

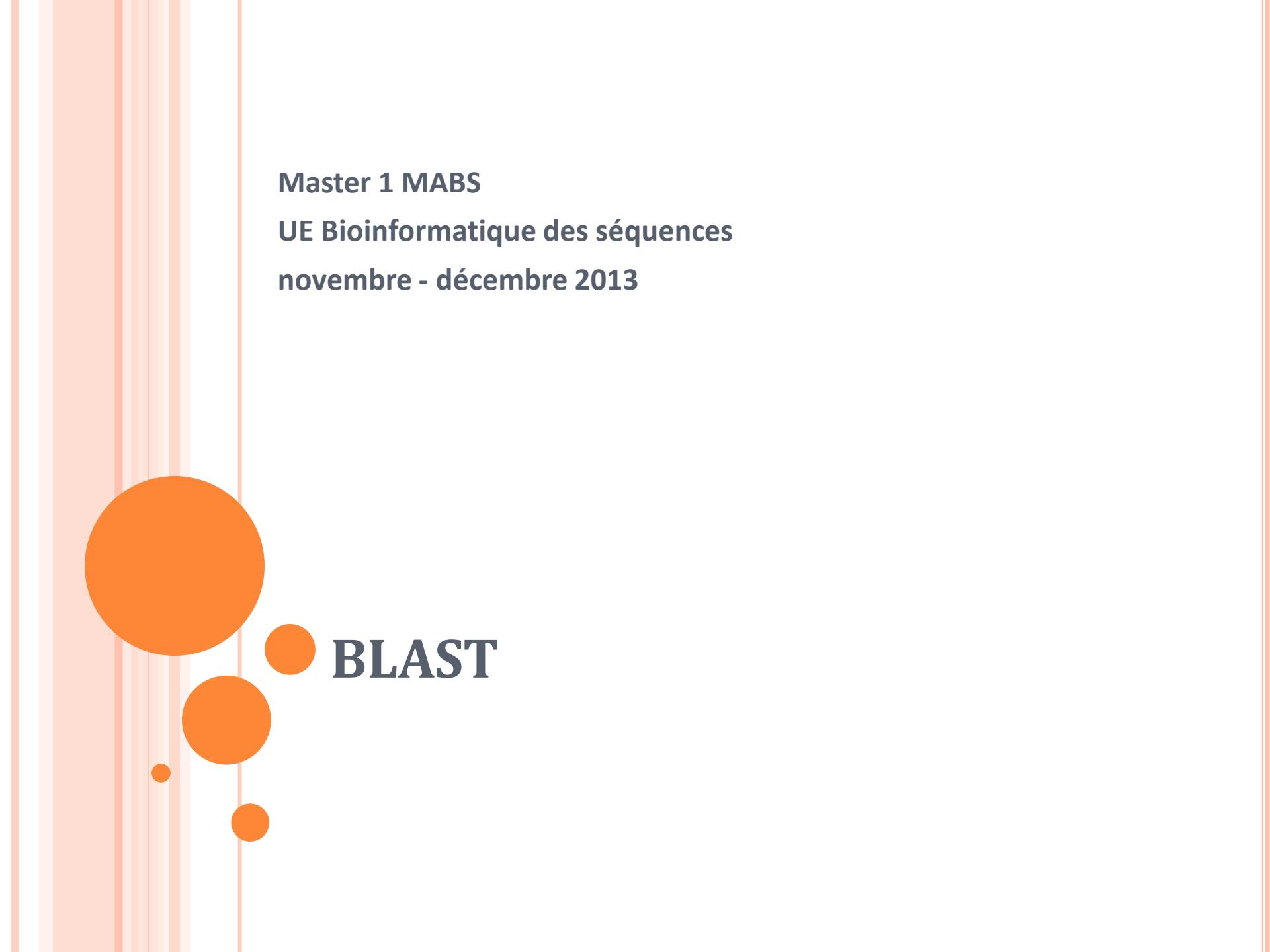


Master 1 MABS

UE Bioinformatique des séquences

novembre - décembre 2013

BLAST ET COMPAGNIE



Master 1 MABS
UE Bioinformatique des séquences
novembre - décembre 2013

BLAST

BLAST ALTSCHUL ET AL. 1990

- L'idée sous-jacente à **l'heuristique** de Blast (**Basic Local Alignment Search Tool**) est que les bons alignements doivent contenir quelque part des **petits segments strictement identiques**.
- Ces éléments constituent les **points d'ancrage** a partir desquels l'alignement est étendu.
- Blast2 est une version de Blast qui autorise les insertions et les délétions, c'est la version à utiliser.

Heuristique = algorithme imparfait donnant rapidement une bonne réponse, mais pas forcément la meilleure

- Exemple d'heuristique : dans un labyrinthe, suivre un mur avec sa main gauche permet de trouver la sortie, mais ce n'est pas forcément le chemin le plus court !



BLAST ALTSCHUL ET AL. 1990

- De plus, des **filtres** (programmes SEG et XNU) ont été conçus pour éliminer les régions répétitives et segments de "**faible complexité**" qui bruitent les résultats. Pour cela, la séquence requête est tout d'abord comparée à une banque de données contenant des séquences représentatives de faible complexité.
- Les fragments de la séquence requête appartenant à ces familles sont alors **masqués** avant d'effectuer la recherche de similitude sur la banque complète.
Exemple: Queue PolyA, PolyProline, etc...



LES ÉTAPES DU BLAST EN « FRANÇAIS »

BLAST recherche des régions sans insertions / délétions riches en similarité entre une séquence protéique et toutes les séquences d'une banque

- ➊ 1) On travaille sur des « mots » de 3 acides aminés
 - ➏ $20 \times 20 \times 20 = 8000$ mots de 3 acides aminés possibles
 - ➏ 8000 est un nombre petit pour un ordinateur !
 - ➏ => on crée les 8000 mots possibles

Index :

Q T Y

Q V L

Q V S

A L Q

L Q V

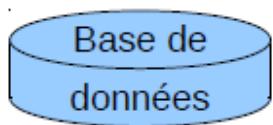
Q V N

...



LES ÉTAPES DU BLAST EN « FRANÇAIS »

- 2) Chaque séquence présente dans la base est découpée en mots de 3 acides aminés (« hachage »)
 - On indexe chaque séquence en fonction des mots qu'elle contient



Séquence n°455 : ...

Séquence n°456 :

A L Q V S P Y T A...

↓
hachage

A L Q
L Q V
Q V S
V S P
...

Index :

Q T Y
Q V L

Q V S : présent dans la séquence
n°456 en position 3

A L Q
L Q V
Q V N
...



LES ÉTAPES DU BLAST EN « FRANÇAIS »

- 2) Chaque séquence présente dans la base est découpée en mots de 3 acides aminés (« hachage »)
 - On obtient un index qui indique pour chaque mot dans quelles séquences il se trouve, et en quelles positions
 - NB : cette étape prend beaucoup de temps, mais elle est réalisé à l'avance, lors de la constitution de la base de données !

Index :

Q T Y : présent dans la séquence n°98 en position 51,...

Q V L : présent dans la séquence n°122 en position 14,...

Q V S : présent dans la séquence n°456 en position 3,
 présent dans la séquence n°518 en position 74,...

A L Q : présent dans la séquence n°456 en position 1,...

L Q V : présent dans la séquence n°456 en position 2,...

Q V N : présent dans la séquence n°542 en position 32,...

...



LES ÉTAPES DU BLAST EN « FRANÇAIS »

- 3) On découpe la séquence recherchée en mots de 3 acides aminés (« hachage »)

L Q V S P L T A...

↓ hachage

L Q V
Q V S
V S P
S P L
...



LES ÉTAPES DU BLAST EN « FRANÇAIS »

- 4) Pour chacun de ces mots, on va chercher dans l'index les séquences de la base où est présent soit ce mot, soit un mot proche (score d'alignement supérieur à un seuil fixé avec BLOSUM62) => des « hits »

L Q V S P L T A...



L Q V
Q V S →
V S P
S P L
...

Index

Séquence n°456, position 3 :

A L Q V S P Y T A...

Score d'alignement : 18

Séquence n°518, position 74 :

A V V I Q V S P Y...

Score d'alignement : 18

séquence n°542, position 32 :

Z R L Q V N P T T...

