étendre*** dans les deux directions pour obtenir une "***High-scoring Segment Pair***" (***HSP***).
  - Le programme retourne les alignements avec des HSP significatifs.

HSP 1

```
GGTAGCAAATGTCCTGTCTGTACTGTACATGGTCAAACTGGTGAAT
      ||||||||||||
TGTATCAAATGTCCTGTGTGAATGGTAGATGGTCAAACTGGTCAAT
```

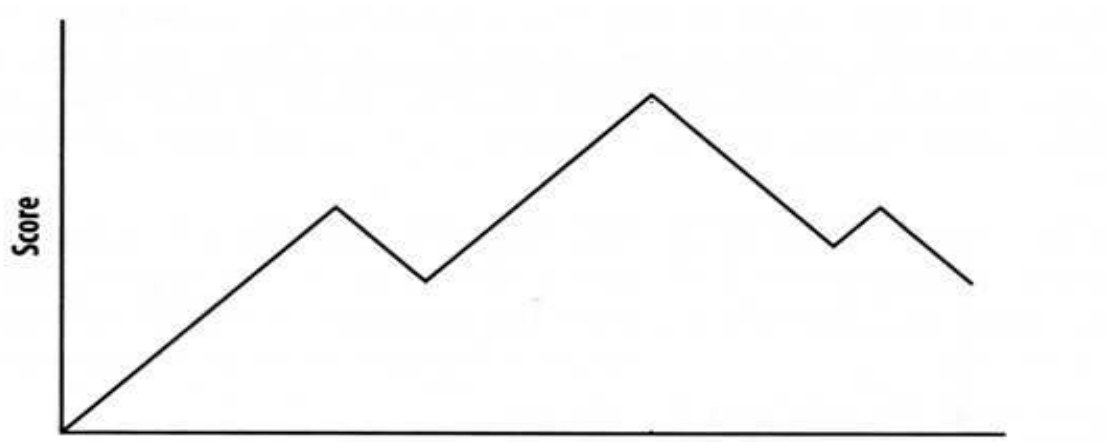
BLAST - Elongation de l'alignement

- Imaginons un exemple simplifié: nous voulons déterminer le plus long segment similaire entre ces deux phrases.
- Nous définissons (arbitrairement) les scores suivants
  - Identité: 1
  - Substitution: -1

```
The quick brown fox jumps over the lazy dog
The quiet brown cat purrs when she sees him
```

BLAST - Elongation de l'alignement

- Nous identifions le segment aligné sans substitution (***HSP*** pour "*highest scoring pair*"), et nous l'étendons
  - en ajoutant un point chaque fois que les deux lettres sont identiques;
  - en retirant un point chaque fois qu'elles diffèrent.
- Le graphique indique le score cumulé en fonction de la position sur la séquence.



Score d'identité = 1
Score de substitution = -1

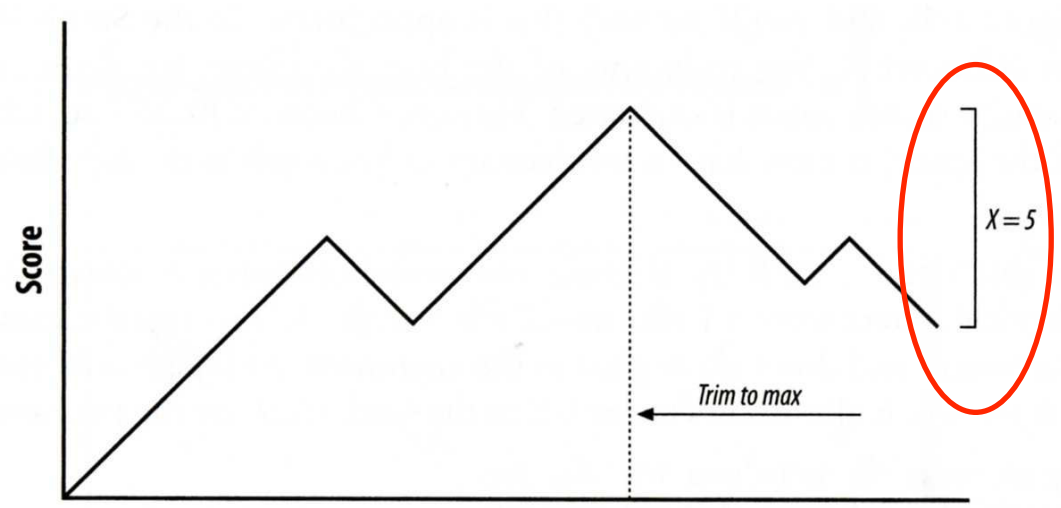
```
The quick brown fox jump
The quiet brown cat purr
123 45654 56789 876 5654   <= SCORE
```

BLAST - Elongation de l'alignement

- On calcule ensuite, pour chaque position, la différence entre le score actuel et le meilleur score en amont dans la séquence.
- L'élongation s'arrête si

  max(score) – score > x

  où x est une limite prédéfinie (x = 5 dans ce cas-ci)

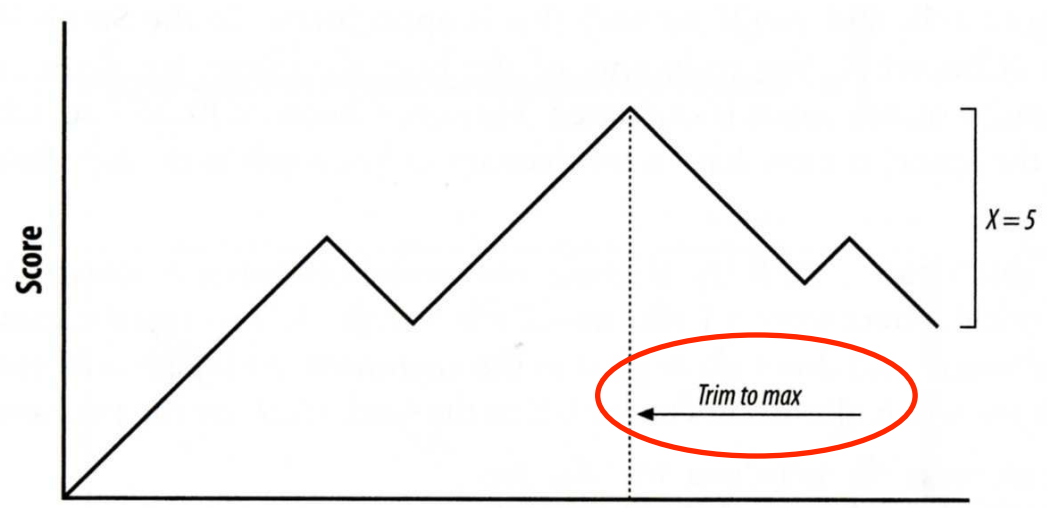


Score d'identité = 1
Score de substitution = -1

```
The quick brown fox jump
The quiet brown cat purr
123 45654 56789 876 5654   <= SCORE
000 00012 10000 123 4345   <=(SCORE(max)-SCORE)
```

BLAST - Elongation de l'alignement

- On remonte ensuite l'alignement jusqu'au dernier max(score)
- On peut ensuite procéder de façon similaire pour allonger l'alignement de l'autre côté du *HSP*.



HSP (High Scoring Pair):

```
The quick brown
|||||||  ||||||
The quiet brown
```

```
The quick brown fox jump
The quiet brown cat purr
123 45654 56789 876 5654   <= SCORE
000 00012 10000 123 4345   <=(SCORE(max)-SCORE)
```

BLAST - Elongation de l'alignement

- Identification des HSP
- Elongation de l'alignement de deux côtés à partir des mots du dictionnaire
  - L'élongation s'arrête si le score diminue en-deçà d'une limite prédéfinie par rapport au dernier maximum.
  - L'alignement est écourté jusqu'au dernier score maximal

BLAST - Exercice

- Faites un alignement local entre ces deux séquences en suivant l'algorithme de BLAST version1
- Scores
  - Identité: 1
  - Substitution: -1
  - Différence maximale entre le score actuel et le score maximal: 5

- Etape 1: identifiez le "HSP": segment identique maximal (sans substitution ni gap)

```
Position 1          2          3
123456789012345678901234567890123456789
TAAATGGTCATGTGATGGTCCTGACTGATGCTGCCTGA
GAAATGGTCATGTGATGGTCGTAACGATGCAATTGGGC
```

BLAST - Exercice

- Faites un alignement local entre ces deux séquences en suivant l'algorithme de BLAST version1
- Scores
  - Identité: 1
  - Substitution: -1
  - Différence maximale entre le score actuel et le score maximal: 5

- Etape suivante: procédez à l'élongation à droite

```
Position 1          2          3
12345678901234567890123456789012345678
TAAATGGTCATGTGATGGTCCTGACTGATGCTGCCTGA
GAAATGGTCATGTGATGGTCGTAACGATGCAATTGGGC
```

Noyau (HSP)

BLAST - Exercice

- Scores: Identité: 1: Substitution: -1
- Différence maximale entre le score actuel et le score maximal: x = 5
- Résultats
  - Maxima locaux: **19** (position 20); **20** (position 25)

- Etape suivante: calculer la différence entre chaque score et le maximum local précédent.

```
Position 1          2          3
123456789012345678901234567890123456 78
TAAATGGTCATGTGATGGTCCTGACTGATGCTGCCTGA Seq1
GAAATGGTCATGTGATGGTCGTAACGATGCAATTGGGC Seq2
 123456789111111111111211111
Score     012345678989890987 65
```

BLAST - Exercice

- Scores: Identité: 1: Substitution: -1
- Différence maximale entre le score actuel et le score maximal: x = 5
- Résultats
  - Maxima locaux: **19** (position 20); **20** (position 25)
- Interruption de l'élongation: position 30 (score =**15** < **20** – 5)

- Etape suivante: remontez l'alignement jusqu'au score maximum local précédent

```
Position 1          2          3
123456789012345678901234567890123456678
TAAATGGTCATGTGATGGTCCTGACTGATGCTGCCTGA Seq1
GAAATGGTCATGTGATGGTCGTAACGATGCAATTGGGC Seq2
 123456789111111111111211111
Score     012345678989098765
 000000000000000000010100012345 Score(max)-Score
```

BLAST - Exercice

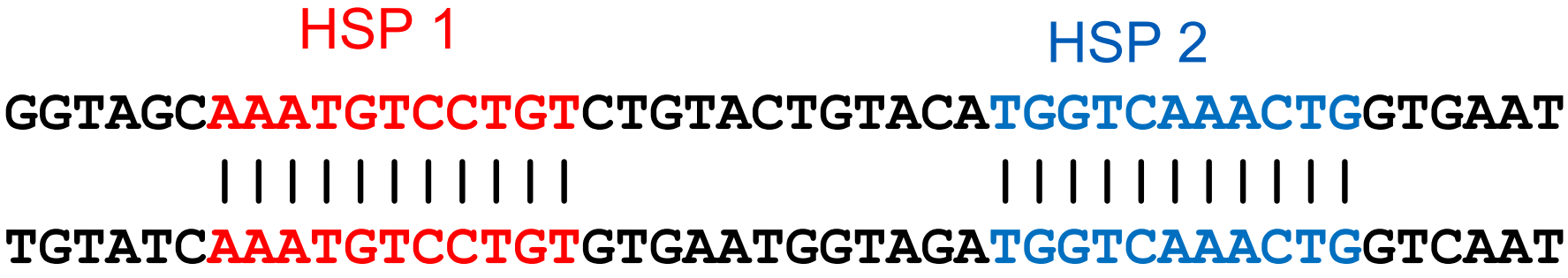
- Scores: Identité: 1: Substitution: -1
- Différence maximale entre le score actuel et le score maximal: x = 5
- Résultats
  - Maxima locaux: **19** (position 20); **20** (position 25)
- Interruption de l'élongation: position 30 (score =**15** < **20** – 5)
- Fin de l'alignement local: position 25

- **Le programme retourne l'alignement s'étendant des positions 2 à 25**

```
Position 1          2          3
123456789012345678901234567890123456 78
TAAATGGTCATGTGATGGTCCTGACTGATGCTGCCTGA Seq1
 |||||||||||||||||||| | ||
GAAATGGTCATGTGATGGTCGTAACGATGCAATTGGGC Seq2
 1234567891111111111111211111
Score     012345678989098765
 000000000000000000010012345 Score(max)-Score
```

Stratégies de BLAST (Altschul et al., 1990; 1997)

- Version 1 (1990): BLAST sans gaps
  - ***Indexation*** préalable de tous les mots (k-mères) de la base de données (formatdb).
  - Au démarrage de la requête, construction d'un ***dictionnaire de mots*** trouvés dans la séquence requête.
  - Utilisation d'une ***matrice de substitution*** (par ex BLOSUM) pour calculer le score entre chaque mot de la séquence requête et tous les mots trouvés dans la base de données.
  - Sélection des mots avec un score suffisant (***seuil*** sur le score de paires de mots).
  - Chaque fois qu'un mot du dictionnaire passe le seuil (***hit***), ***étendre*** dans les deux directions pour obtenir une "***High-scoring Segment Pair***" (***HSP***).
  - Le programme retourne les alignements avec des HSP significatifs.
- Version 2 (1997)
  - Utilisation de mots , mais ne procéder à l'extension que si l'on trouve deux hits sur la même diagonale.
  - L'extension repose sur la programmation dynamique -> permet d'inclure des gaps
  - L'extension coûte donc plus de temps de calcul, mais elle est initiée beaucoup moins fréquemment.



Stratégies de BLAST (Altschul et al., 1990; 1997)

- Version 1 (1990): BLAST sans gaps
  - ***Indexation*** préalable de tous les mots (k-mères) de la base de données (formatdb).
  - Au démarrage de la requête, construction d'un ***dictionnaire de mots*** trouvés dans la séquence requête.
  - Utilisation d'une ***matrice de substitution*** (par ex BLOSUM) pour calculer le score entre chaque mot de la séquence requête et tous les mots trouvés dans la base de données.
  - Sélection des mots avec un score suffisant (***seuil*** sur le score de paires de mots).
  - Chaque fois qu'un mot du dictionnaire passe le seuil (***hit***), ***étendre*** dans les deux directions pour obtenir une "***High-scoring Segment Pair***" (***HSP***).
  - Le programme retourne les alignements avec des HSP significatifs.
- Version 2 (1997)
  - Utilisation de mots , mais ne procéder à l'extension que si l'on trouve deux hits sur la même diagonale.
  - L'extension repose sur la programmation dynamique -> permet d'inclure des gaps
  - L'extension coûte donc plus de temps de calcul, mais elle est initiée beaucoup moins fréquemment.
- PSI-BLAST (également dans l'article de 1997)
  - Un traitement secondaire après avoir fait tourner un BLAST normal (avec gap).
  - Alignement multiple des séquences retournées par BLAST, et construction d'un profil.
  - Scanning de la base de données avec ce motif, pour collecter un nouveau jeu de séquences.
  - Répétition de ce processus
    - Collecte de séquences > construction de profil -> collecte de séquences -> ...