Score d'alignement : 14



LES ÉTAPES DU BLAST EN « FRANÇAIS »

- 5) Pour chaque « hit », on cherche à étendre l'alignement avant et après les 3 acides aminés
 - Si le score de l'alignement devient trop faible, on s'arrête
 - Si le score de l'alignement devient important, on considère qu'il y a alignement local

L Q V S P L T A...



Séquence n°456, position 3 :

A L Q V S P Y T A...
| | | | | | | |
L Q V S P L T A...

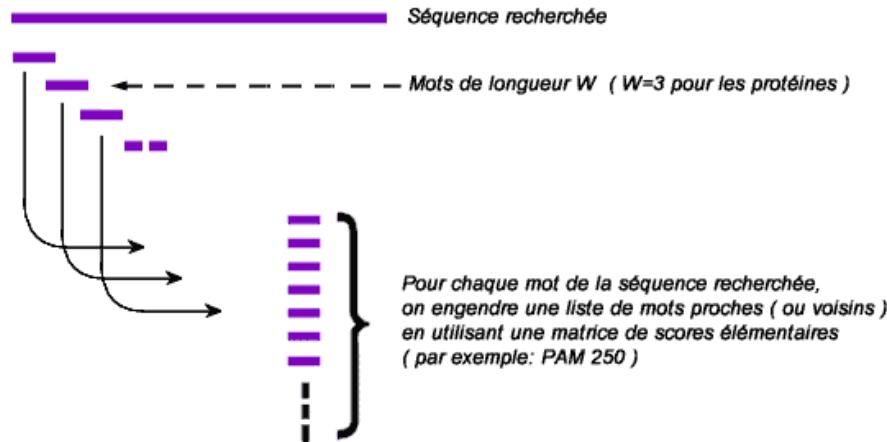
Séquence n°518, position 74 :

...
séquence n°542, position 32 :
...

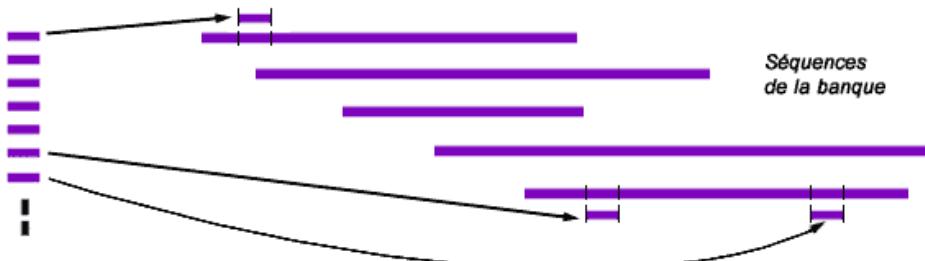


BLAST PRINCIPE

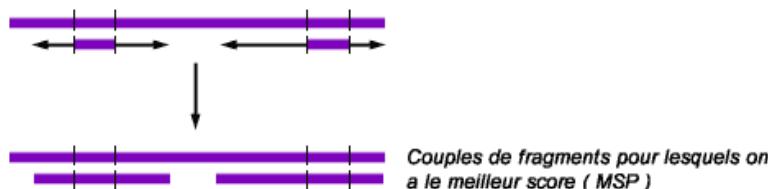
1) Détermination de la liste des mots de longueur W et de leurs proches



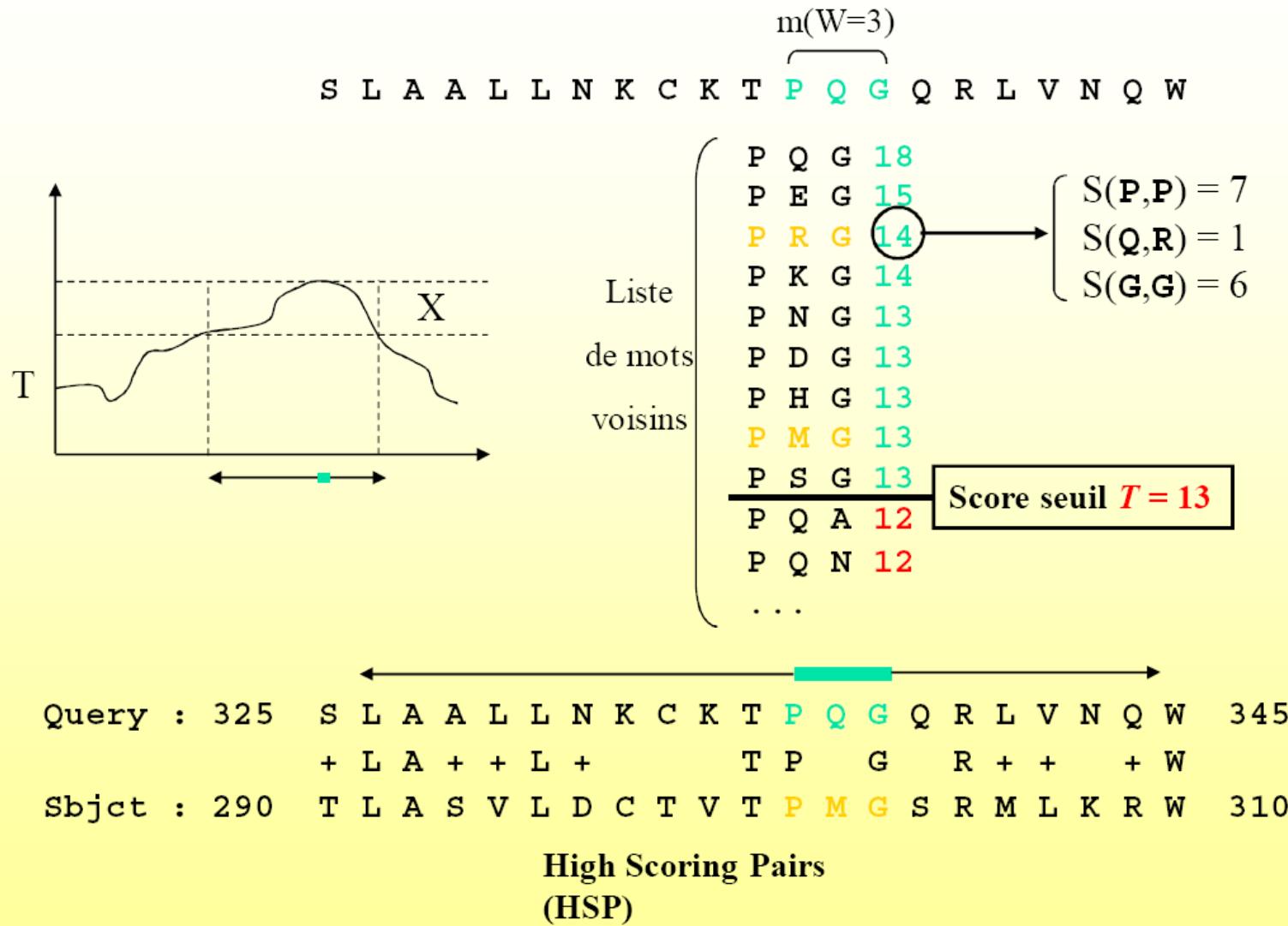
2) Identification des mots de la liste sur les séquences de la banque



3) Pour chaque mot trouvé sur une séquence de la banque, on essaie d'étendre la zone de similitude à gauche et à droite



BLAST PRINCIPE



BLAST ÉTAPES DU CALCUL

- Recherche de tous les mots de taille W communs aux séquences avec un score de similitude supérieur à T
 - Hit Blast
 - W = 11 pour ADN
 - W = 3 pour protéines
 - la valeur de W est ajustable
- T = score seuil au-delà duquel la ressemblance entre deux mots de taille W n'est pas due au hasard.
 - T est ajustable



BLAST ÉTAPES DU CALCUL

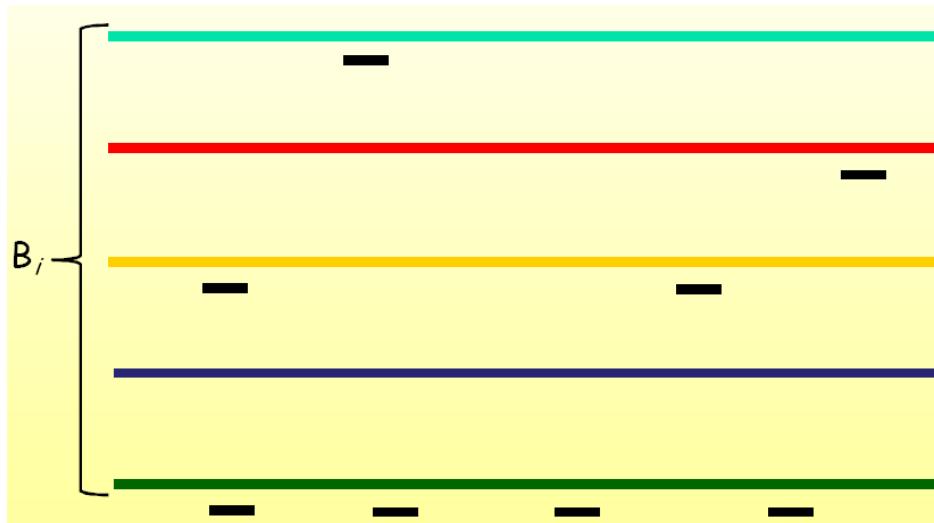
- Recherche de régions sans insertions/délétions riches en similarités
- Détermination d'une longueur de mot : $W = 3$ acides amines pour les protéines
- **Hachage** de la séquence « requête » (query) en mot de taille W



- Liste de mots voisins de longueur W ayant un score supérieur à un seuil T fixe par rapport au mot m .

BLAST ÉTAPES DU CALCUL

- Chaque mot similaire au mot **m** est comparé à chaque mot de taille **W** pris dans chaque séquence **B_i** de la banque de données.
- Lorsqu'un mot d'une séquence **B_i** est identique à un mot de la liste de mots voisins, un **hit** est enregistré.

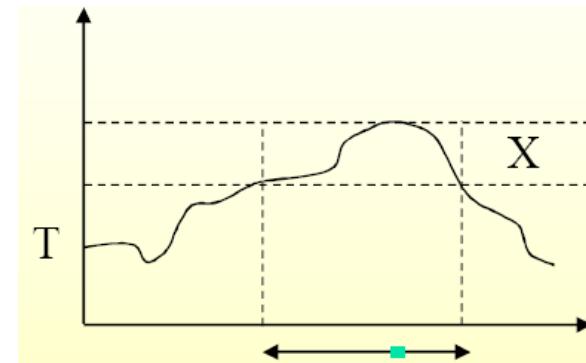


BLAST ÉTAPES DU CALCUL

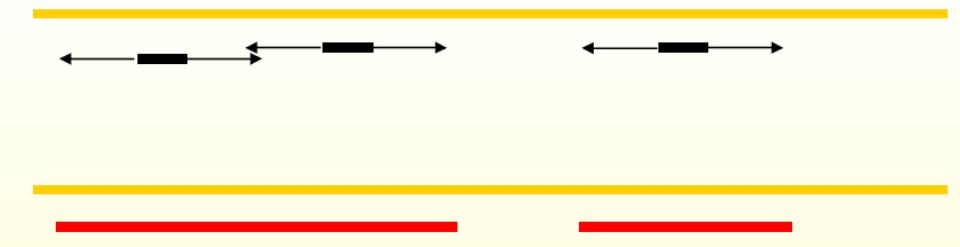
- **Extension** des mots trouvés dans **les deux directions** pour trouver les régions de similitude les plus longues possibles ayant un score supérieur ou égal à un score seuil **S**

HSP, High-scoring Segment Pair

- Arrêt de l'extension **S_i** lorsque :
 - Diminution de **X** du score cumulé par rapport au maximum atteint
 - Score cumulé ≤ 0
 - Fin d'une des séquences



BLAST ÉTAPES DU CALCUL

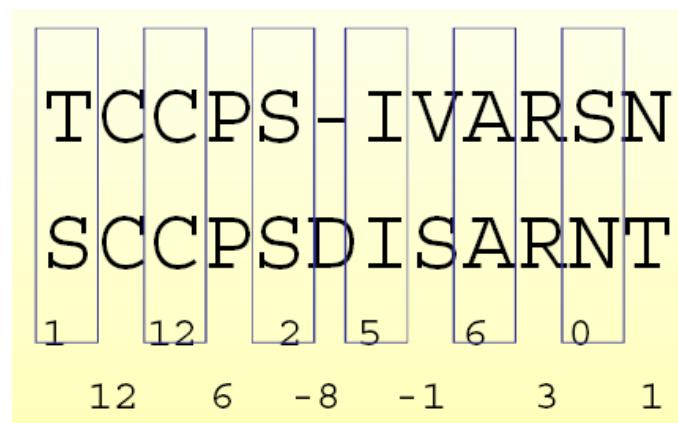


- Pour chaque **hit**, le programme effectue une extension de l'alignement dans les deux sens. (en gros alignement local de type Smith et Waterman).
- L'extension s'arrête quand le score du mot étendu diminue au-delà d'un seuil fixe.
- Les segments ayant un score de similarité supérieur à un score S seuil fixe sont retenus (**High Scoring Pairs = HSP**).

BLAST CALCUL DU SCORE

- Le Score d'un alignement est la somme des scores de toutes les paires de résidus dans l'alignement :
- **score HSP = $\sum S_{i,j} - \text{gaps}$**

- Séquence 1:
- Séquence 2:



- => alignement score = 39



SIGNIFICATIVITÉ DES ALIGNEMENTS

- 3 mesures sont données pour comprendre la significativité de l'alignement obtenu:
- le bit score,
- la P-value,
- la E-value



SIGNIFICATIVITÉ DES ALIGNEMENTS

- Le Score S'
- le score S' est dérivé du score brut de l'alignement.
- Il a été normalisé et peut donc être utilisé pour comparer des scores provenant de recherches différentes.

$$S' = (\lambda_{S-\ln K}) / \ln 2 \quad \text{“bit score”}$$

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

λ : paramètre lié au système de score



ATTENTION AUX SCORES

BLAST SIGNIFICATION DE LA E-VALUE

- **E-value (Expect)** = nombre d'alignements attendus par hasard ayant un score supérieur au score S obtenu pour l'HSP dans la banque considérée
 - Plus la valeur est faible, plus l'alignement est fiable
 - Dépend de la taille de la banque de données utilisée !
 - Valeurs non comparables entre deux banques

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

m : longueur seq 1

n : longueur seq 2

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

Contre une banque de séquences :

$n \Rightarrow N$ (longueur totale de la base)



BLAST SIGNIFICATION DE LA P-VALUE

○ P-value (probability)

- **P(N)**: Probabilité du score observe. Plus cette valeur est faible, plus l'HSP est significatif.
- Nombre de HSP dont score $\geq S$ suit une loi de Poisson.
- Probabilité de n'avoir aucune HSP de score $\geq S$ est :

$$e^{-E} \quad \text{Pour } a \text{ HSP : } e^{-E} (E^a / a!)$$

- Donc la probabilité d'avoir au moins une telle HSP est :

$$P = 1 - e^{-E}$$

E-value	P-value
5	0.993
10	0.99995



BLAST SIGNIFICATION D'UN ALIGNEMENT

- Taille de la base de données = 20×10^6 lettres

peptide	nombre présents par hasard
A	1×10^6
AP	50000
IAP	2500
LIAP	125
WLIAP	6
KWLIAP	0,3
KWLIAPY	0,015



BLAST SIGNIFICATION DE LA E-VALUE

- Exemple BlastP du NCBI de P17538.1 contre la base de données SwissProt
- seq query= 263 Aa

	score	% identité	e-value
P17538.1 Chymotrypsinogen B;	533	100%	3e-151
Q6GPI1.2 Chymotrypsinogen B2;	525	100%	1e-148
.../...			
P42882.1 Protein NMT1 homolog	30.8	15%	8.7
Q7LZF5.1 Thrombin-like enzyme catroxobin-1;	30.8	6%	9.8

- La valeur de la e-value pourrait signifier:
- Dans une banque de donnée quelconque de même taille que SwissProt, je m'attends à trouver **9,8 séquences** qui ressembleront au moins autant que la séquence Q7LZF5 avec ma séquence query.
- La e-value de Blast **n'est en aucun cas** un nombre qui vous dit si l'alignement de deux séquences est "**biologiquement significatif**" ou non, c'est un **outil d'aide à la décision**.