Quelques pièges pour les recherches avec BLAST

- Domaines ubiquitaires
  - Certains domaines se retrouvent dans un grand nombre de protéines. Ceci ne signifie pas que ces protéines ont la même une fonction.
  - La longueur des alignements doit être analysé pour établir si la région alignée couvre l'ensemble de la séquence, ou seulement un segment délimité.
- Régions de faible complexité (séquences répétitives).
  - Certaines séquences se retrouvent répétées à divers endroits du génome, sans qu'on puisse pour autant leur attribuer une fonction spécifique.
  - Le génome humain comporte différents types de séquences répétées : Alu, LINES, SINES, …
  - Ces séquences posent des problèmes pour les statistiques de mots, qui reposent sur une hypothèse d'indépendance.
  - BLAST est muni d'un filtre permettant d'ignorer les régions de faible complexité.
- Vecteurs de clonage
  - Certaines entrées des bases de données de séquences contiennent, par erreur d'encodage, des fragments des vecteurs de clonage.
  - Ceci peut susciter des résultats non pertinents, où la région de similarité est restreinte au vecteur de clonage.
- … quelques autres pièges à découvrir par la pratique

Scores d'alignements

Statistiques d'alignements – le score brut (raw score S)

- Le ***score brut*** est calculé en faisant la somme des scores de la matrice de substitution pour chaque paire de résidus ($r_{1,i}$ and $r_{2,i}$) tout au long de l'alignement (L).

		Ala A	Arg R	Asn N	Asp D	Cys C	Gln Q	Glu E	Gly G	His H	Ile I	Leu L	Lys K	Met M	Phe F	Pro P	Ser S	Thr T	Trp W	Tyr Y	Val V
Ala	A	4																			
Arg	R	-1	5																		
Asn	N	-2	0	6																	
Asp	D	-2	-2	1	6																
Cys	C	0	-3	-3	-3	9															
Gln	Q	-1	1	0	0	-3	5														
Glu	E	-1	0	0	2	-4	2	5													
Gly	G	0	-2	0	-1	-3	-2	-2	6												
His	H	-2	0	1	-1	-3	0	0	-2	8											
Ile	I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4
		Ala A	Arg R	Asn N	Asp D	Cys C	Gln Q	Glu E	Gly G	His H	Ile I	Leu L	Lys K	Met M	Phe F	Pro P	Ser S	Thr T	Trp W	Tyr Y	Val V

$$S = \sum_{i=1}^{L} s_{r_{1,i} r_{2,i}}$$

R L A S V E T D M P L T L R Q H

T L T S L Q T T L K A H L G T H

Statistiques d'alignements – calcul du score brut (raw score S)

- Le ***score brut*** est calculé en faisant la somme des scores de la matrice de substitution pour chaque paire de résidus ($r_{1,i}$ and $r_{2,i}$) tout au long de l'alignement (L).

		Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	A	4																			
Arg	R	-1	5																		
Asn	N	-2	0	6																	
Asp	D	-2	-2	1	6																
Cys	C	0	-3	-3	-3	9															
Gln	Q	-1	1	0	0	-3	5														
Glu	E	-1	0	0	2	-4	2	5													
Gly	G	0	-2	0	-1	-3	-2	-2	6												
His	H	-2	0	1	-1	-3	0	0	-2	8											
Ile	I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4
		Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V

$$S = \sum_{i=1}^{L} s_{r_{1,i} r_{2,i}}$$

```
R  L  A  S  V  E  T  D  M  P  L  T  L  R  Q  H
.  |  .  |  :  :  |  .  :  .  .  .  |  .  .  |
T  L  T  S  L  Q  T  T  L  K  A  H  L  G  T  H
-1 +4 +0 +4 +1 +2 +5 -1 +2 -1 -1 -2 +4 -2 -1 +8 = 21
```

Exemple d'alignement retourné par BLAST

```
>gi|16127996|ref|NP_414543.1| bifunctional: aspartokinase I
          (N-terminal); homoserine dehydrogenase I (C-terminal)
          [Escherichia coli K12]
         Length = 820

 Score =  344 bits (882), Expect = 2e-95
 Identities = 247/821 (30%), Positives = 410/821 (49%), Gaps = 44/821 (5%)

Query: 16  KFGGSSLADVKCYLRVAGIMAEYSQPDDMM-VVSAAGSTTNQLINWLKLSQTDRLSAHQV 74
           KFGG+S+A+ + +LRVA I+   ++   +  V+SA    TN L+  ++ + + + +   +
Sbjct: 5   KFGGTSVANAERFLRVADILESNARQGQVATVLSAPAKITNHLVAMIEKTISGQDALPNI 64

Query: 75  QQTLRRYQCDLISGLLPAEEADSL--ISAFVSDLERLAALLDSGIN------DAVYAEVV 126
               R +  +L++GL   A+    L  +  FV         +  GI+      D++ A ++
Sbjct: 65  SDAERIF-AELLTGLAAAQPGFPLAQLKTFVDQEFAQIKHVLHGISLLGQCPDSINAALI 123

Query: 127 GHGEVWSARLMSAVLNQQGLPAAWLDAREFLRAER---AAQPQVDEGLSYPLLQQLLVQH 183
             GE  S  +M+ VL  +G     +D  E L A      +   + E        ++   H
Sbjct: 124 CRGEKMSIAIMAGVLEARGHNVTVIDPVEKLLAVGHYLESTVDIAESTRRIAASRIPADH 183

Query: 184 PGKRLVVTGFISRNNAGETVLLGRNGSDYSATQIGALAGVSRVTIWSDVAGVYSADPRKV 243
               +++ GF + N  GE V+LGRNGSDYSA  + A        IW+DV GVY+ DPR+V
Sbjct: 184 ---MVLMAGFTAGNEKGELVVLGRNGSDYSAAVLAACLRADCCEIWTDVDGVYTCDPRQV 240

Query: 244 KDACLLPLLRLDEASELARLAAPVLHARTLQPVSGSEIDLQLRCSYTPDQ-----GSTRI 298
            DA LL  +   EA EL+   A VLH RT+ P++  +I   ++ +  P       G++R
Sbjct: 241 PDARLLKSMSYQEAMELSYFGAKVLHPRTITPIAQFQIPCLIKNTGNPQAPGTLIGASRD 300

Query: 299 ERVLASGTGARIVTSHDDVCLIEFQVPASQDFKLAHKEIDQILKRAQVRPLAVGVHNDRQ 358
           E  L      + +++ +++ +     P  +        +   + RA++  + +   +
Sbjct: 301 EDELP----VKGISNLNNMAMFSVSGPGMKGMVGMAARVFAAMSRARISVVLITQSSSEY 356

Query: 359 LLQFCYTSEVADSALKILDEA-------GLPGELRLRQGLALVAMVGAGVTRNPLHCHRF 411
            + FC        A + + E        GL   L + + LA++++VG G+        +F
Sbjct: 357 SISFCVPQSDCVRAERAMQEEFYLELKEGLLEPLAVTERLAIISVVGDGMRTLRGISAKF 416

Query: 412 WQQLKGQPVEFTW--QSDDGISLVAVLRTGPTESLIQGLHQSVFRAEKRIGLVLFGKGNI 469
           +  L    +      Q     S+  V+      + ++  HQ +F  ++ I + + G G +
Sbjct: 417 FAALARANINIVAIAQGSSERSISVVVNNDDATTGVRVTHQMLFNTDQVIEVFVIGVGGV 476

Query: 470 GSRWLELFAREQSTLSARTGFEFVLAGVVDSRRSLLSYDGLDASRALAEFNDEAVEQDEE 529
```

- A partir du score brut et du résultat de l'alignement, BLAST dérive une série de scores qui quantifient la qualité de l'alignement.
- Example
  - Score brut 882
  - Bit score 244
  - Expect 2e-95
  - Identities 247
  - % identities 30%
  - Positives 410
  - % positives 49%
  - Gaps 44
  - % gaps 5%
- Questions
  - Comment interpréter ces scores ?
  - Quel(s) scores peut-on considérer comme pertinent(s) ?
  - A partir de quel(s) seuil(s) l'alignement est-il significatif ?

Note pour les étudiants de biologie + bioingénieurs

- Le détail des statistiques d'alignement (les formules) ne fait pas partie de la matière d'examen.
- Cependant, je vous suggère de lire attentivement les commentaires de ces statistiques.
- **Ce que vous devez savoir**
  - **Comment calculer le score brut d'un alignement (avec et sans gap) ?**
  - **Comment interpréter la e-valeur (diapos suivantes) ?**
- Dans les diapos qui suivent, vous pouvez ignorer les formules de calcul de ces probabilités, mais vous devez ensuite savoir quels critères sont pris en compte pour évaluer la significativité d'un alignement.

P-valeur d'un segment aligné (MSP) et score en bits

- A partir du score brut (*S*), on peut calculer la ***p-valeur***, qui représente **la probabilité d'obtenir par hasard un score au moins égal à *S***.
  - Interprétation de la P-valeur: estimation du risque de faux-positif.
  - Karlin and Altschul (1990) définissent les statistiques de calcul de la p-valeur d'un segment aligné (matching segment pair, MSP).
  - La p-valeur suit une distribution expontentielle à deux paramètres: ***lambda*** et ***K***.
    - Ces deux paramètres dépendent de la matrice de substitution.
    - On peut les calculer de façon exacte uniquement pour les alignements sans gaps.
    - Pour les alignements avec gaps, Altschul et al (1997) proposent de les estimer de façon empiriques (alignements de séquences non-apparentées).

$$Pval_S^{MSP} = P(X \geq S) = Ke^{-\lambda S}$$

Score en bits

- **Score bit d'un alignement**
  - Karlin and Altschul (1990) proposent de convertir la p-valeur en « ***bit score » (S')***.
  - Le score en bits (*S'*) est plus interprétable que le score brut (*S*), car la p-valeur peut être directement retrouvée à partir du score de bits.
  - La conversion de bits en p-valeurs repose sur la même formule, indépendamment de la matrice de substitution utilisée.

$$Pval_S^{MSP} = P(X \geq S) = Ke^{-\lambda S}$$

$$S' = \frac{\lambda S - \ln(K)}{\ln(2)}$$

$$Pval_S^{MSP} = Ke^{-\lambda S} = Ke^{-\ln(2)S' + \ln(K)} = 2^{-S'}$$

Statistiques d'alignements – la e-valeur (expect)

- Imaginons qu'on aligne deux fragments de séquences choisis au hasard. Le score sera généralement faible.
- Cependant, si on répète cette opération des milliards de fois, certains scores élevés pourraient sortir occasionnellement, par hasard.
- Lors d'une recherche de similarité, chaque position de la séquence requête est comparé à chaque position de la base de données.
- FastA et BLAST estiment, pour chaque score, le nombre de correspondances attendues au hasard, étant donné la taille de la base de données. CE nombre est appelé la ***e-valeur*** (« ***expect*** » sur la page de résultats de BLAST).
  - La e-valeur est le produit de la p-valeur nominale (le risque de faux positifs pour une seule comparaison de deux positions) par la taille de l'espace de recherche.
  - Pour une requête de taille *m* (par exemple 300aa), et une base données de taille *n* (par exemple 12×10^9), l'espace de recherche est donc
    - $N=nm=300*12*10^9=3.6*10^{12}$
- Pour un score S donné, la e-valeur augmente donc avec la taille de données.

$$N = n \cdot m$$

$$\begin{aligned} E &= m \cdot n \cdot Pval \\ &= m \cdot n \cdot K \cdot e^{-\lambda S} \\ &= N \cdot K \cdot e^{-\lambda S} \\ &= N / 2^{S'} \end{aligned}$$

Choix du seuil sur la e-valeur

- Plus la e-valeur est faible, plus l'alignement est significatif.
- Des e-valeurs élevées (>1)
  - indiquent donc qu'un alignement a de fortes chances de résulter du hasard, et ne devrait pas être considéré comme pertinent (il ne correspond vraisemblablement pas à une homologie).
- Une e-valeur très basse (ex: 1e-21)
  - indique que l'alignement n'a quasiment aucune chance de résulter du hasard. Il est dès lors vraisemblable qu'il résulte d'une origine ancestrale commune entre les deux séquences alignées. Dans ce cas, on admet donc l'hypothèse d'homologie.
- Un paramètre essentiel pour BLAST et FastA est le ***seuil sur la e-valeur*** (***expect threshold***).
- Attention
  - Sur le serveur BLAST du NCBI, la valeur seuil par défaut vaut 10.
  - Ceci signifie que chaque requête pourrait retourner 10 alignements par hasard.
  - Si on se fie à ce seuil, on doit s'attendre à ***~10 faux positifs par requête***.
  - Il est donc recommandé de diminuer le seuil d'e-valeur (par exemple à 0.001), pour obtenir des résultats significatifs.

Distribution de probabilité des scores d'alignement

- Quand on effectue une recherche de similarités, la distribution de scores suit une distribution très différente de la normale.
- Il s'agit d'une ***distribution de valeurs extrêmes***.
- Cette distribution est asymétrique, et ne doit donc en aucun cas être modélisée par une distribution gaussienne.