PERTINENCE/INTERPRÉTATION BIOLOGIQUE

- La significativité statistique ne suffit pas !
- Elle dépend de la taille de la banque interrogée...
- Tout dépend de ce que l'on cherchait... petites ou grandes régions +/- conservées
- Mismatches conservatifs ? Domaine protéique, CDS et usage du code
- Importance structurale et/ou fonctionnelle des régions alignées ou non



BLAST EXEMPLE D'UN HSP SEQUENCE QUERY= 256 AA

>| sp|Q2KJ63.1|KLKB1 BOVIN [G] RecName: Full=Plasma kallikrein; AltName: Full=Plasma prekallikrein; AltName: Full=Kininogenin; AltName: Full=Fletcher factor; Contains: RecName: Full=Plasma kallikrein heavy chain; Contains: RecName: Full=Plasma kallikrein light chain; Flags: Precursor Length=636

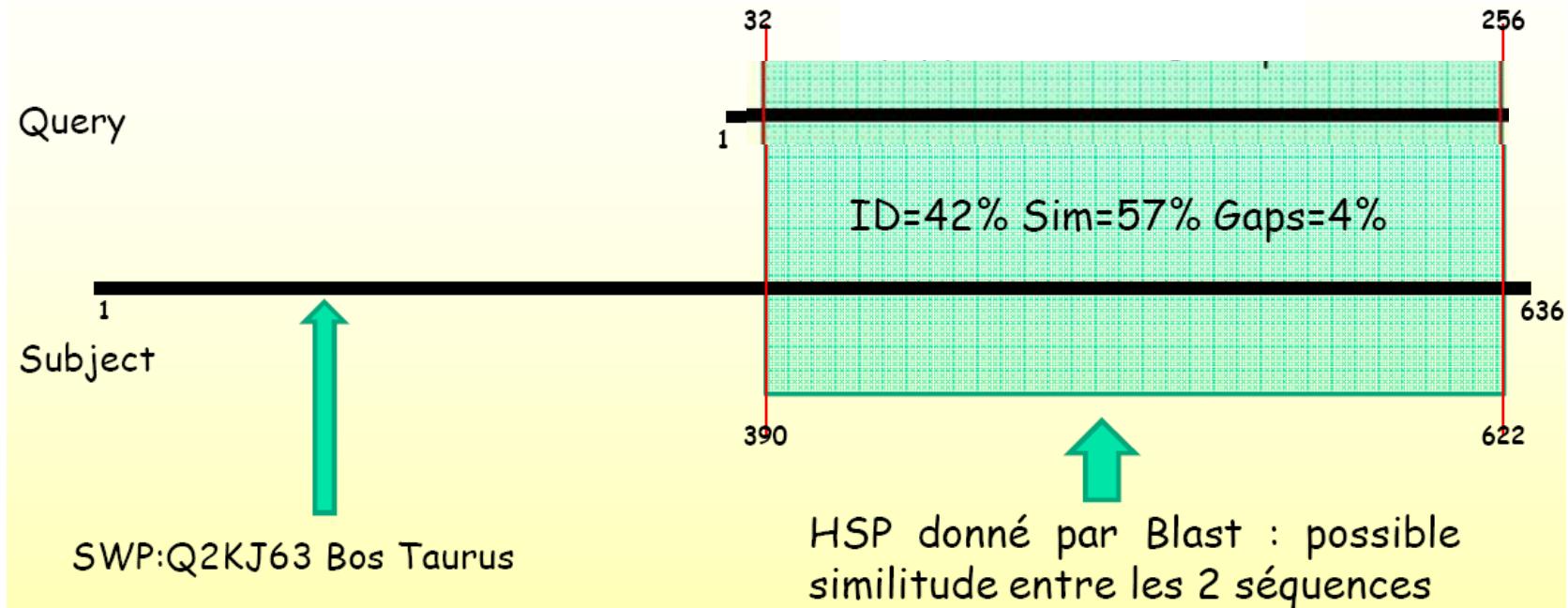
GENE ID: 533547 KLKB1 | kallikrein B, plasma (Fletcher factor) 1 [Bos taurus]

Score = 183 bits (465), Expect = 7e-46, Method: Compositional matrix adjust.
Identities = 100/234 (42%), Positives = 134/234 (57%), Gaps = 10/234 (4%)

Query	32	SRIVNGEDAVPGSWPWQVSLQ--DKTGFHFCGGSLISEDWVVTAAHC--GVRTSDV--VV	85
	+RIV G +A G WPWQVSLQ	+ H CGGS+I WV+TAAHC G+ S++ +	
Sbjct	390	TRIVGGTNASWGEWPWQVSLQVKQRAQSHLGGSIIGRQWVLTAAHCFDGLLLSNIWRIY	449
Query	86	AGEFDQGSDEENIQVLKIAKVFKNPKFSILTVNNNDITLLKLATPARFSQTVSAVCLPSAD	145
	G + +I ++ +P + I ++DI L+KL P F+ A+CLPS D		
Sbjct	450	GGILNLSEITTETSFSQLIKEIIVHPNYKISEGSHDIALIKLEAPLNFTDLQKAICLPSKD	509
Query	146	DDFPAGTLCATWGKGTKYNANKTPDKLQQAALPLLSNAECKKSW-GRRITDVMICAGAS	204
	D P T C TGWG T+ K + LQ+A +PL+SN EC+KS+ +IT M I CAG		
Sbjct	510	DTKPVYTDGWITGWGFTE-EKGKIQNTLQKANIPLISNEECQKSYRDYKITKQMICAGYK	568
Query	205	--GVSSCMGDSGGPLVCQKDGAWTLVGIVSWGSDTCSTSSPGVYARVTKLIPWV	256
	G +C GDSGGPLVCQ + W LVGI SWG PGVY +V + + W+		
Sbjct	569	EGGKDACKGDSGGPLVCQHEETWHLVGITSWGEGCARREQPGVYTKVAEYVDWI	622

Domaines structuraux importants

BLAST EXEMPLE D'UN HSP



- Que peut-on conclure à propos de la séquence query ?
- Est-elle homologue à la séquence de la banque (Q2KJ63 bovins)?

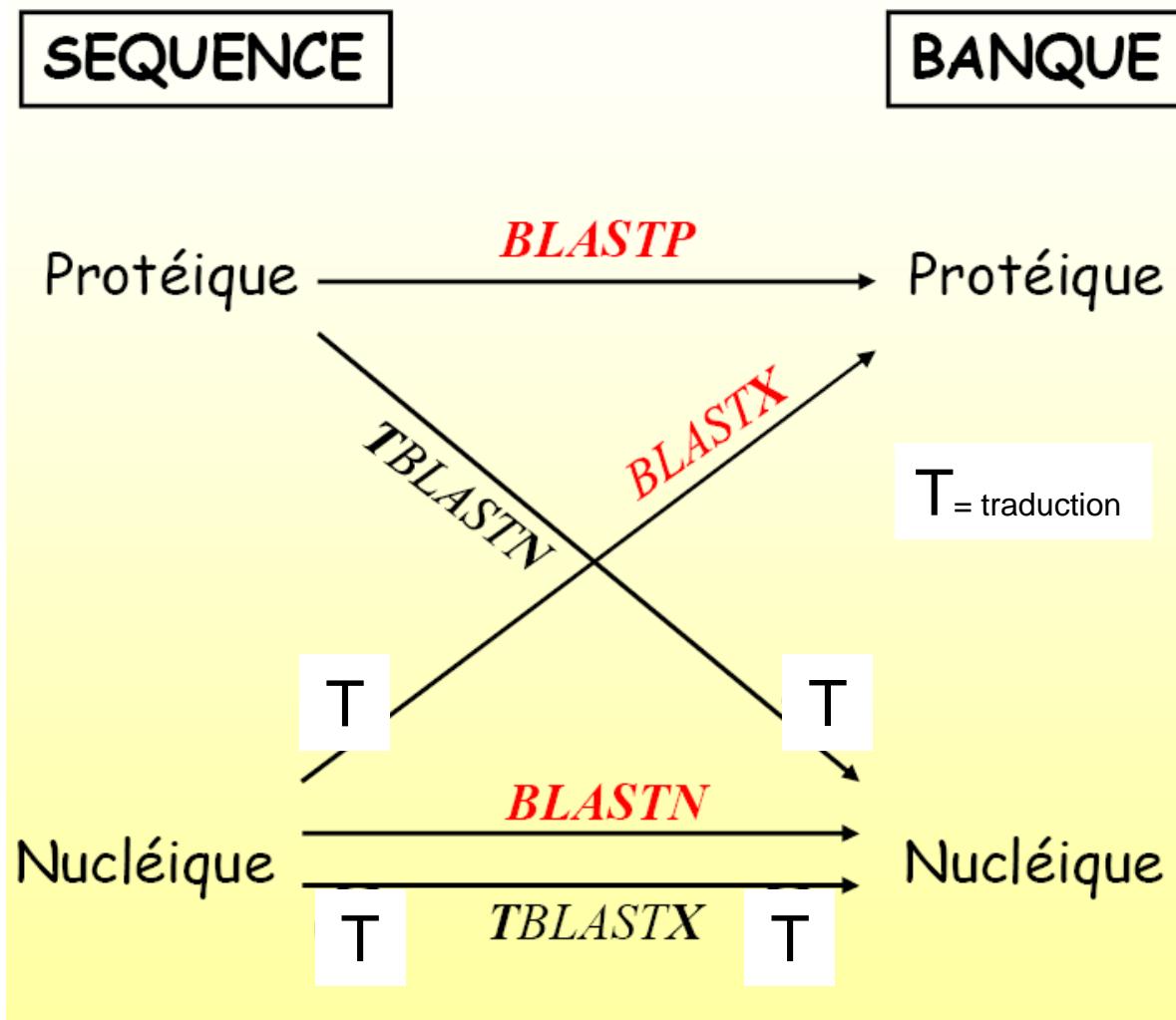
Peu d'acides aminés identiques mais ce sont des acides aminés très « structurant » (P, W, C, F, Y sont conservés), peut-être observons nous ici un domaine structural conservé. L'alignement n'est pas à rejeter, hypothèse à creuser.