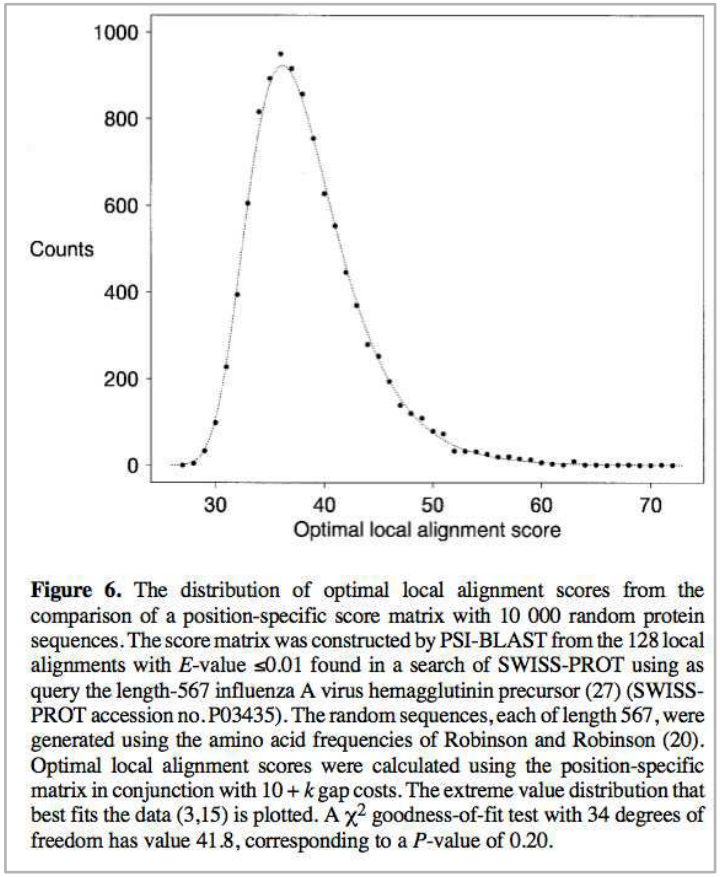
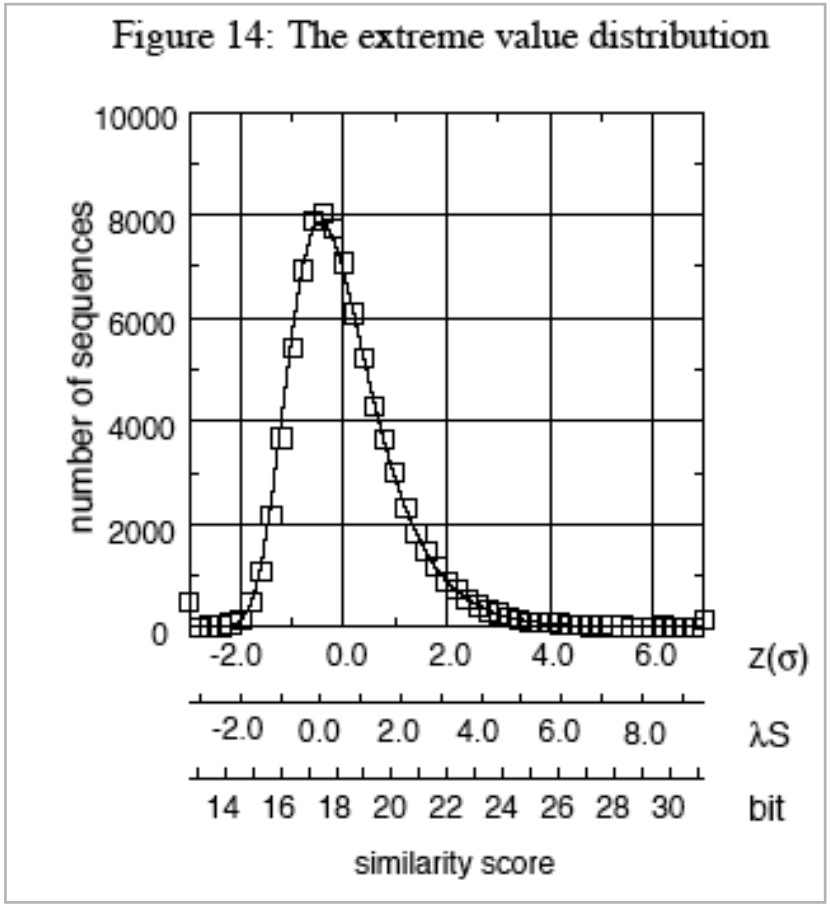


Figure 6. The distribution of optimal local alignment scores from the comparison of a position-specific score matrix with 10 000 random protein sequences. The score matrix was constructed by PSI-BLAST from the 128 local alignments with *E*-value ≤0.01 found in a search of SWISS-PROT using as query the length-567 influenza A virus hemagglutinin precursor (27) (SWISS-PROT accession no. P03435). The random sequences, each of length 567, were generated using the amino acid frequencies of Robinson and Robinson (20). Optimal local alignment scores were calculated using the position-specific matrix in conjunction with $10 + k$ gap costs. The extreme value distribution that best fits the data (3,15) is plotted. A χ^2 goodness-of-fit test with 34 degrees of freedom has value 41.8, corresponding to a *P*-value of 0.20.

Source: Altschul et al. (1997).Nucl Acids Res 25: 3389–3402



Source: W.P.Pearson (2000). Protein sequence comparison and Protein evolution. ISMB Tutorial.

Statistiques d'alignement – p-valeur à échelle de la base de données (~FWER=Family-Wise Error Rate)

- A partir de la e-valeur (E), on peut estimer la probabilité d'observer au hasard ***au moins X alignements*** qui passent le seuil donnée.
- Il s'agit d'une simple application de la distribution de Poisson: calculer la probabilité d'observer *X* succès d'un événement attendu *E* fois (E est utilisé ici comme estimation du paramètre lambda de la Poisson).

$$P(X \geq x) = \sum_{i=x}^{N} \frac{e^{-E} E^{i}}{i!} = 1 - \sum_{i=0}^{x-1} \frac{e^{-E} E^{i}}{i!}$$

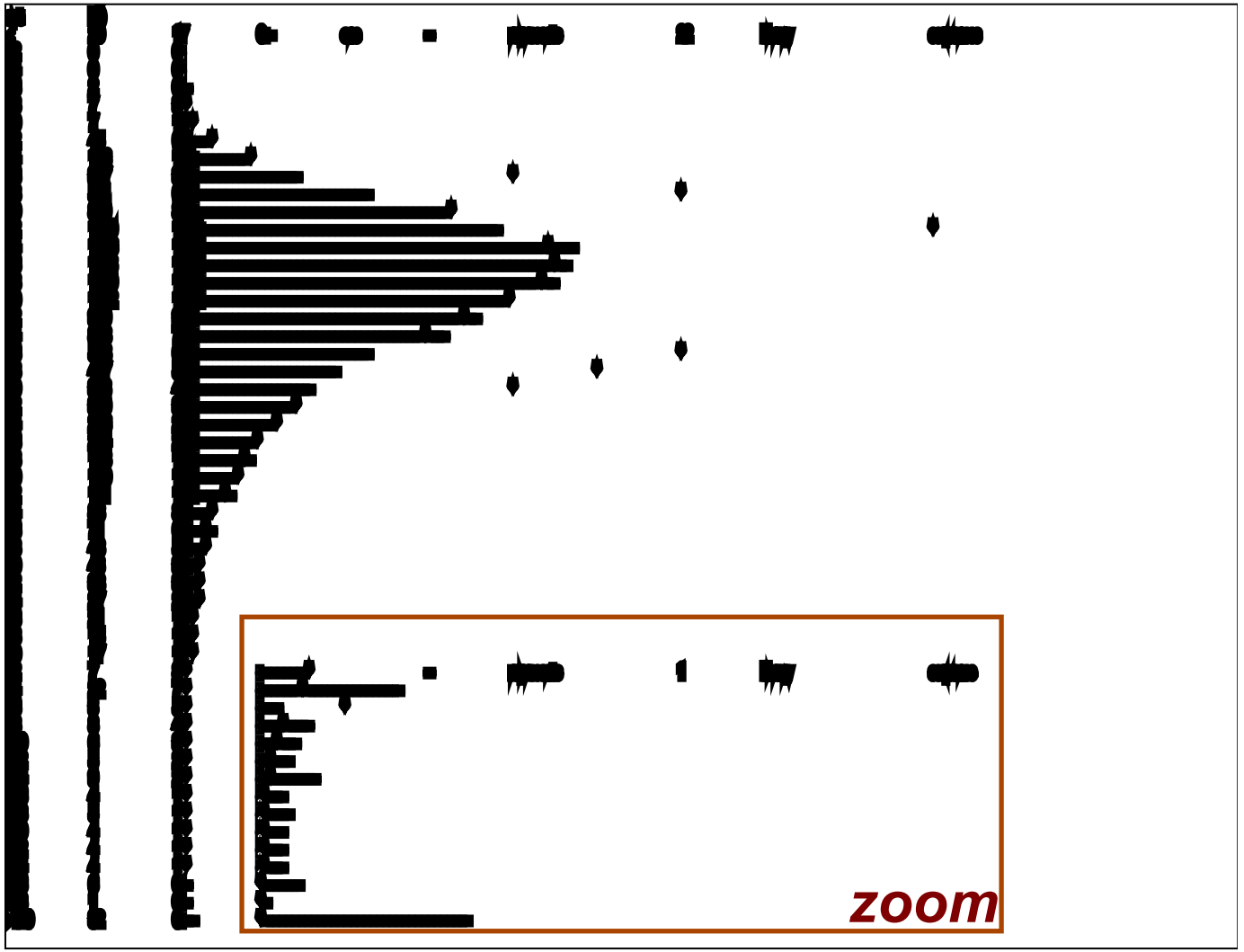
- Cas particulier: probabilité d'observer au moins un résultat par hasard
  - P(X>=1).
- Cette probabilité est généralement appelée ***Family-Wise Error Rate (FWER)***.
- Dans le cas de recherches de similarités, on peut l'appeler ***P-valeur à l'échelle de la base de données***.
- Cette p-valeur représente la probabilité de trouver au moins un alignement par chance dans l'ensemble de la base de données, étant donné le seuil de e-valeur choisi.

$$\begin{aligned} Pval^{DB} &= P(X \geq 1) \\ &= 1 - P(X = 0) \\ &= 1 - \frac{e^{-E} E^{0}}{0!} \\ &= 1 - e^{-E} \end{aligned}$$

Interprétation des résultats d'une recherche par similarité

Distribution de score

- L'histogramme indique le nombre de séquences trouvées dans une base de données pour chaque valeur de score.
- Pour les scores >= 92, on observe un très petit nombre de résultats.
- L'encadré indique la queue de l'histogramme avec une échelle plus fine.
- Les astérisques indiquent les nombres de hits attendus au hasard (E-valeur).



FastA output from Pearson (2000)

Exemple de résultat de BLAST

```
BLASTP 2.2.6 [Apr-09-2003]


Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database search
programs",  Nucleic Acids Res. 25:3389-3402.

Query= metL gi|16131778|ref|NP_418375.1| aspartokinase II and
homoserine dehydrogenase II; bifunctional: aspartokinase II
(N-terminal); homoserine dehydrogenase II (C-terminal) [Escherichia
coli K12]
         (810 letters)

Database: /Users/jvanheld/rsa-
tools/data/genomes/Escherichia_coli_K12/genome/NC_000913.faa
           4242 sequences; 1,351,322 total letters

Searching.........done

                                                                 Score    E
Sequences producing significant alignments:                      (bits) Value

gi|16131778|ref|NP_418375.1| aspartokinase II and homoserine deh...  1596   0.0
gi|16127996|ref|NP_414543.1| bifunctional: aspartokinase I (N-te...   344   2e-95
gi|16131850|ref|NP_418448.1| aspartokinase III, lysine sensitive...   122   7e-29
gi|16128228|ref|NP_414777.1| gamma-glutamate kinase [Escherichia...    31   0.28

>gi|16131778|ref|NP_418375.1| aspartokinase II and homoserine
           dehydrogenase II; bifunctional: aspartokinase II
           (N-terminal); homoserine dehydrogenase II (C-terminal)
           [Escherichia coli K12]
          Length = 810

 Score = 1596 bits (4132), Expect = 0.0
 Identities = 810/810 (100%), Positives = 810/810 (100%)

Query: 1   MSVIAQAGAKGRQLHKFGGSSLADVKCYLRVAGIMAEYSQPDDMMVVSAAGSTTNQLINW 60
```

- The text shows the result of a BLAST search,
- Query: the *E.coli* protein MetL, a bifunctional enzyme combining aspartokinase and homoserine dehydrogenase activities.
- Database: all proteins from *Escherichia coli K12.*
- The BLAST result file starts with a summary of
  - the parameters used for the search
  - The matching sequences and the score of each match.

BLAST result - first match

```
>gi|16131778|ref|NP_418375.1| aspartokinase II and homoserine
           dehydrogenase II; bifunctional: aspartokinase II
           (N-terminal); homoserine dehydrogenase II (C-terminal)
           [Escherichia coli K12]
          Length = 810

 Score = 1596 bits (4132), Expect = 0.0
 Identities = 810/810 (100%), Positives = 810/810 (100%)

Query: 1   MSVIAQAGAKGRQLHKFGGSSLADVKCYLRVAGIMAEYSQPDDMMVVSAAGSTTNQLINW 60
           MSVIAQAGAKGRQLHKFGGSSLADVKCYLRVAGIMAEYSQPDDMMVVSAAGSTTNQLINW
Sbjct: 1   MSVIAQAGAKGRQLHKFGGSSLADVKCYLRVAGIMAEYSQPDDMMVVSAAGSTTNQLINW 60

Query: 61  LKLSQTDRLSAHQVQQTLRRYQCDLISGLLPAEEADSLISAFVSDLERLAALLDSGINDA 120
           LKLSQTDRLSAHQVQQTLRRYQCDLISGLLPAEEADSLISAFVSDLERLAALLDSGINDA
Sbjct: 61  LKLSQTDRLSAHQVQQTLRRYQCDLISGLLPAEEADSLISAFVSDLERLAALLDSGINDA 120

Query: 121 VYAEVVGHGEVWSARLMSAVLNQQGLPAAWLDAREFLRAERAAQPQVDEGLSYPLLQQLL 180
           VYAEVVGHGEVWSARLMSAVLNQQGLPAAWLDAREFLRAERAAQPQVDEGLSYPLLQQLL
Sbjct: 121 VYAEVVGHGEVWSARLMSAVLNQQGLPAAWLDAREFLRAERAAQPQVDEGLSYPLLQQLL 180

Query: 181 VQHPGKRLVVTGFISRNNAGETVLLGRNGSDYSATQIGALAGVSRVTIWSDVAGVYSADP 240
           VQHPGKRLVVTGFISRNNAGETVLLGRNGSDYSATQIGALAGVSRVTIWSDVAGVYSADP
Sbjct: 181 VQHPGKRLVVTGFISRNNAGETVLLGRNGSDYSATQIGALAGVSRVTIWSDVAGVYSADP 240

Query: 241 RKVKDACLLPLLRLDEASELARLAAPVLHARTLQPVSGSEIDLQLRCSYTPDQGSTRIER 300
           RKVKDACLLPLLRLDEASELARLAAPVLHARTLQPVSGSEIDLQLRCSYTPDQGSTRIER
Sbjct: 241 RKVKDACLLPLLRLDEASELARLAAPVLHARTLQPVSGSEIDLQLRCSYTPDQGSTRIER 300

Query: 301 VLASGTGARIVTSHDDVCLIEFQVPASQDFKLAHKEIDQILKRAQVRPLAVGVHNDRQLL 360
           VLASGTGARIVTSHDDVCLIEFQVPASQDFKLAHKEIDQILKRAQVRPLAVGVHNDRQLL
Sbjct: 301 VLASGTGARIVTSHDDVCLIEFQVPASQDFKLAHKEIDQILKRAQVRPLAVGVHNDRQLL 360

Query: 361 QFCYTSEVADSALKILDEAGLPGELRLRQGLALVAMVGAGVTRNPLHCHRFWQQLKGQPV 420
           QFCYTSEVADSALKILDEAGLPGELRLRQGLALVAMVGAGVTRNPLHCHRFWQQLKGQPV
Sbjct: 361 QFCYTSEVADSALKILDEAGLPGELRLRQGLALVAMVGAGVTRNPLHCHRFWQQLKGQPV 420

Query: 421 EFTWQSDDGISLVAVLPTGPTESLLQGLHQSVERAEKRIGLVLEGKGNIGSRWLELEARE 480
```

- The first match is the query sequence itself (metL). This is not surprising since we scanned the set of all E.coli proteins with a protein from E.coli.
- The E-value (0) means that, with this level of similarity; one would expect 0 false positive by chance.

BLAST result - second match

```
>gi|16127996|ref|NP_414543.1| bifunctional: aspartokinase I
          (N-terminal); homoserine dehydrogenase I (C-terminal)
          [Escherichia coli K12]
         Length = 820

 Score =  344 bits (882), Expect = 2e-95
 Identities = 247/821 (30%), Positives = 410/821 (49%), Gaps = 44/821 (5%)

Query: 16  KFGGSSLADVKCYLRVAGIMAEYSQPDDMM-VVSAAGSTTNQLINWLKLSQTDRLSAHQV 74
           KFGG+S+A+ + +LRVA I+   ++   +  V+SA    TN L+  ++ + + + +   +
Sbjct: 5   KFGGTSVANAERFLRVADILESNARQGQVATVLSAPAKITNHLVAMIEKTISGQDALPNI 64

Query: 75  QQTLRRYQCDLISGLLPAEEADSL--ISAFVSDLERLAALLDSGIN------DAVYAEVV 126
               R +  +L++GL  A+    L  +  FV         +  GI+      D++ A ++
Sbjct: 65  SDAERIF-AELLTGLAAAQPGFPLAQLKTFVDQEFAQIKHVLHGISLLGQCPDSINAALI 123

Query: 127 GHGEVWSARLMSAVLNQQGLPAAWLDAREFLRAER---AAQPQVDEGLSYPLLQQLLVQH 183
             GE  S  +M+ VL  +G     +D  E L A      +   + E         ++   H
Sbjct: 124 CRGEKMSIAIMAGVLEARGHNVTVIDPVEKLLAVGHYLESTVDIAESTRRIAASRIPADH 183

Query: 184 PGKRLVVTGFISRNNAGETVLLGRNGSDYSATQIGALAGVSRVTIWSDVAGVYSADPRKV 243
              +++ GF + N  GE V+LGRNGSDYSA  + A        IW+DV GVY+ DPR+V
Sbjct: 184 ---MVLMAGFTAGNEKGELVVLGRNGSDYSAAVLAACLRADCCEIWTDVDGVYTCDPRQV 240

Query: 244 KDACLLPLLRLDEASELARLAAPVLHARTLQPVSGSEIDLQLRCSYTPDQ-----GSTRI 298
            DA LL  +   EA EL+   A VLH RT+ P++  +I   ++ +  P       G++R
Sbjct: 241 PDARLLKSMSYQEAMELSYFGAKVLHPRTITPIAQFQIPCLIKNTGNPQAPGTLIGASRD 300

Query: 299 ERVLASGTGARIVTSHDDVCLIEFQVPASQDFKLAHKEIDQILKRAQVRPLAVGVHNDRQ 358
           E  L      + +++ +++ +     P  +        +   + RA++  + +   +
Sbjct: 301 EDELP----VKGISNLNNMAMFSVSGPGMKGMVGMAARVFAAMSRARISVVLITQSSSEY 356

Query: 359 LLQFCYTSEVADSALKILDEA-------GLPGELRLRQGLALVAMVGAGVTRNPLHCHRF 411
            + FC        A + + E        GL   L + + LA++++VG G+        +F
Sbjct: 357 SISFCVPQSDCVRAERAMQEEFYLELKEGLLEPLAVTERLAIISVVGDGMRTLRGISAKF 416

Query: 412 WQQLKGQPVEFTW--QSDDGISLVAVLRTGPTESLIQGLHQSVFRAEKRIGLVLFGKGNI 469
```

- The second match is another bifunctional protein, product of the gene thrA.
- This protein contains the same two domains as metA (aspartokinase and homoserine dehydrogenase).
- The alignment covers almost the complete sequences (820 aa), with 30% identities and 49% similarity.
- The E-value is very low (2e-95), indicating that thrA and metL are likely to be true homologs.

BLAST result - third match

```
>gi|16131850|ref|NP_418448.1| aspartokinase III, lysine sensitive;
           aspartokinase III, lysine-sensitive [Escherichia coli
           K12]
          Length = 449

 Score =  122 bits (307), Expect = 7e-29
 Identities = 121/452 (26%), Positives = 194/452 (42%), Gaps = 25/452 (5%)

Query: 16  KFGGSSLADVKCYLRVAGIMAEYSQPDDMMVVSAAGSTTNQLINWLK-LSQTDRLSAHQV 74
           KFGG+S+AD     R A I+   +    ++V+SA+   TN L+   + L   +R    +
Sbjct: 8   KFGGTSVADFDAMNRSADIVLSDANVR-LVVLSASAGITNLLVALAEGLEPGERF---EK 63

Query: 75  QQTLRRYQCDLISGLLPAEEADSLISAFVSDLERLAALLDSGINDAVYAEVVGHGEVWSA 134
              +R  Q  ++  L         I   + ++  LA       + A+  E+V HGE+ S
Sbjct: 64  LDAIRNIQFAILERLRYPNVIREEIERLLENITVLAEAAALATSPALTDELVSHGELMST 123

Query: 135 RLMSAVLNQQGLPAAWLDAREFLRA-ERAAQPQVDEGLSYPLLQQLLVQHPGKRLVVT-G 192
            L   +L ++ + A W D R+ +R  +R  + + D      L    L+    + LV+T G
Sbjct: 124 LLFVEILRERDVQAQWFDVRKVMRTNDRFGRAEPDIAALAELAALQLLPRLNEGLVITQG 183

Query: 193 FISRNNAGETVLLGRNGSDYSATQIGALAGVSRVTIWSDVAGVYSADPRKVKDACLLPLL 252
           FI   N G T  LGR GSDY+A  +      SRV IW+DV G+Y+ DPR V  A  +  +
Sbjct: 184 FIGSENKGRTTTLGRGGSDYTAALLAEALHASRVDIWTDVPGIYTTDPRVVSAAKRIDEI 243

Query: 253 RLDEASELARLAAPVLHARTLQPVSGSEIDLQLRCSYTPDQGSTRI---------ERVLA 303
              EA+E+A   A VLH  TL P   S+I + +  S  P  G T +          R LA
Sbjct: 244 AFAEAAEMATFGAKVLHPATLLPAVRSDIPVFVGSSKDPRAGGTLVCNKTENPPLFRALA 303

Query: 304 SGTGARIVTSHDDVCLIEFQVPASQDFKLAHKEIDQILKRAQVRPLAVGVHNDRQLLQFC 363
                ++T H    L      A     LA   I   L       +A+        L
Sbjct: 304 LRRNQTLLTLHSLNMLHSRGFLAEVFGILARHNISVDLITTSEVSVAL-------TLDTT 356

Query: 364 YTSEVADSAL--KILDEAGLPGELRLRQGLALVAMVGAGVTRNPLHCHRFWQQLKGQPVE 421
            ++   D+ L   +L E      + + +GLALVA++G  +++        +  L+   +
Sbjct: 357 GSTSTGDTLLTQSLLMELSALCRVEVEEGLALVALIGNDLSKACGVGKEVFGVLEPFNIR 416

Query: 422 FTWQSDDGISLVAVLRTGPTESLIQGLHQSVF 453
```

- The third match is the product of the gene lysC: aspartokinase III.
- This protein contains the aspartokinase domain, but not the homoserine dehydrogenase.
- Consequently, the alignment only extends over the first half of the query protein (453aa).
- On this segment, there is a good level of identity (26%) and similarity (42%).
- The E-value is very low (7e–29), indicating that the two domains are likely to be true homologs.
-

BLAST result - fourth match

```
>gi|16128228|ref|NP_414777.1| gamma-glutamate kinase [Escherichia
           coli K12]
          Length = 367

 Score = 31.2 bits (69), Expect = 0.28
 Identities = 17/56 (30%), Positives = 29/56 (51%)

Query: 194 ISRNNAGETVLLGRNGSDYSATQIGALAGVSRVTIWSDVAGVYSADPRKVKDACLL 249
           I+ N+A  T  +    +D  +     LAG  ++ + +D  G+Y+ADPR    A L+
Sbjct: 133 INENDAVATAEIKVGDNDNLSALAAILAGADKLLLLTDQKGLYTADPRSNPQAELI 188
```

- The fourth match is a gamma-glutamate kinase, product of proB.
- The match has the same level of identity (30%) and similarity (51%) as the second match (thrA).
- However, this match only extends over 56aa, whereas the alignment between thrA and metL extends over 821aa.
- The significance of the match is thus much lower: the E-value is quite high (0.28) suggesting that the similarity could be an artefact.
- **This clearly illustrates the fact that the important parameter to evaluate the significance of an alignment is the E-value, not the percentage of similarity !**

BLAST result - summary

```
 Database: /Users/jvanheld/rsa-
  tools/data/genomes/Escherichia_coli_K12/genome/NC_000913.faa
    Posted date:  Sep 8, 2004 12:13 PM
  Number of letters in database: 1,351,322
  Number of sequences in database:  4242

Lambda     K      H
   0.320    0.136    0.397

Gapped
Lambda     K      H
   0.267   0.0410    0.140

Matrix: BLOSUM62
Gap Penalties: Existence: 11, Extension: 1
Number of Hits to DB: 2,199,628
Number of Sequences: 4242
Number of extensions: 96525
Number of successful extensions: 290
Number of sequences better than  1.0: 4
Number of HSP's better than  1.0 without gapping: 4
Number of HSP's successfully gapped in prelim test: 0
Number of HSP's that attempted gapping in prelim test: 279
Number of HSP's gapped (non-prelim): 5
length of query: 810
length of database: 1,351,322
effective HSP length: 92
effective length of query: 718
effective length of database: 961,058
effective search space: 690039644
effective search space used: 690039644
T: 11
A: 40
X1: 16 ( 7.4 bits)
X2: 38 (14.6 bits)
X3: 64 (24.7 bits)
S1: 41 (21.8 bits)
S2: 65 (29.6 bits)
```

- The last part of the BLAST result gives some statistics about the search:
  - Number of hits
  - Number of sequences in the DB
  - …

Critères de significativité: quel(s) score(s) choisir ?

Critères de significativité: quel(s) score(s) choisir ?

- BLAST, Fasta et quelques autres programmes d'alignement retournent une série de scores fournissant des indications complémentaires concernant la qualité de l'alignement:
  - Longueur de l'alignement
  - Score brut somme des scores de substitution + pénalités de gaps
  - Score en bit s bits = log2(p-valeur nominale)
  - Identities Nombre et pourcentage d'identité.
  - Positives Nombre et pourcentage de similarités (scores positifs entre paires de résidus alignés, d'après la matrice de substitution choisie).
  - Gaps Nombre et pourcentage de résidus alignés face à un gap ("-")
  - Eval E-valeur (nombre de faux positifs attendus étant donné le score en bits)
- Quel(s) critère(s) choisir pour décider si un alignement est significatif ou non ?
- Dans certains manuels de bioinformatique, on trouve des critères basés sur le pourcentage d'identité (par exemple "au moins 30% d'identités").
  - Problème: un taux élevé d'identités sur un tout petit alignement (par exemple 50% sur 10 résidus) n'est pas significatif.
- On trouve parfois des critères mixtes ("au moins 30% d'identités sur au moins 100 résidus")
  - Problème: d'où proviennent ces valeurs de seuil? Pourquoi 100 résidus et pas 97 ou 101 ?
  - Extension du problème précédent: le même taux d'identité est plus ou moins significatif selon la longueur de l'alignement.
- Exercices: dans les diapos suivantes, analysez l'ensemble des scores et déterminez, pour chaque alignement, s'il est ou non significatif.