BLAST VALEURS INDICATRICES

- **Identities** = nombre paires d'identités / nombre total paires de lettres alignées similitude calculée à partir de la matrice unitaire
- **Positives** = nombre paires avec poids positif / nombre total paires similitude calculée à partir de la matrice de substitution utilisée pour la recherche
- **Gaps** = nombre (insertions ou délétions) / nombre total paires



BLAST CHOIX DU PROGRAMME



BLAST EN PRATIQUE: LES PROGRAMMES

- **Blastn**: acides nucléiques contre acides nucléiques
- **Blastp**: protéines contre protéines
- **Blastx**: acides nucléiques traduits dans les 6 phases contre protéines
- **Tblastn**: protéines contre acides nucléiques traduits dans les 6 phases
- **Tblastx**: acides nucléiques traduits dans les 6 phases contre acides nucléiques traduits dans les 6 phases



BLAST EN PRATIQUE: LES PROGRAMMES (SUITE)

- **PSI-Blast** (Position-Specific Iterated BLAST): Blast relancé plusieurs fois par itération, à chaque itération une séquence consensus est déterminée à partir des résultats et utilisée comme séquence source pour l'itération suivante. **Utile pour comparer des séquences avec de grandes distances évolutives.**
- **PHI-Blast** (Pattern Hit Initiated Blast): Le programme utilise comme source une séquence protéique et un motif, celui-ci est utilisé comme point de départ des recherches de similarité avec les séquences présentes dans les bases de données.
- **MegaBLAST**: Pour séquences nucléiques, optimisé pour séquences peu différentes (erreurs séquençage) - $W = 28$
- **Discontiguous MegaBLAST**: Pour séquences divergentes. Cherche des mots discontinus, dans un ensemble. Considère codant/ non-codant

BLAST EN PRATIQUE

- Base de données: nr (non redondant) est automatiquement sélectionné en version "protéines" ou "acides nucléiques" selon qu'on utilise blastp ou blastn.
- Masquage: les régions risquant de produire des solutions non spécifiques peuvent être remplacées par des X
 - Régions de basse complexité
 - Séquences répétées eucaryotes



ATTENTION

- Pièges: vecteurs
 - Attention: il reste souvent des séquences de vecteurs attachées aux ARNm
- Pièges: basse complexité
 - Séquences riches en AT, riches en résidus hydrophobes, etc.
- Pièges: Alu
 - Séquences répétées = 40% du génome humain
- Pièges: transmembrane
 - Séquences transmembranaires (hydrophobes): même comportement que basse complexité



EFFET DE LA PÉNALITÉ D'INDELS

```

# 1: ILV1_TOBAC
# 2: ILVB_ARATH
# Matrix: EPAM60
# Gap_penalty: 2
# Extend_penalty: 2
#
# Length: 715
# Identity: 531/715 (74.3%)
# Similarity: 586/715 (82.0%)
# Gaps: 93/715 (13.0%)
# Score: 3415

```

Sequence alignment between ILV1_T and ILVB_A showing the effect of gap penalties. The alignment highlights identical residues with colons and similar residues with dots. Orange boxes indicate positions where gaps are present in one sequence.

	10	20	30	40
ILV1_T	MAAAAESP-SSS-AFS-KTLPSSSTSSTLLP-RSTF-PFP-HHPHK			
ILVB_A	MAAAATTTTSSSISFSTKP-SPSSSKSP-L-PISR-FSLFSLN-PNK			
	10	20	30	40

	50	60	70	
ILV1_T	TTPPPLHLTHIHIHSQRRR-F-T-ISNVIST-NQKV-SQT			
ILVB_A	SS-S-S-RRRGIGKSSSPSSISAVLNTTTN-VTTTPSPPT			
	50	60	70	

	80	90	100	110	120
ILV1_T	EK-T-ETFVSRFAPDEPRKGSDVLVEALEREGV-TDVFAYPGGASMEIH				
ILVB_A	-KPTKEETFISRFAPDQPRKGADILVEALERQGVET-VFAYPGGASMEIH				
	80	90	100	110	120

	130	140	150	160	170
ILV1_T	QALTRSS-IIRNVLPRHEQGGVFAAEQYARATIG-FPGVCIATSGPGATNL				
ILVB_A	QALTRSSSI-RNVLPRHEQGGVFAAEQYARSSGK-PGICIAATSGPGATNL				
	130	140	150	160	170

	180	190	200	210	220
ILV1_T	VSGLADALLDSVPIVAITGQVPRRMIGTDQETPIVEVTRSITKHNYLV				
ILVB_A	VSGLADALLDSVPVLAITGQVPRRMIGTDQETPIVEVTRSITKHNYLV				
	180	190	200	210	220

	230	240	250	260	270
ILV1_T	MDVEDIPVVRE-AFFLA-RSGRPGLIDVPKDIQQQLVIPDWDPQPMRL				
ILVB_A	MDVEDIPPHII-EAAFFLAT-SGRGPVLVDVPKDIQQQLAIPNWEQAMRL				
	230	240	250	260	270

```

# 1: ILV1_TOBAC
# 2: ILVB_ARATH
# Matrix: EPAM60
# Gap_penalty: 12
# Extend_penalty: 2
#
# Length: 683
# Identity: 520/683 (76.1%)
# Similarity: 575/683 (84.2%)
# Gaps: 29/683 ( 4.2%)
# Score: 3275

```

Sequence alignment between ILV1_T and ILVB_A with a higher gap penalty (12). The alignment highlights identical residues with colons and similar residues with dots. Orange boxes indicate positions where gaps are present in one sequence.

	10	20	30	40
ILV1_T	MAAAAESP-PSAAFSTKLPSSTSSTLLPRSTFPFHHPHKTPPPL			
ILVB_A	MAAAATTTTSSSISFSTKPSPSSSKSP-LISRFSLPFSLNPNKSSS			
	10	20	30	40

	50	60	70	80
ILV1_T	HLTHTHIIHIHSQRRR-FTISNVISTNQKV-SQTE-KTEF			
ILVB_A	-SSRRGIKSSSPSSISAVLNTTNVTTTPSPTKPTKPETF			
	50	60	70	80

	90	100	110	120	130
ILV1_T	VSRFAPDEPRKGSDVLVEALEREGVTDVFAYPGGASMEIHQALTRSSIIR				
ILVB_A	ISRFAPDQPRKGADILVEALERQGVETVFAYPGGASMEIHQALTRSSSIR				
	90	100	110	120	130

	140	150	160	170	180
ILV1_T	NVLPRHEQGGVFAAEQYARATGFPGVCIATSGPGATNLVSGLADALLDSV				
ILVB_A	NVLPRHEQGGVFAAEQYARSSGKPGICIAATSGPGATNLVSGLADALLDSV				
	140	150	160	170	180

	190	200	210	220	230
ILV1_T	PIVAITGQVPRRMIGTDQETPIVEVTRSITKHNYLVMDVEDIPRVVRE				
ILVB_A	PLVAITGQVPRRMIGTDQETPIVEVTRSITKHNYLVMDVEDIPRIIEE				
	190	200	210	220	230