Séquence requête: P04000 OPSR_HUMAN (364 aa)

```
Rhodopsin [Cataglyphis bombycinus] (Length 378)
Score: 122 bits                Expect: 8e-30
Identities: 80/284(28%)        Positives: 137/284(48%)
Gaps: 16/284(5%)

Query  67   SVFTNGLVLAATMKFKKLRHPLNWILVNLAVADLAETVIASTISIVNQVSGYFVLGHPMC  126
            SV  NG+V+      K LR P N +++NLA++D    +  S   ++N     +VLG  +C
Sbjct  68   SVIGNGMVIYIFTTTKSLRTPSNLLVINLAISDFLMMLSMSPAMVINCYYETWVLGPLVC  127

Query  127  VLEGYTVSLCGITGLWSLAIISWERWLVVCKPFGNVRFDAKLAIVGIAFSWIWSAVWTAP  186
             L G T SL G   +W++ +I+++R+ V+ K           A++ I   W +S  WT
Sbjct  128  ELYGLTGSLFGCGSIWTMTMIAFDRYNVIVKGLSAKPMTINGALLRILGIWFFSLGWTIA  187

Query  187  PIFGWSRYWPHGLKTSCGPDVFSGSSYPGVQSYMIVLMVTCCIIPLAIIMLCYLQVWLAI  246
            P+FGW+RY P G  T+CG D  +       +SY++V    C  +PL +I+  Y  +  A+
Sbjct  188  PMFGWNRYVPEGNMTACGTDYLTKDLLS--RSYILVYSFFCYFLPLFLIIYSYFFIIQAV  245

Query  247  RAVAKQQKE------------SESTQKAEKEVTRMVVVMIFAYCVCWGPYTFFACFAAA  293
             A  K  +E                +  AE ++ ++ ++ I  + + W PY     +A
Sbjct  246  AAHEKNMREQAKKMNVASLRSAENQSTSAECKLAKVALMTISLWFMAWTPYLVIN-YAGI  304

Query  294  NPGYAFHPLMAALPAYFAKSATIYNPVIYVFMNRQFRNCILQLF  337
                  +PL     + FAK+  +YNP++Y   + ++R  + Q F
Sbjct  305  FETVKINPLFTIWGSLFAKANAVYNPIVYGISHPKYRAALFQRF  348
```

- Pourcentage d'identités < 30% -> similarité due au hasard ?
- Ce pourcentage est observé sur un alignement long (378 positions alignées)
- **E-valeur (expect) = $8e^{-30}$**
  - Au hasard, on s'attend à observer un aussi bon alignement une fois sur $8e^{-30}$
  - **L'hypothèse d'homologie est donc, de très loin, la plus vraisemblable !!!**

Séquence requête: P01308 Insuline humaine, longueur 110aa

```
serine-type enodpeptidase, putative [Aedes aegypti]
Score: 31.2 bits                Expect: 2.4
Identities:14/45(31%)  Positives: 23/45(51%)
Gaps: 2/45(4%)

Query 25 FVNQHLCGSHLVEALYLVCGERGFFYTP--KTRREAEDLQVGQVE 67
         FV  HLCG  ++   +++   + FF  P  + R +A  L + Q E
Sbjct 49 FVTTHLCGGSILNNFHVITAAQCFFSNPSGRFRVQAGKLTLNQFE 93
```

- Identité 31% -> homologie vraisemblable ?
- Longueur faible -> similarité due au hasard ?
- Gaps 4% -> homologie possible ?
- % de couverture: (67-24-2)/110=39% => similarité due au hasard ?
- Fonction: insuline versus endopeptidase => similarité due au hasard ?
- **E-valeur = 2,4**
  - Au hasard, on s'attend à >1 faux-positifs pour ce type de score, en tenant compte de la longueur de l'alignement
  - **=> similarité vraisemblablement due au hasard**

Séquence requête: P01308 Insuline humaine, longueur 110aa

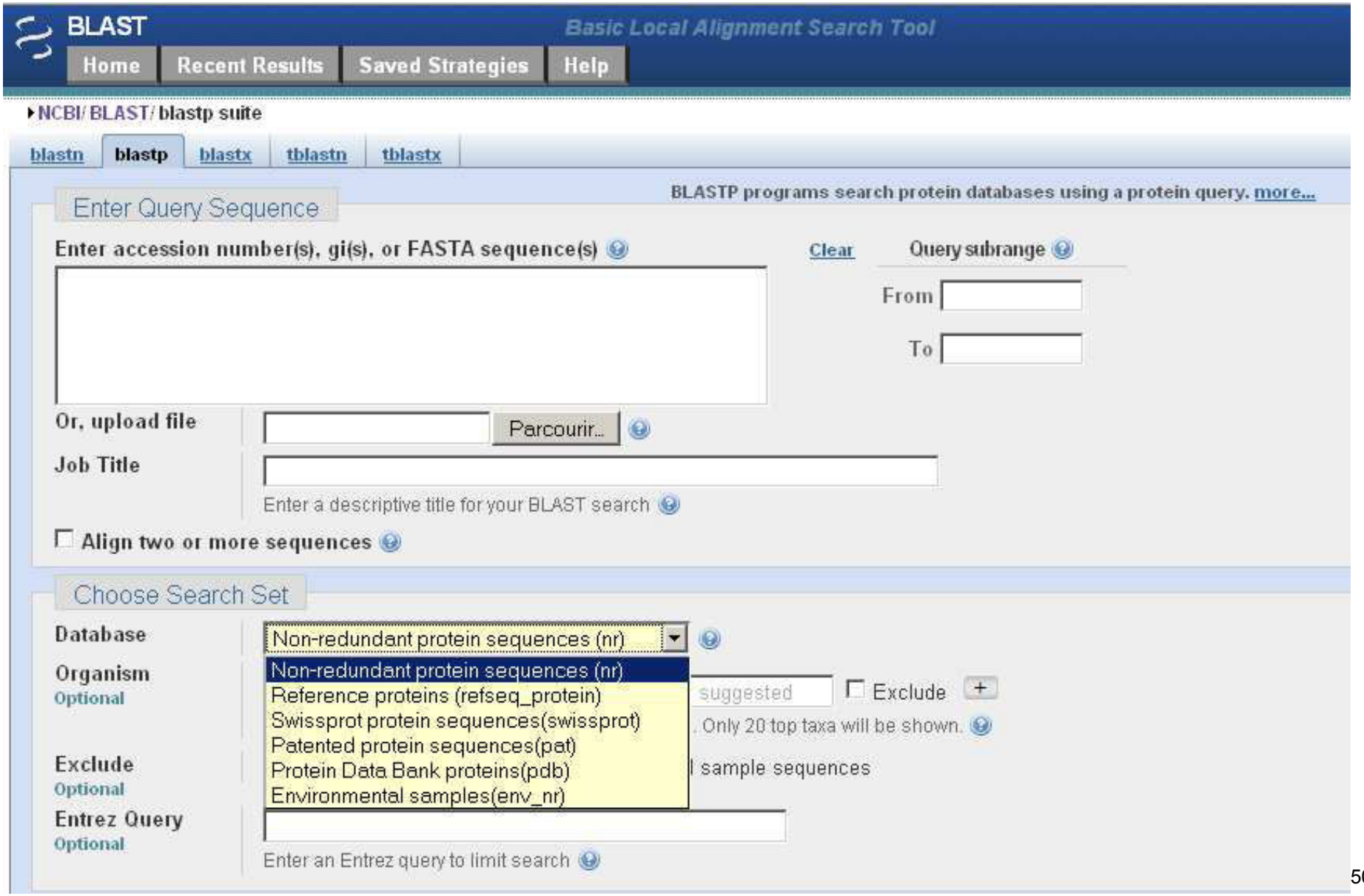
```
Chain A, Enhancing The Activity Of Insulin At
Receptor Edge (Length 21)
Score: 44,7 bits                Expect: 2e-04
Identities: 20/21(95%) Positives: : 20/21(95%)
Gaps: 0/21(0%)

Query 90 GIVEQCCTSICSLYQLENYCN 110
         GIVEQCC SICSLYQLENYCN
Sbjct 1  GIVEQCCHSICSLYQLENYCN 21
```

- Identité 95% -> homologie vraisemblable
- Gaps 0% -> homologie vraisemblable
- % de couverture: 21/21=100% -> homologie vraisemblable
- Fonction: similaire (mais attention, cette fonction a-t-elle été assignée sur base de preuves expérimentales ?)
- **E-valeur = 2e-04**
  - **Nombre de faux-positifs faibles**

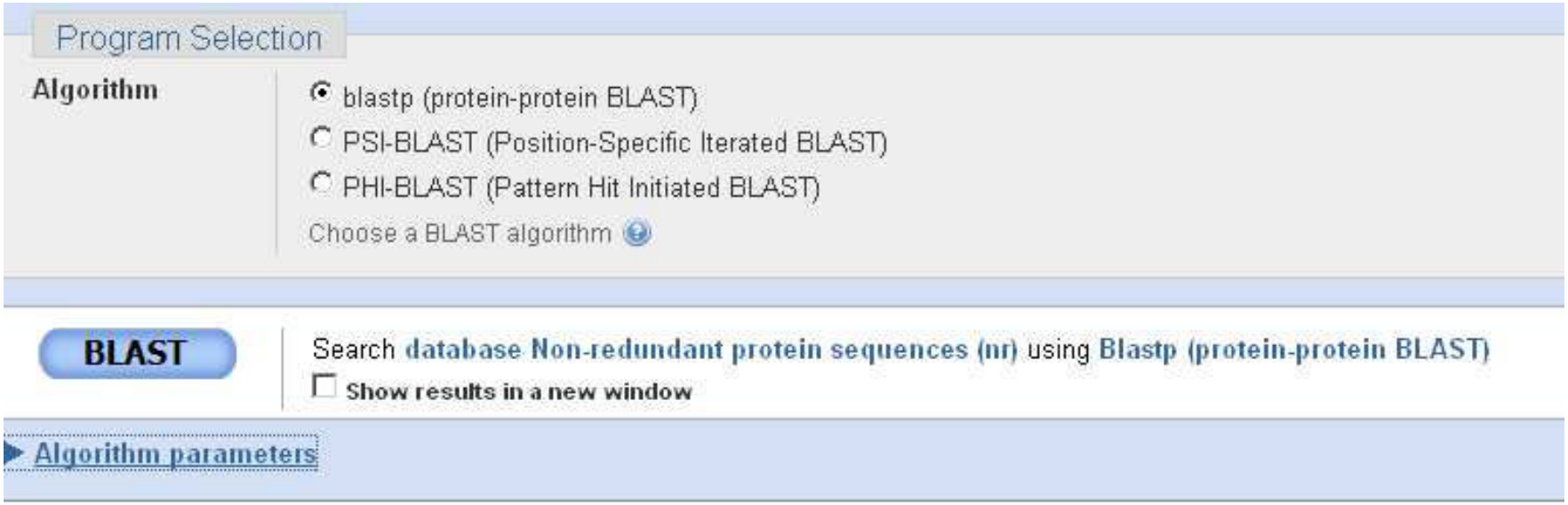
  => homologie vraisemblable

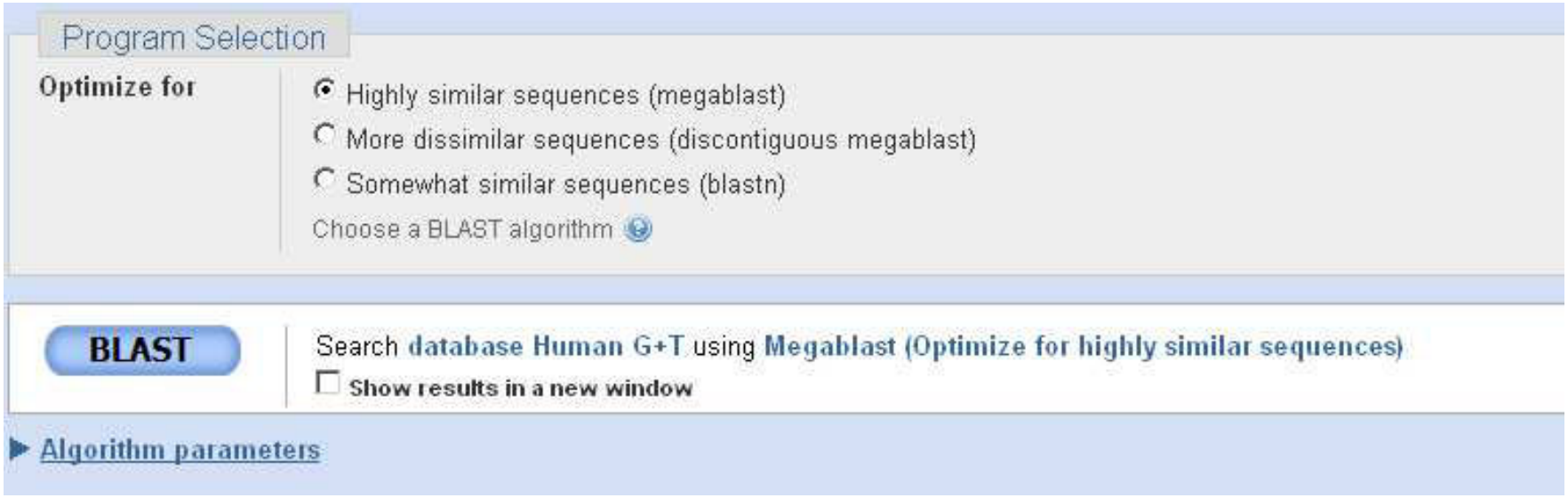
L'outil BLAST du NCBI

BLAST (NCBI)

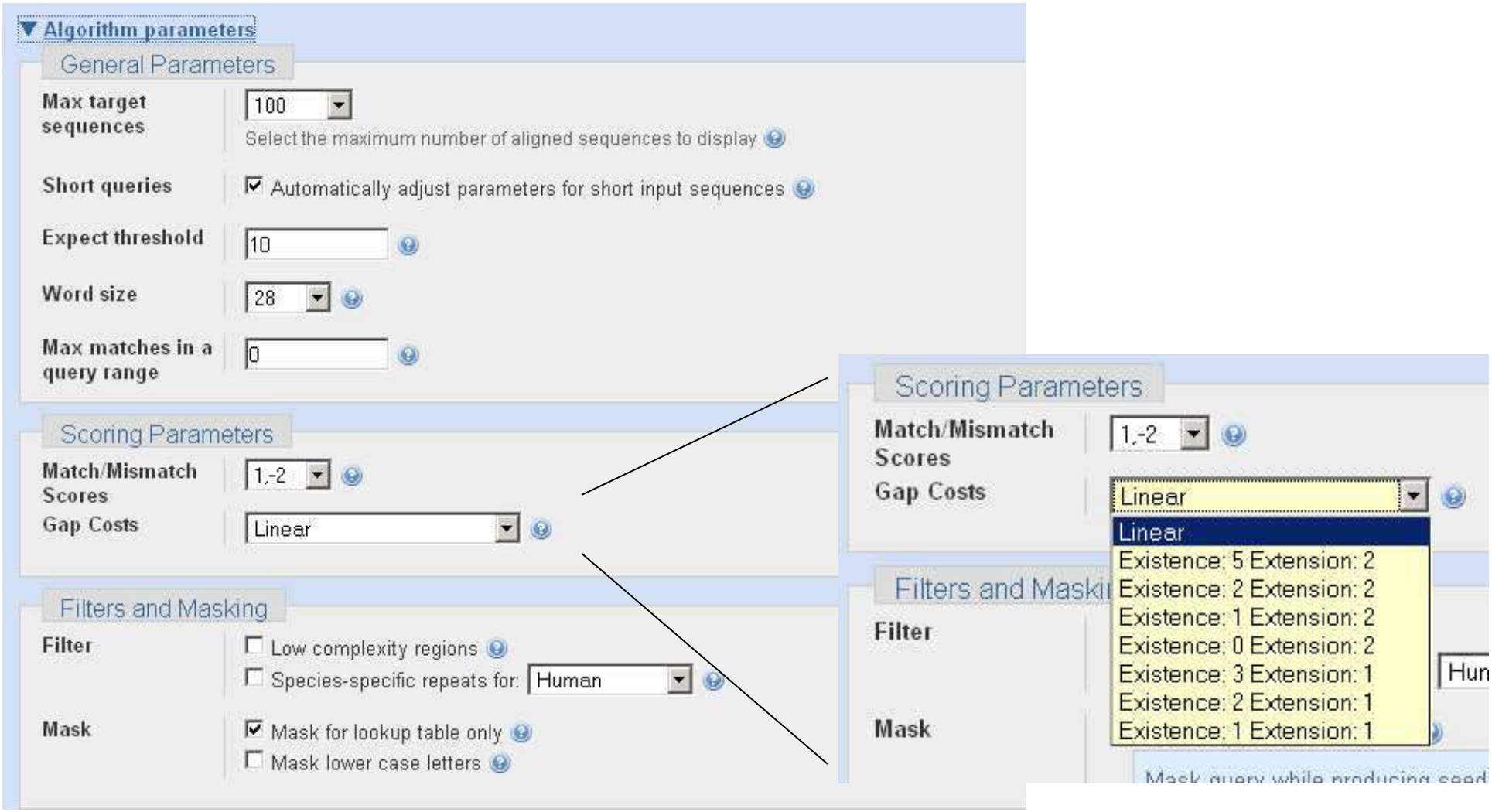

BLAST
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
NCBI/BLAST/blastp suite
blastn
blastp
blastx
tblastn
tblastx
BLASTP programs search protein databases using a protein query. more...
Enter Query Sequence
Enter accession number(s), gi(s), or FASTA sequence(s)
Clear
Query subrange
From
To
Or, upload file
Parcourir...
Job Title
Enter a descriptive title for your BLAST search
Align two or more sequences
Choose Search Set
Database
Non-redundant protein sequences (nr)
Non-redundant protein sequences (nr)
Reference proteins (refseq_protein)
Swissprot protein sequences(swissprot)
Patented protein sequences(pat)
Protein Data Bank proteins(pdb)
Environmental samples(env_nr)
Organism
Optional
suggested
Exclude
Only 20 top taxa will be shown.
Exclude
Optional
sample sequences
Entrez Query
Optional
Enter an Entrez query to limit search

BLAST (NCBI)

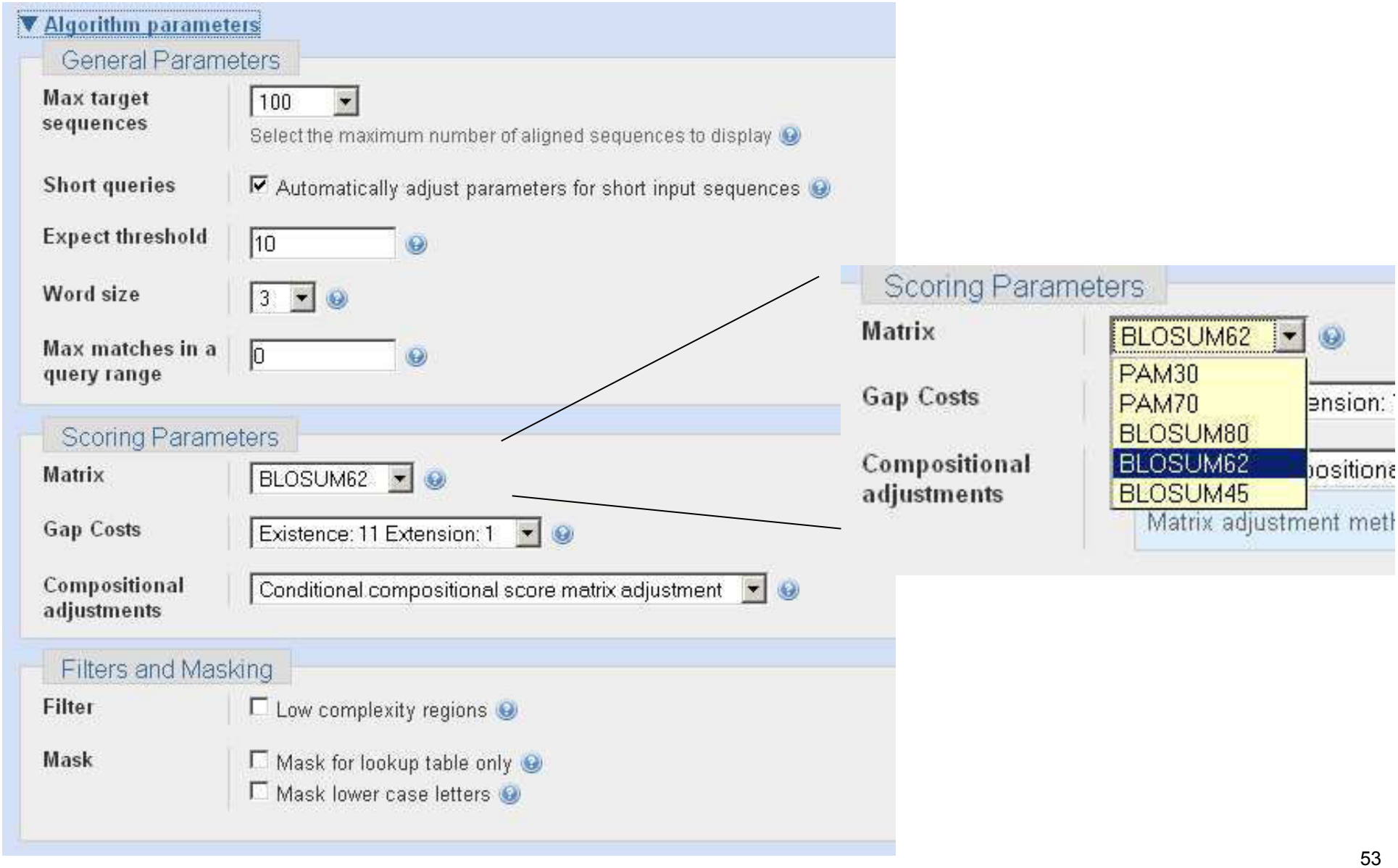

Program Selection
Algorithm
blastp (protein-protein BLAST)
PSI-BLAST (Position-Specific Iterated BLAST)
PHI-BLAST (Pattern Hit Initiated BLAST)
Choose a BLAST algorithm
BLAST
Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)
Show results in a new window
Algorithm parameters

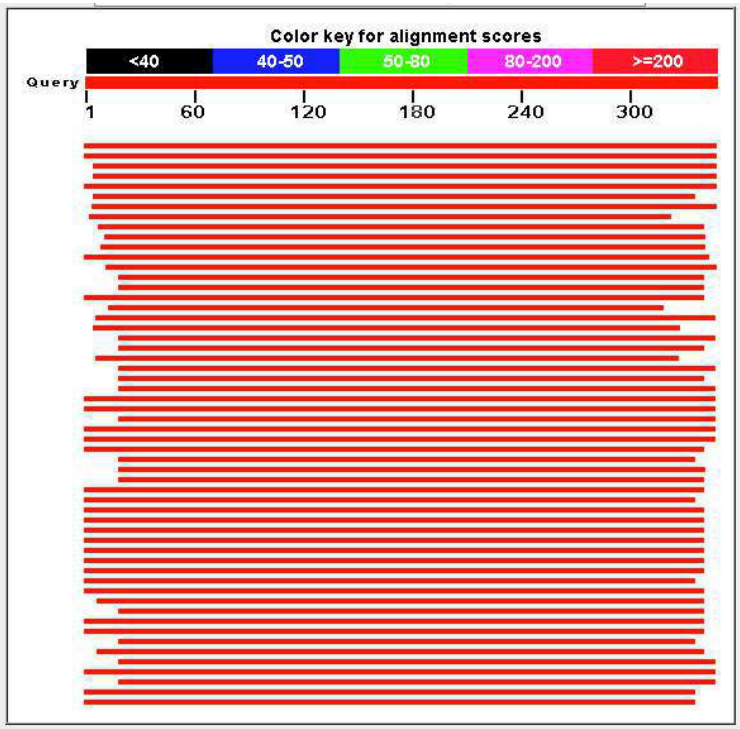

Program Selection
Optimize for
Highly similar sequences (megablast)
More dissimilar sequences (discontiguous megablast)
Somewhat similar sequences (blastn)
Choose a BLAST algorithm
BLAST
Search database Human G+T using Megablast (Optimize for highly similar sequences)
Show results in a new window
Algorithm parameters

BLASTn paramètres



BLASTp paramètres





Descriptions

Legend for links to other resources: U UniGene E GEO G Gene S Structure M Map Viewer PubChem BioAssay

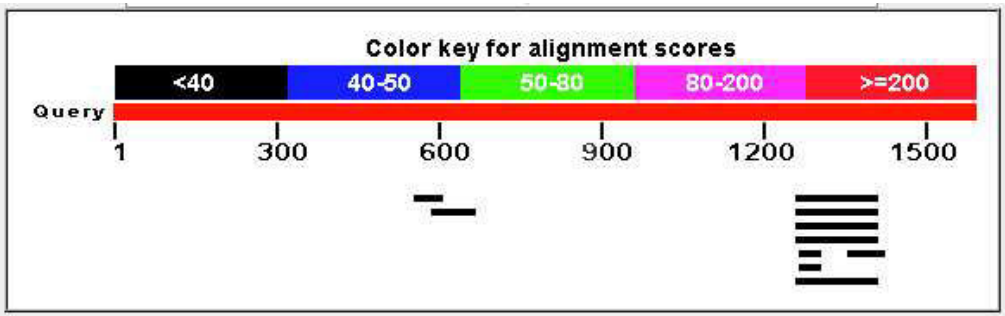
Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
P03999.1	short-wave-sensitive opsin 1 [Homo sapiens] >gi|57114123|ref|NP_0	716	716	100%	0.0	100%	G M
Q13092.1	RecName: Full=Short-wave-sensitive opsin 1; AltName: Full=Blue cor	662	662	100%	0.0	92%	
Q63652.2	RecName: Full=Short-wave-sensitive opsin 1; Short=S opsin; AltNam	627	627	98%	0.0	87%	G M
P51491.1	RecName: Full=Short-wave-sensitive opsin 1; Short=S opsin; AltNam	626	626	98%	0.0	86%	G M
P51490.1	RecName: Full=Short-wave-sensitive opsin 1; AltName: Full=Blue cor	626	626	100%	0.0	86%	G M
O57605.1	RecName: Full=Ultraviolet-sensitive opsin; AltName: Full=Ultraviolet	581	581	95%	0.0	83%	
P28684.1	RecName: Full=Violet-sensitive opsin; AltName: Full=Violet cone opsi	574	574	98%	0.0	80%	G
P51473.1	RecName: Full=Violet-sensitive opsin; AltName: Full=Violet cone opsi	537	537	91%	0.0	78%	G
P87368.1	RecName: Full=Putative violet-sensitive opsin; AltName: Full=KFH-V;	440	440	95%	7e-153	64%	G
Q9W6A9.2	RecName: Full=Opsin-1, short-wave-sensitive 1; Short=Opsin SWS-1	439	439	95%	3e-152	63%	G M
Q90309.1	RecName: Full=Ultraviolet-sensitive opsin; AltName: Full=Ultraviolet	423	423	95%	4e-146	63%	
P32310.1	RecName: Full=Blue-sensitive opsin; AltName: Full=Blue cone photor	351	351	98%	9e-118	49%	
P87365.1	RecName: Full=Blue-sensitive opsin; AltName: Full=Blue cone photor	330	330	96%	3e-109	50%	G
Q8AYM7.1	RecName: Full=Green-sensitive opsin-3; AltName: Full=Green cone p	329	329	92%	3e-109	49%	G M
Q9W6A6.2	RecName: Full=Green-sensitive opsin-4; AltName: Full=Green cone p	329	329	92%	3e-109	48%	G M

Checking expected values with random sequences ("negative control")

Searching a sequence database with a random sequence as query

- Empirical test of the expected value:
  - We generated a random sequence of 1588 aa using the tool random-sequence (http://rsat.ulb.ac.be/rsat/).
  - This random sequence mimics the amino acid and dipeptide composition of yeast proteins (generated with First order Markov model).
- A blast search against the non-redundant database returns several hits.
- These hits however have low scores. Not surprisingly, the corresponding expected values are higher than 1.