	240	250	260	270	280
ILV1_T	AFFLARSGRGPGLIDVPKDIQQQLVIPDWDPQPMRLPGYMSRLPKLPNEM				
ILVB_A	AFFLATSGRGPVPLVDVPKDIQQQLAIPNWEQAMRLPGYMSRMPKPPEDS				
	240	250	260	270	280

EFFET DU CHOIX DE LA MATRICE DE SUBSTITUTION

```
# Aligned_sequences: 2
# 1: PDC1_MAIZE
# 2: ILVB_ARATH
Matrix: EBLOSUM62
# Gap_penalty: 12
# Extend_penalty: 2
#
# Length: 692
# Identity: 133/692 (19.2%)
# Similarity: 244/692 (35.3%)
# Gaps: 104/692 (15.0%)
# Score: -14
```

10				
PDC1_M	METLLAG-----	-NPANGVAKPT		
:		:: .		
ILVB_A	MAAATTTTTSSSISFSTKPS	SSKSPLPISRFSLPFSLNPNKSSSSR		
10	20	30	40	

20	30	40	50	
PDC1_M	CNGVGALPVANSHAIATPAAAATLAPAGAT	-----LGRH-----		
: : : . : . : . :			
ILVB_A	RRGIKSSSPSSISAVLNTTNVTTPSPTKPTKPETFISR	FAPDQPRKG		
60	70	80	90	100

60	70	80	90	100	
PDC1_M	--LARRLVQIGASDVFAVPGDFNLTLDY	LIAEPGLTLVGCCNELNAGYA			
:	.. : . : : : . . : . : . : . :				
ILVB_A	DILVEALERQGVETVFAYPGGASMEIHQALTRSSSIRNVLP	RHEQGGVFA			
110	120	130	140	150	

110	120	130	140	150	
PDC1_M	ADGYARSRGV-GACAVTFTVGG	GLSVLNIAIAGAYSENLPVVCIVGGPNSD			
:	::::: : : : : : :				
ILVB_A	AEGYARSSGKPGICIATSGPGATNLV	SGLADALLSVPLVAITGQVPRRM			
160	170	180	190	200	

160	170	180	190		
PDC1_M	YGTNRILHHTIGLPDFSQELRCFQT	ITCYQAIINNLDDAHEQIDTA--IA			
:	.. : : . . : . :				
ILVB_A	IGTDAFQETPI-----	EVTRSITKHNYLVMVEDIPRIIEAFFLA			
210	220	230	240		

200	210	220	230	240	
PDC1_M	TALRESKPVYISVSCNLAG-LSHPTFS	--RDPVPMFISPRLSNKANLEY			
:	.. : . : . : : :				
ILVB_A	TSGRPG-PVLVDVPKDIQQQLAIPNWEQAMRLPGYMSRMPKPPEDSHLEQ				
250	260	270	280	290	

```
# Aligned_sequences: 2
# 1: PDC1_MAIZE
# 2: ILVB_ARATH
Matrix: EPAM30
# Gap_penalty: 12
# Extend_penalty: 2
#
# Length: 797
# Identity: 173/797 (21.7%)
# Similarity: 216/797 (27.1%)
# Gaps: 314/797 (39.4%)
# Score: -977
```

10	20	30		
PDC1_M	ME---TLLAGNPANGVAKPT-CNGVGALPVA-----NSH-----			
:	: : :: :: . :			
ILVB_A	MAAATTTTTSSSISFSTKPS	SSKSPLPISRFSLPFSLNPNKSSSSR		
20	30	40	50	

Ici on ne serait lequel choisir !

40	50			
PDC1_M	-----AIIATPAAAATLAPAGAT-----LGRHLA---RR-			
:	.. : : : : : : : : : :			
ILVB_A	RRGIKSSSPSSISAVLNTTNVTTPSPTKPTKPETFISR-FAPDQPRKG			
60	70	80	90	

60	70	80	90	100	
PDC1_M	--LVQI---GASDVFAVPGDFNLTLDY	LIAEPGLTLVGCCNELNAGY			
:	.. : : : : . . : : : :				
ILVB_A	ADILVEALERQGVETVFAYPGGASMEIHQALTRSSSIRNVLP	RHEQGGVFA			
100	110	120	130	140	

110	120	130	140		
PDC1_M	AADGYARSRG-VGACAVTFTVGG	GLSVLNIAIAGAYSENLPVVCIVGGPNS			
:	::::: : : : : :				
ILVB_A	AAEGYARSSGKPGICIATSGPGATNLV	SGLADALLSVPLVAI-----			
150	160	170	180	190	

150	160	170	180	
PDC1_M	DYGTNRILHHTIGLPDFSQELRCFQT	--ITCYQAI--NNL---DDA		
:	.. : : : : . . : : . . :			
ILVB_A	--TGQVPRRMIGTDAF-QE-----	TPIVEVT--RSITKHNYLVMVEDI		
200	210	220	230	

190	200	210	220	
PDC1_M	HEQIDTA--IATALRESKPVYISVSCN	--LA-----GLSHPTF-SRD		
:	.. : : : : :			
ILVB_A	PRIIEEAFFLATSGRPG-PVLVDVPKDIQQQLAIPNWEQAMRLPGYMSR			
240	250	260	270	

EFFET DU CHOIX DE LA MATRICE DE SUBSTITUTION

```
# Aligned_sequences: 2
# 1: PDC1_MAIZE
# 2: ILVB_ARATH
# Matrix: EBLOSUM62
# Gap_penalty: 12
# Extend_penalty: 2
#
# Length: 692
# Identity: 133/692 (19.2%)
# Similarity: 244/692 (35.3%)
# Gaps: 104/692 (15.0%)
# Score: -14
```

			10		
PDC1_M	METLLAG-----		NPANGVAKPT		
:	:		::	.	
ILVB_A	MAAATTTTSSSISFSTKPSPLPISRLFSLNPNSKSSSSR				
	10	20	30	40	50
	20	30	40	50	
PDC1_M	CNGVGALPVANSHAIATPAAAAATLAPAGAT-----LGRH-----				

ILVB_A	RRGIKSSSPSSISAVLNNTTNVTTPSPTKPKTFISRFA	P	DQPRKGA		

			10		
PDC1_M	--LARR-----				
:	:				
ILVB_A	DILVEALERQGVETVFAYPGGASMEIHQALTRSSSIRNVLP	RHEQGGVFA			
	110	120	130	140	150
	110	120	130	140	150
PDC1_M	ADGYARSRGV-GACAVTFTVGGLSVLNAIAGAYSENLPVVCIVGGPNSND				

ILVB_A	AEGYARSSGKPGICIATSGPGATNLVSGLADALLDSVPLVAITGQVPRRM				
	160	170	180	190	200
	160	170	180	190	
PDC1_M	YGTNRILHHTIGLPDFSQEELRCFQTITCYQAIINNLDAAHEQIDTA--IA				

ILVB_A	IGTDAFQETPI-----VEVTRSITKHNYLVMDVEDIPRIIEAFFLA				
	210	220	230	240	
	200	210	220	230	240
PDC1_M	TALRESKPVYISVSCNLAG-LSHPTFS--RDPVPMFISPRLSNKANLEY				

ILVB_A	TSGRPG-PVLVDVPKDIQQQLAIPNWEQAMRLPGYMSRMPKPPEDSHLEQ				
	250	260	270	280	290

```
# Aligned_sequences: 2
# 1: PDC1_MAIZE
# 2: ILVB_ARATH
# Matrix: EPAM350
# Gap_penalty: 12
# Extend_penalty: 2
#
# Length: 700
# Identity: 133/700 (19.0%)
# Similarity: 360/700 (51.4%)
# Gaps: 120/700 (17.1%)
# Score: 396
```

			10		
PDC1_M	METLLAGNPANGV-----AKPT-CNGVGALPVAN-----				
:	:	
ILVB_A	MAAATTTTSSSISFSTKPSPLPISRLFSLNPNSKSSSSR				
	10	20	30	40	50
	30	40	50		
PDC1_M	-----SHAIATPAAAAATLAPAGAT-----LGRH-----				

ILVB_A	RRGIKSSSPSSISAVLNNTTNVTTPSPTKPKTFISRFA	P	DQPRKGA		

			100		
ILVB_A	DILVEALERQGVETVFAYPGGASMEIHQALTRSSSIRNVLP	RHEQGGVFA			
	110	120	130	140	150
PDC1_M	ADGYARSRG-VGACAVTFTVGGLSVLNAIAGAYSENLPVVCIVGGPNSND				