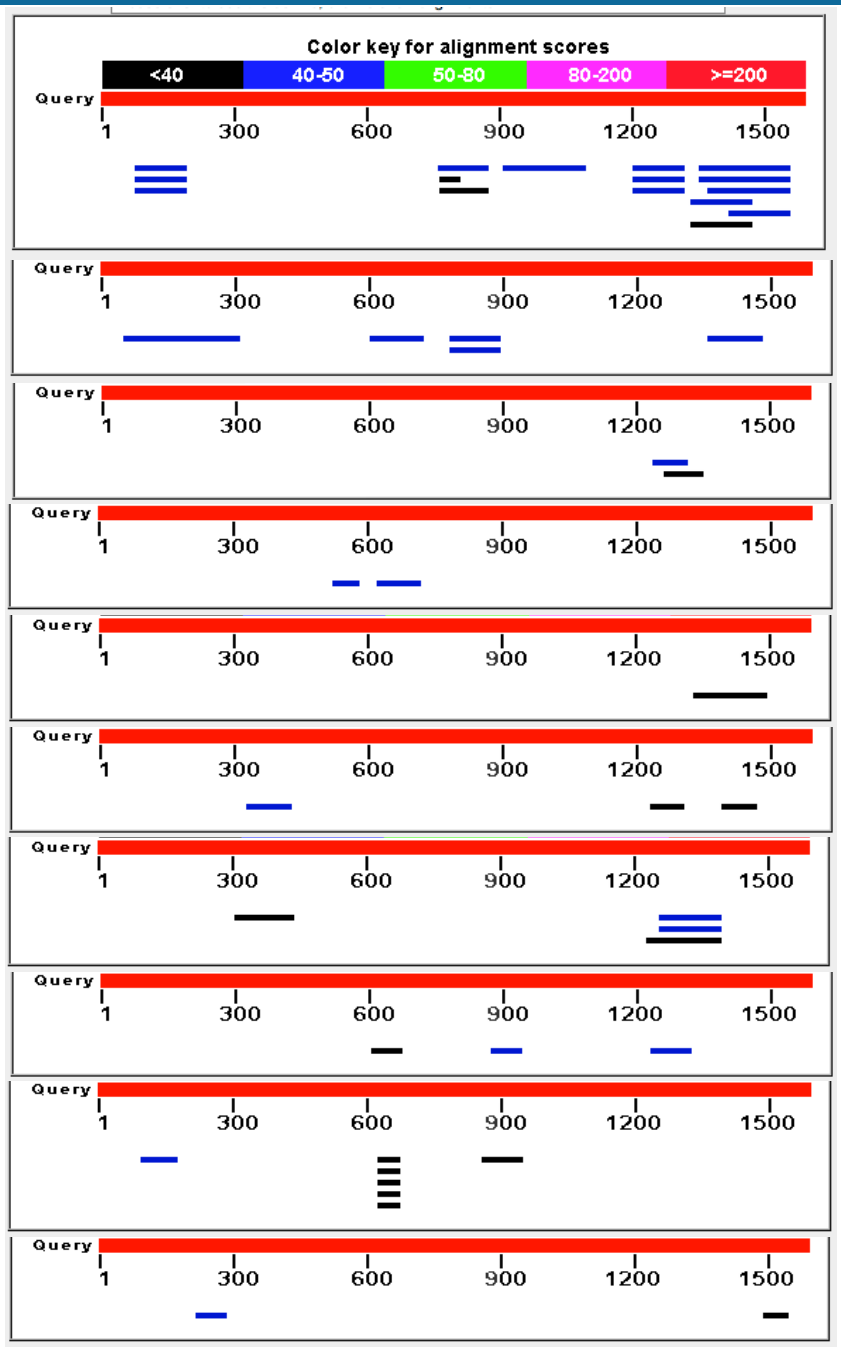
Descriptions

Legend for links to other resources: U UniGene E GEO G Gene S Structure M Map Viewer PubChem BioAssay

Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
Q31606.2	RecName: Full=UPF0413 protein yjbH	34.7	34.7	3%	5.6	31%	G
Q2KN98.1	RecName: Full=Cytospin-A; AltName: Full=SPECC1-like protein; AltNa	35.0	35.0	9%	6.2	22%	GM
Q2KN99.1	RecName: Full=Cytospin-A; AltName: Full=SPECC1-like protein; AltNa	35.0	35.0	9%	6.7	22%	GM
Q69YQ0.2	RecName: Full=Cytospin-A; AltName: Full=Renal carcinoma antigen N	35.0	35.0	9%	6.8	21%	GM
Q2KNA0.1	RecName: Full=Cytospin-A; AltName: Full=SPECC1-like protein; AltNa	34.7	34.7	9%	7.4	21%	GM
Q931S1.1	RecName: Full=Regulatory protein msrR	33.9	33.9	2%	8.9	38%	G
Q5HG57.1	RecName: Full=Regulatory protein msrR >sp|Q99Q02.1|MSRR_STAAM	33.9	33.9	2%	8.9	38%	G
Q2KNA1.1	RecName: Full=Cytospin-A; AltName: Full=SPECC1-like protein; AltNa	34.7	34.7	9%	8.9	21%	GM
Q2SSW4.1	RecName: Full=Alanine--tRNA ligase; AltName: Full=Alanyl-tRNA synt	34.7	34.7	4%	9.1	30%	G
Q196X0.1	RecName: Full=Probable DNA-directed RNA polymerase II subunit RPE	34.3	34.3	5%	10.0	33%	

blastp with random sequences



- pblast of 10 random sequences against the non-redundant database.
  - Between 1 and 15 matches per trial.
  - Was this to be expected ?
- This corresponds pretty well to the expectation
  - On the NCBI BLAST web server, the default threshold on expect has been set to 10.
  - We thus expect, for each request, an average of 10 matches by chance.
  - We indeed observe this order of magnitude when submitting random sequences.

The modalities of BLAST

DNA versus protein searches

- When the query is a coding DNA sequence, it is recommended to apply searches with the translated rather than raw DNA sequences
  - This allows to introduce a substitution matrix (PAM, BLOSUM, ...), which better reflects the evolutionary changes.
  - It has been shown that some distant relationships can be detected with translated searches, but escape detection with the DNA search.
  - It is easier to filter out low complexity regions from proteins than from DNA sequences.

Traduction d'une séquence nucléique dans les 6 phases

- Si l'on dispose d'une séquence nucléique, on peut facilement déduire la séquence de la protéine qui pourrait être produite par sa traduction, sur chacun des 6 brins.
- Si cette séquence n'est pas codante, on s'attend à trouver des codons stop assez fréquemment (3 codons sur 64).
- Cependant, rien n'empêche d'aligner les 6 séquences ainsi produites avec d'autres séquences peptidiques.

Traduction sur 6 phases

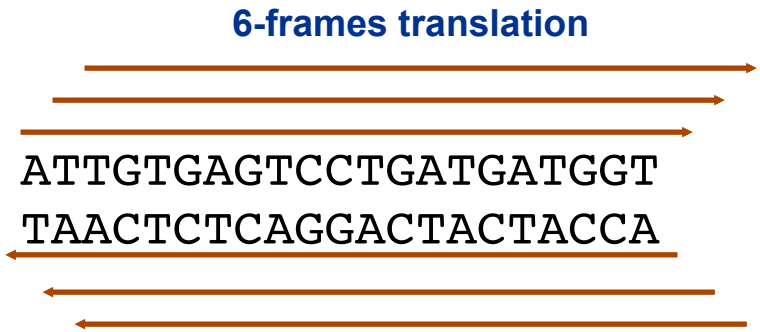
Résultat

```
F1       I  V  S  P  D  D  G
F2        L  *  V  L  M  M  V
F3         C  E  S  *  *  W  X
      1 ATTGTGAGTCCTGATGATGGT 21
        ----:----|----:----|-
      1 TAACACTCAGGACTACTACCA 21
F6        X  T  L  G  S  S  P
F5        X  Q  S  D  Q  H  H
F4         N  H  T  R  I  I  T
```

- Outil: http://www.ebi.ac.uk/Tools/st/emboss_sixpack/

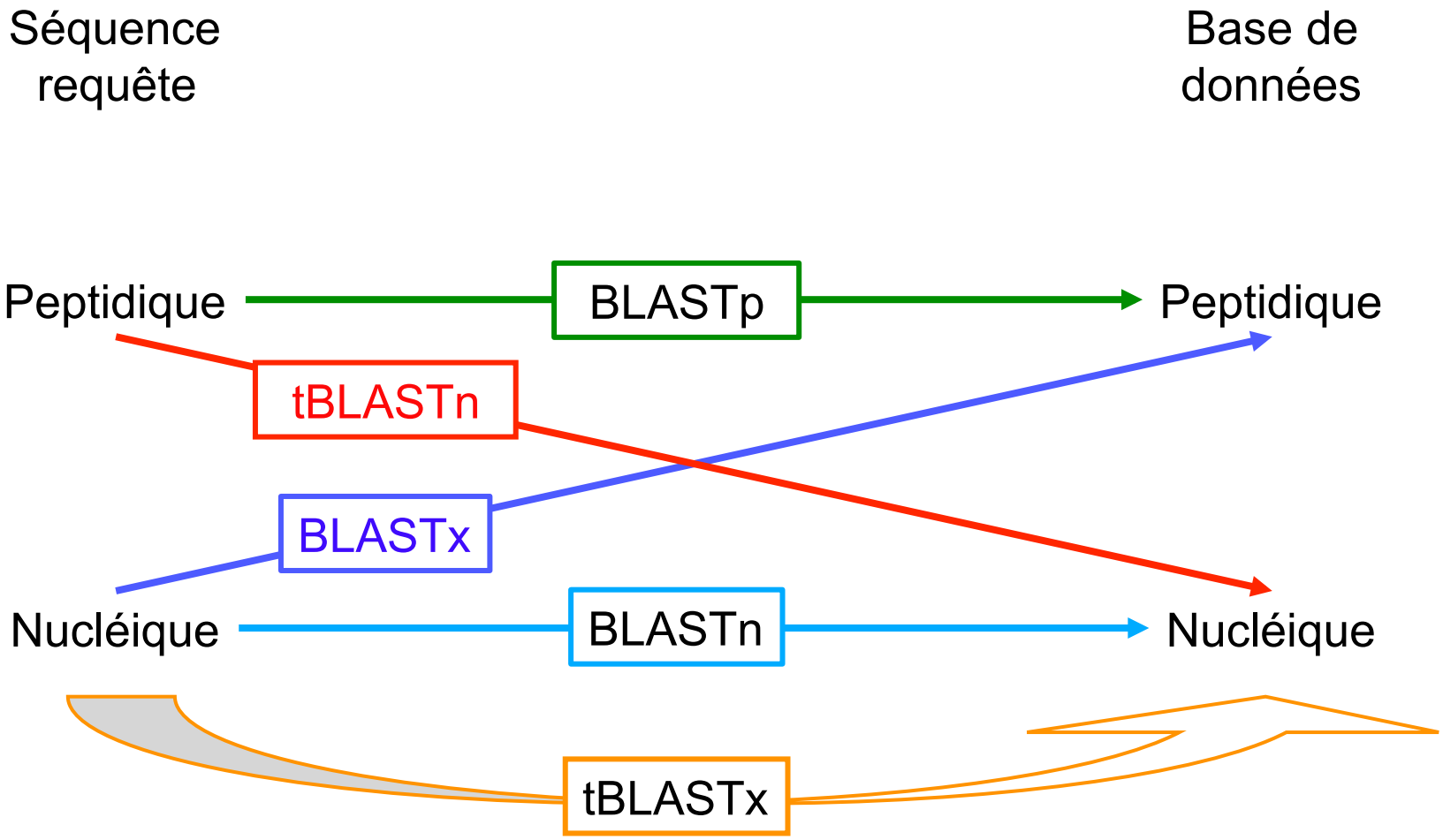
BLAST - a family of purpose-specific programs

- Different program names exist, depending on the type (protein or nucleic acid) of query and database sequences.
- For comparison between nucleic acids and proteins, the nucleic acid is translated in the 6 frames (3 frames per strand)



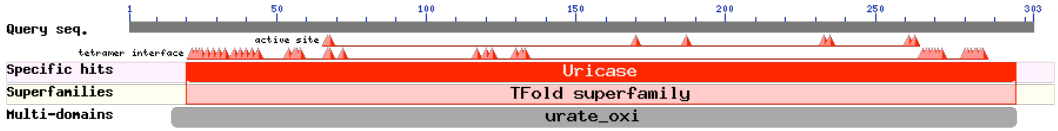
Query	Database	Program	Application examples	Study cases
protein	protein	blastp	Starting from a protein of known function detect putative homologs in the whole Uniprot database.	Collect sequences similar to the blue-sensitive opsin in all human proteins.
nucleic acid	nucleic acid	blastn	Match RNAi against a genome. Match mRNA (or EST) against a genome.	
nucleic acid (translated)	protein	blastx	After having sequenced a piece of DNA, search potentially coding fragments + their putative homologs without any prior knowledge of gene positions in the query sequence.	
protein	nucleic acid (translated)	tblastn	- Identify a genomic region likely to code for an homolog of a protein of interest. - Identify pseudo-genes (defective genes, with many stop codons) for a protein of interest in a genome.	Do cats see colors ? Get Human blue-sensitive opsin protein, connect UCSC genome browser, use BLAT to find similarities in Cat genome
nucleic acid (translated)	nucleic acid (translated)	tblastx		

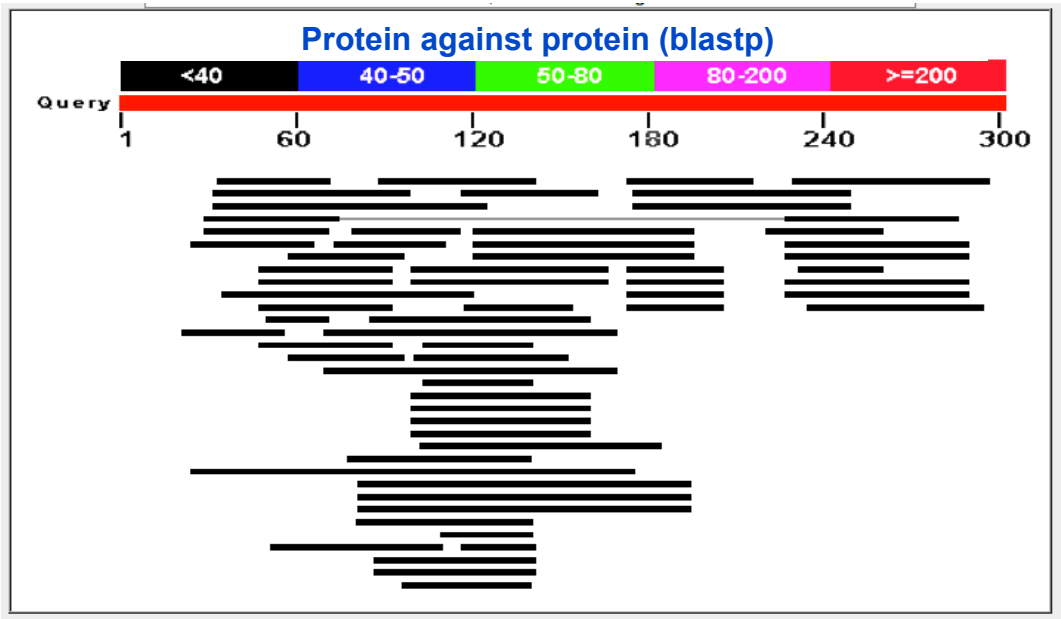
Les modalités de BLAST

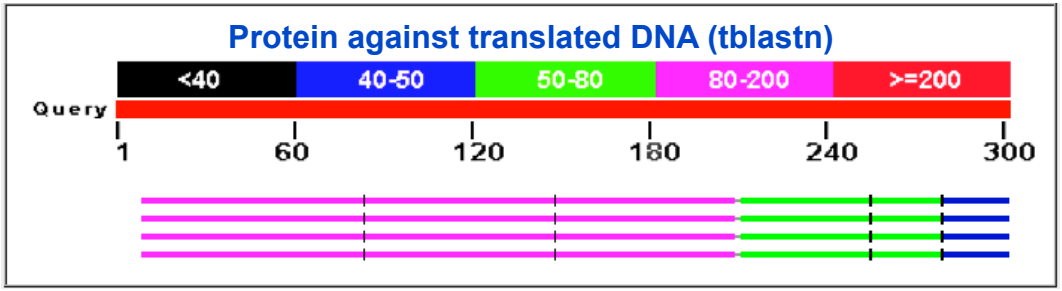


Scanning 6-frames translated genomes with a protein sequence

- The mouse urate oxidase enzyme (P25688) contains an uricase domain (EC 1.7.3.3), catalyzing the urate degradation:
  - Urate + O(2) + H(2)O <=> 5-hydroxyisourate + H(2)O(2).
- Top: **blastp** result (protein against protein):
  - Homo sapiens reference proteins (Refseq) scanned with urate oxidase peptidic sequence.
  - The scan only returns weakly scoring matches (lowest E-value = 0.004).
- Bottom: **tblastn** (search translated nucleotide database using protein query)
  - The scan returns 4 very high-scoring genome locations.
- Question: why did we identify very good matches with tblastn, and not with blastp ?