ILVB_A	AEGYARSSGKPGICIATSGPGATNLVSGLADALLDSVPLVAITG				
	160	170	180	190	
PDC1_M	YGTNRILHHTIGLPDFSQE--LRCFQTITCYQAIINNLDAAHEQIDTA--IA				

ILVB_A	QVPRRMIGTDAFQETPIVEVTRSITKHNYLVMDVEDIPRIIEAFF				
	200	210	220	230	240
PDC1_M	IATALRESKPVYISVSCNLAG-LSHPTFSRD-PVPMFISPRLSNKANLEY				

ILVB_A	LATSGRPG-PVLVDVPKDIQQQLAIPNWEQAMRLPGYMS-RMPKPPE-DS				
	250	260	270	280	

Ici on gagne nettement en similarité notamment en Nter

Le reste de la séquence semble assez comparable entre les deux matrices choisies

ALIGNEMENT GLOBAL VERSUS ALIGNEMENT LOCAL

```
# Aligned_sequences: 2
# 1: frag_new
# 2: ILV1_TOBAC
# Matrix: EBLOSUM45
# Gap_penalty: 12
# Extend_penalty: 2
#
# Length: 667
# Identity: 40/667 ( 6.0%)
# Similarity: 56/667 ( 8.4%)
# Gaps: 576/667 (86.4%)
# Score: -1062
```

```
frag_n M-----ETLL-----
:       :::
ILV1_T MAAAAPSPSSSAFSKTLSPSSSTSILLPRSTFPFPFHPHKTPPPLHLT
10      20      30      40      50
```

```
frag_n -----
ILV1_T HTHIHIHSQRRTFTISNVISTNQKVQTEKTETFVSRFAPDEPRKGSDVL
60      70      80      90      100
```

```
frag_n -----
ILV1_T VEALEREGVTDFAYPGGASMEIHQLTRSSIIRNVLPKHEQGGVFAAEG
110     120     130     140     150
```

```
10
frag_n ---AGNPA----NGVS-----IG-
:       :   :::::       :::
ILV1_T YARATGPGVCIATSGPGATNLVSGLADALLDSVPIVAITGQVPRRMIGT
160     170     180     190     200
```

```
frag_n -----
ILV1_T DAFQETPIVEVTRSITKHNLYLVMVEDIPRVVREAFFLARSGRGPILID
210     220     230     240     250
```

```
frag_n -----WS-----
: 
ILV1_T VPKDIOQQQLVIPDWDQPMRRLPGYMSRLPKLPNEMLLEQIVRLISESKKPV
260     270     280     290     300
```

```
20          30
frag_n -----VGATLGYAGAV-----
:   :::   :::
ILV1_T LYVGGGCSQSSEDLRRFVELTGIPVASTLMGLGAFPTGDELSLSMLGMHG
310     320     330     340     350
```

```
40          50
frag_n TTFCAEIVESADAYLFAGPPIFND-----
:   .   :::::   :   :   :::
ILV1_T TVYANYAVDSSLALLAFGVRFDDRTVGKLEAFASRAKIVHIDIDSAEIGK
360     370     380     390     400
```

```
frag_n -----YSSWQEN-----
:   :::::
ILV1_T NKQPHVSICADIKLALQGLNSILESKEGKLKLDFAWRQELTEQVKVHPL
410     420     430     440     450
```

```
frag_n -----DQCP--Y-----RT
:   :   :
ILV1_T NFKTFGDAIPPQYAIQVLDELTNGNAIISTGVGQHQMWAAQYYKYRKPRQ
460     470     480     490     500
```

```
70
frag_n W-----HITSITT---
: 
ILV1_T WLTSGLGAMGFLPAAIGAAVGRPDEVVVDIDGDGSFIMNVQELATIKV
510     520     530     540     550
```

```
80
frag_n -----NDYAHV-----EAB-----CK
:   :::   :::::   :::
ILV1_T ENLPVKIMLLNNQHLGMVVQWEDRFYKANRAHTYLGPNPSNEAEIFPNMLK
560     570     580     590     600
```

```
90
frag_n F-----ERME-----
: 
ILV1_T FAEACGVPAARVTHRDDLRAAIQKMLDTPGPYLLDVIVPHQEHVLPMIPS
610     620     630     640     650
```

```
frag_n -----
ILV1_T GGAFKDVITEGDRSSY
660
```

ALIGNEMENT GLOBAL VERSUS ALIGNEMENT LOCAL

```
# Aligned_sequences: 2
# 1: frag_new
# 2: ILV1_TOBAC
# Matrix: EBLOSUM45
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 97
# Identity:      25/97 (25.8%)          Frag_n : 83 aa
# Similarity:    37/97 (38.1%)          ILV1_T : 667 aa
# Gaps:          16/97 (16.5%)
# Score:         72.5
#
#
#=====
#
#          10          20          30          40
frag_n LAGNPANGVSIGWSVGA-----TLGYAGAVSTTFCAEIVESADAYLFA
        : : : . : . : . : : . : . : . : . : . : . : . : . : .
ILV1_T LTGIPVASTLMG--LGAFPTGDELSSLGMLGMHGTVYANYAVDSSDLLLAF
        320        330        340        350        360
#
#          50          60          70          80
frag_n GPIFNDYSSWQ-ENDQCPYRTWHI----TSITTNDYAHVE--ABCKF
        : : : . . : . . : . : : . : : . : . : . : .
ILV1_T GVRFDDDRVGTGKLEAFASRAKIVHIDIDS A E I G K N K Q P H V S I C A D I K L
        370        380        390        400        410
```



PROBLÈMES ET LIMITES DE LA RECHERCHE DE SIMILITUDES

- **Les gènes inconnus**

Quand un gène ne ressemble à aucun autre, on le dit "**orphelin**". Quand le génome de la levure a été obtenu, près de la moitié de ses gènes n'avaient pas d'homologues connus dans les banques.

- **Les erreurs**

Les informations présentes dans les banques peuvent être erronées, **il est indispensable de vérifier attentivement les résultats.**



PROBLÈMES ET LIMITES DE LA RECHERCHE DE SIMILITUDES

○ Les gènes homologues : orthologues et paralogues

Une fois une certaine similitude mise en évidence, il est nécessaire de séparer les gènes **orthologues** des **paralogues**.

- Quand le gène est transmis à deux espèces filles : ils sont **orthologues**.
- Il est fréquent que certains gènes se dupliquent. Un exemplaire du gène conserve généralement sa fonction première, le ou les autres (ce sont les **paralogues**) peuvent évoluer indépendamment et acquérir des fonctions complètement différentes.
- Seule une analyse de leur évolution via la **construction d'arbres phylogénétiques** permet de différencier ces deux cas.

PROBLÈMES ET LIMITES DE LA RECHERCHE DE SIMILITUDES

○ Le "bricolage de l'évolution"

Une autre difficulté de la recherche de fonctions provient des réarrangements qui s'opèrent lors des étapes séparant le gène de la protéine fonctionnelle :

- **L'épissage alternatif** : pour un même gène et dans un même organisme, l'élimination des introns peut être différente selon la cellule concernée. Ainsi, pour un même gène, l'ARNm sera différent et donnera naissance à une protéine différente.
- Par ailleurs, l'association de fragments provenant de gènes différents permet l'émergence de fonctions totalement nouvelles (cassettes fonctionnelles).



PROBLÈMES ET LIMITES DE LA RECHERCHE DE SIMILITUDES

La maturation post-traductionnelle de la protéine

- Les protéines vont migrer grâce à des signaux d'adressage spécifiques vers les mitochondries, les lysosomes, les membranes...
- Elles peuvent aussi traverser le réticulum endoplasmique et passer par l'appareil de Golgi pour être secrétées dans le milieu extracellulaire.
- Une fois traduite, la protéine peut subir une maturation post-traductionnelle (**glycosylation, hydroxylation, ...**) les modifiant profondément, de telle sorte que la protéine finale est bien différente de la molécule directement codée par le génome.



PROBLÈMES ET LIMITES DE LA RECHERCHE DE SIMILITUDES

Pour toutes ces raisons, les résultats produits par les logiciels ne constituent que des hypothèses qui doivent être vérifiées par une démarche expérimentale en laboratoire. Notamment par observation des effets de l'altération ou de la délétion du gène dans l'organisme.