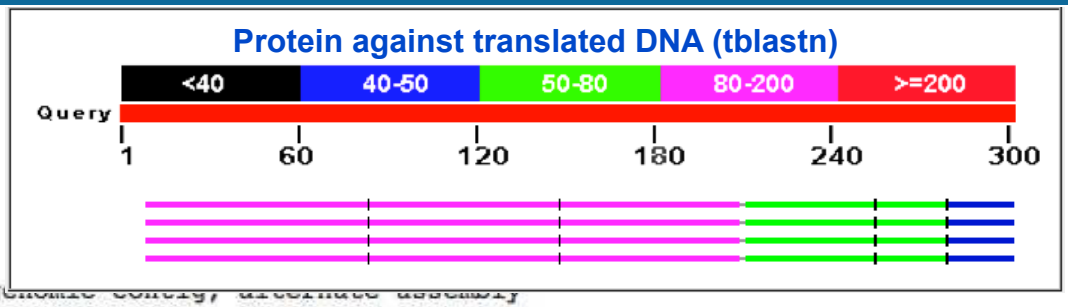






tblastn result : mouse urate oxidase agains Human genome

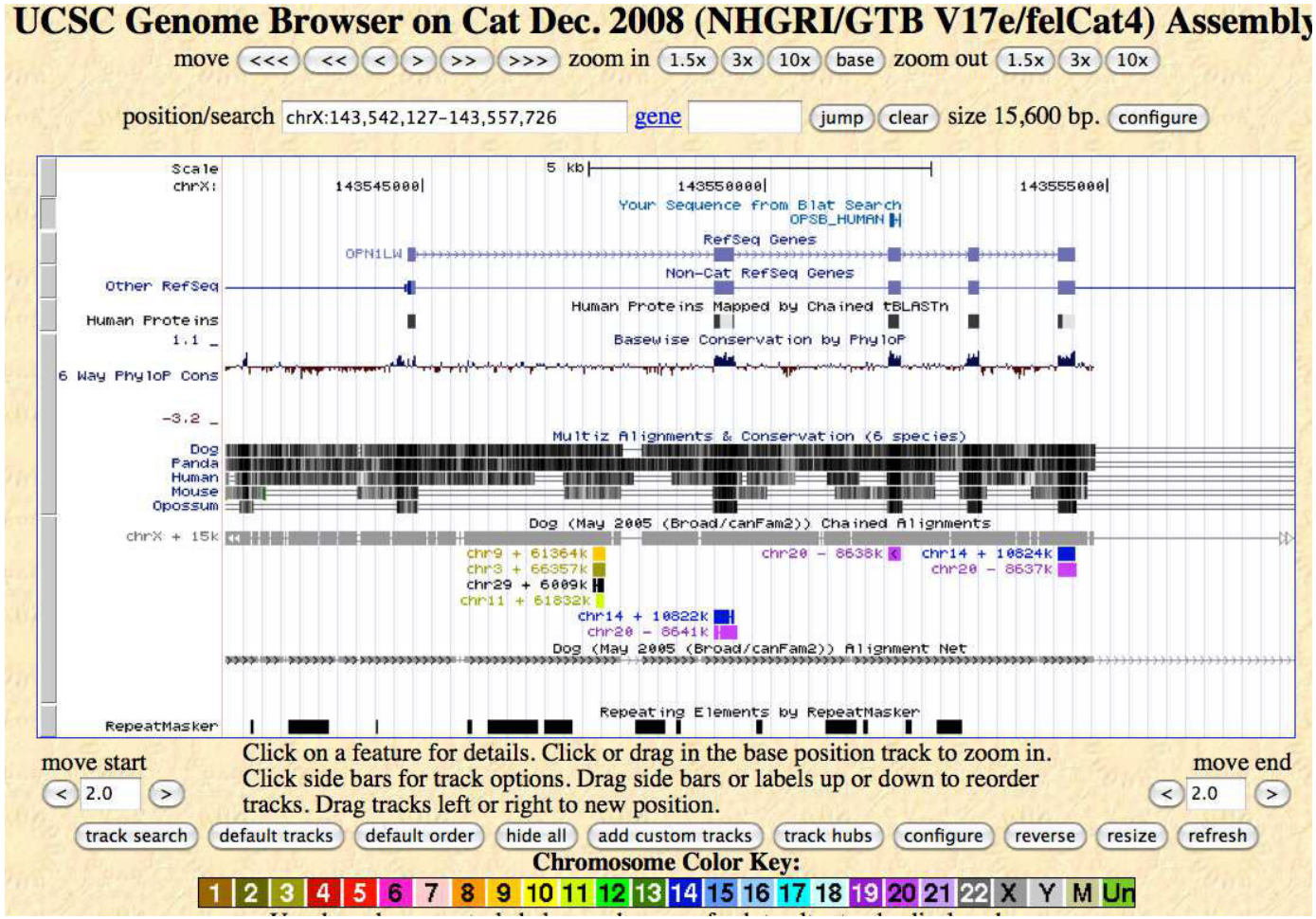
- Some aligned fragments.




```
>ref|NW_001838589.2| Homo sapiens chromosome 1 genomic contig, alternate assembly
HuRef SCAF_1103279188310, whole genome shotgun sequence
Length=20217283

                                                    Sort alignments for this subject sequence by:
                                                      E value  Score  Percent identity
                                                      Query start position  Subject start position
 Features flanking this part of subject sequence:
   27555 bp at 5' side: deoxyribonuclease-2-beta isoform 2
   34809 bp at 3' side: sterile alpha motif domain-containing protein 13 isoform 2

 Score =   137 bits (346),  Expect = 6e-33, Method: Compositional matrix adjust.
 Identities = 67/75 (89%), Positives = 71/75 (95%), Gaps = 0/75 (0%)
 Frame = +1

Query  10        KNDEVEFVRTGYGKDMVKVLHIQRDGKYHSIKEVATSVQLTLRSKKDYLHGDNSDIIPTD  69
                 +NDEVEFVRTGYGK+MVKVLHIQ DGKYHSIKEVATSVQLTL SKKDYLHGDNSDIIPTD
Sbjct  19269451  QNDEVEFVRTGYGKEMVKVLHIQ*DGKYHSIKEVATSVQLTLSSKKDYLHGDNSDIIPTD  19269630

Query  70        TIKNTVHVLAKLRGI  84
                 TIKNTVHVLAK + +
Sbjct  19269631  TIKNTVHVLAKFKEV  19269675


 Features flanking this part of subject sequence:
   30259 bp at 5' side: deoxyribonuclease-2-beta isoform 2
   32033 bp at 3' side: sterile alpha motif domain-containing protein 13 isoform 2

 Score =   103 bits (256),  Expect = 1e-21, Method: Compositional matrix adjust.
 Identities = 51/99 (52%), Positives = 59/99 (60%), Gaps = 33/99 (33%)
 Frame = +2

Query  84        IRNIETFAMNICEHFLSSFNHVTRAHVYVEEVPWKRFEK----------------------  122
                 I++IE F +NICEHFLSSFNHV RA VY+EE+PWK   K
Sbjct  19272155  IKSIEAFGVNICEHFLSSFNHVIRAQVYMEEIPWKHLGKVNSLICALSLIKEIGFMA*TE  19272334

Query  123       ------------NGIKHVHAFIHTPTGTHFCEVEQMRNG  149
                            NG+KHVHAFIHTPTGTHFCEVEQ+R+G
Sbjct  19272335  IFLEIICFPNFQNGVKHVHAFIHTPTGTHFCEVEQLRSG  19272451
```

tblastn example – do cat see colors ?

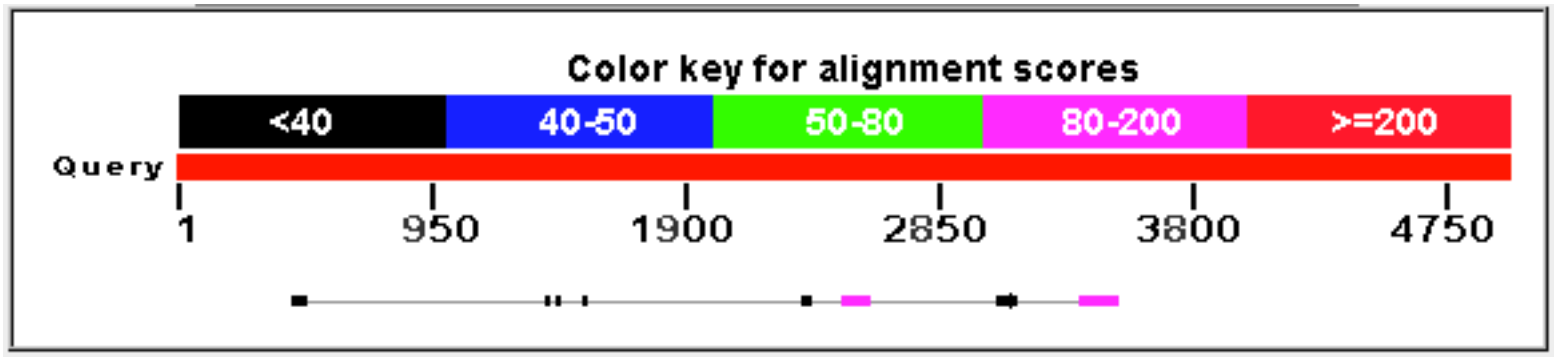
- Approach: match the peptidic sequence of the Human Short-wave-sensitive opsin (blue) against the complete cat genome (6-frames translated).
- Tool: BLAT tool at UCSC genome browser (http://genome.ucsc.edu/)
- Result: 3 matches in cat genome
  - Short wave (blue) sensitive opsin
  - rhodopsin
  - Long wave (red) sensitive opsin. Partial match with 1 exon, but sufficient to “fish out” the cat OPN1LW gene.



Blastx example: scanning a genomic sequence with a protein sequence

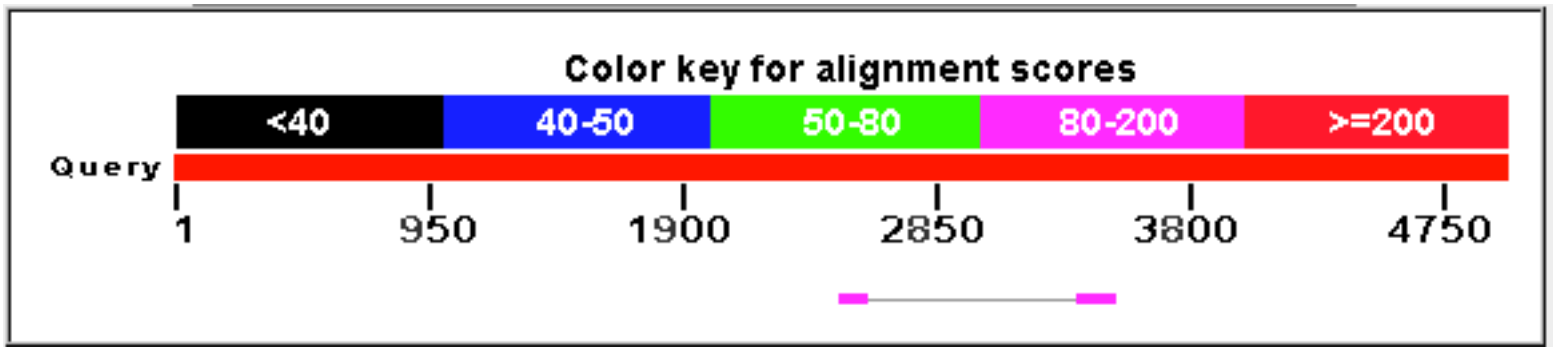
Insuline – scan de la séquence génomique avec la protéine

- expect threshold = 10; low-complexity filter ON
- Note the surprising matches on reverse complementary strand



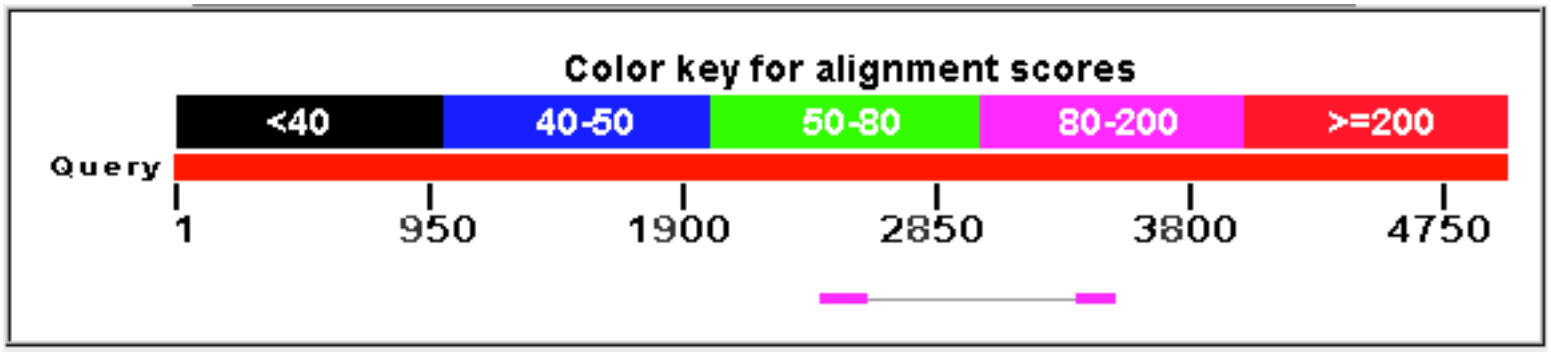
Insuline – scan de la séquence génomique avec la protéine

- expect threshold = 1e-5; low-complexity filter ON



Insuline – scan de la séquence génomique avec la protéine

- expect threshold = 1e-5; low-complexity filter OFF



Insuline – scan de la séquence génomique avec la protéine

- expect threshold = 10; low-complexity filter OFF