LES PRINCIPAUX SERVEURS BLAST

- **NCBI**

<http://www.ncbi.nlm.nih.gov/BLAST/>

Le plus souvent utilisé mais aux USA (donc risque d'encombrement)

- **EBI**

<http://www.ebi.ac.uk/blast/>

Blast-Wu, développement un peu différent du NCBI, paramètres différents mais en Europe.

- **SwissProt**

<http://www.expasy.org/tools/blast/>

Dédié aux protéines essentiellement (BlastP)

Chaque serveur a son **propre Blast avec ses propres paramètres et différents choix de bases de données**. Il est souvent utile (nécessaire) de comparer les résultats entre les serveurs pour affirmer/infirmer des hypothèses.

CE QUE VOUS ALLEZ VOIR SUR UN SERVEUR BLAST OUTPUT NCBI

 **BLAST®** *Basic Local Alignment Search Tool*

Home Recent Results Saved Strategies Help

► [NCBI/BLAST Home](#)

BLAST finds regions of similarity between biological sequences. [more...](#)

New [DELTABLAST](#), a more sensitive protein-protein search [Go](#)

BLAST Assembled RefSeq Genomes

Choose a species genome to search, or [list all genomic BLAST databases](#).

<input type="checkbox"/> Human	<input type="checkbox"/> Oryza sativa	<input type="checkbox"/> Gallus gallus
<input type="checkbox"/> Mouse	<input type="checkbox"/> Bos taurus	<input type="checkbox"/> Pan troglodytes
<input type="checkbox"/> Rat	<input type="checkbox"/> Danio rerio	<input type="checkbox"/> Microbes
<input type="checkbox"/> Arabidopsis thaliana	<input type="checkbox"/> Drosophila melanogaster	<input type="checkbox"/> Apis mellifera

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a nucleotide database using a nucleotide query <i>Algorithms:</i> blastn, megablast, discontiguous megablast
protein blast	Search protein database using a protein query <i>Algorithms:</i> blastp, psi-blast, phi-blast, delta-blast
blastx	Search protein database using a translated nucleotide query
tblastn	Search translated nucleotide database using a protein query
tblastx	Search translated nucleotide database using a translated nucleotide query

Choix du BLAST



CE QUE VOUS ALLEZ VOIR SUR UN SERVEUR BLAST OUTPUT NCBI

BLAST Basic Local Alignment Search Tool My NCBI [Sign In] [Register]

Home Recent Results Saved Strategies Help

NCBI/BLAST/blastp suite: BLASTP programs search protein databases using a protein query. more... Reset page Bookmark

Enter Query Sequence

Enter accession number, gi, or FASTA sequence Clear Query subrange
From To Séquence requête

Or, upload file Parcourir... Job Title
Enter a descriptive title for your BLAST search

Choose Search Set

Database Non-redundant protein sequences (nr) Choix de la base de données

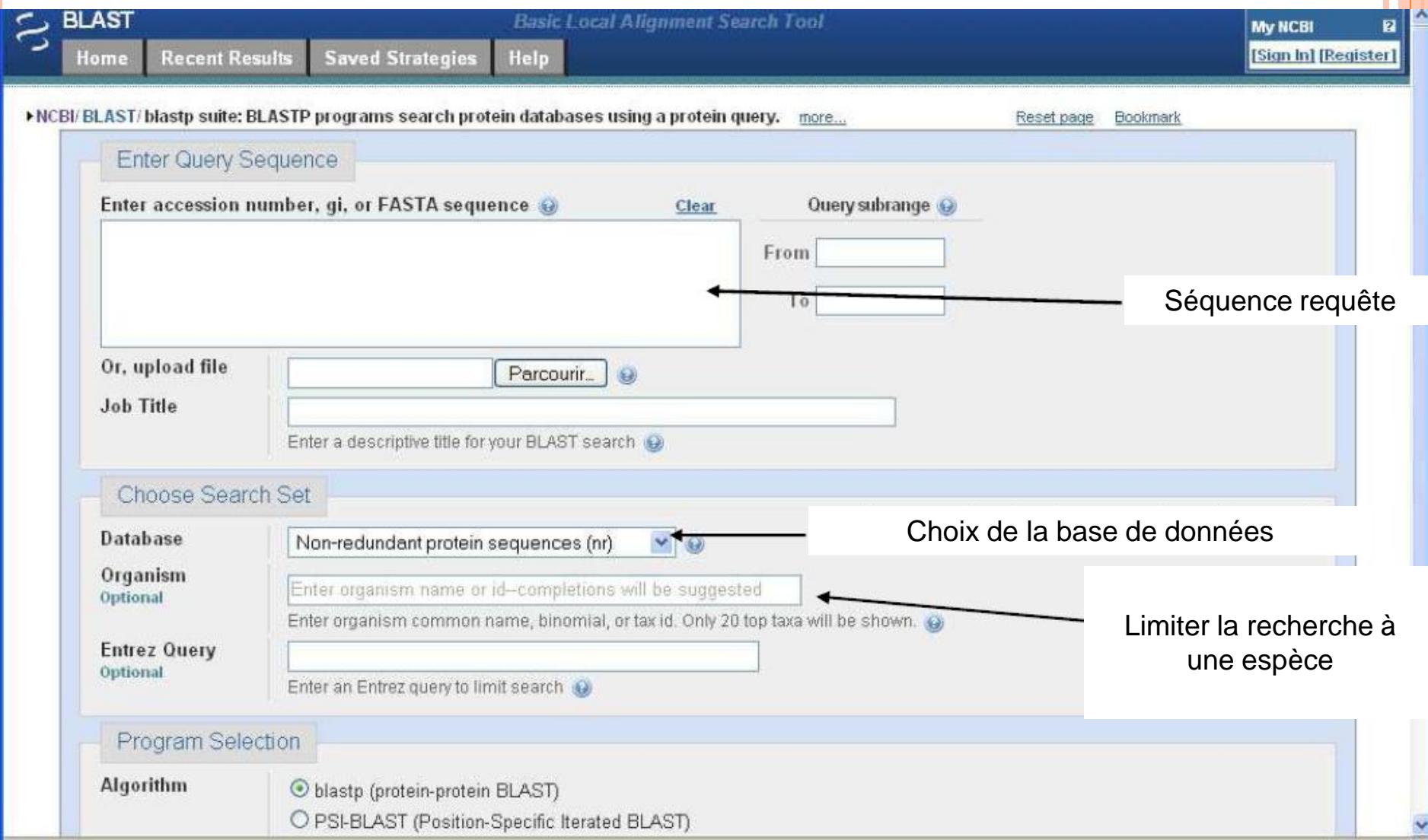
Organism Optional Enter organism name or id—completions will be suggested
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Entrez Query Optional Enter an Entrez query to limit search

Program Selection

Algorithm blastp (protein-protein BLAST)
 PSI-BLAST (Position-Specific Iterated BLAST)

Limiter la recherche à une espèce



CE QUE VOUS ALLEZ VOIR SUR UN SERVEUR BLAST OUTPUT NCBI

The screenshot shows the NCBI BLAST search interface. At the top left, there is a field labeled "Entrez Query Optional" with a placeholder "Enter an Entrez query to limit search". Below it is a "Program Selection" section. Under "Algorithm", three options are listed: "blastp (protein-protein BLAST)" (selected), "PSI-BLAST (Position-Specific Iterated BLAST)", and "PHI-BLAST (Pattern Hit Initiated BLAST)". A link "Choose a BLAST algorithm" is also present. In the center, there is a "BLAST" button and a search input field "Search database nr using Blastp (protein-protein BLAST)". A checkbox "Show results in a new window" is available. At the bottom left, a link "Algorithm parameters" is circled in green. The footer contains links for "Copyright | Disclaimer | Privacy | Accessibility | Contact | Send feedback on new interface" and the NCBI logo.

Attention, pour accéder aux différents paramètres, il faut cliquer sur « Algorithm parameters »

CE QUE VOUS ALLEZ VOIR SUR UN SERVEUR BLAST OUTPUT NCBI

▼ Algorithm parameters

General Parameters

Max target sequences: 100 Select the maximum number of aligned sequences to display

Short queries: Automatically adjust parameters for short input sequences

Expect threshold: 10 Limite de la e-value

Word size: 3 Taille W du mot m

Scoring Parameters

Matrix: BLOSUM62 Choix de la matrice de substitution
Gap Costs: Existence: 11 Extension: 1 Et gestion des indels

Compositional adjustments: Composition-based statistics

Filters and Masking

Filter: Low complexity regions Filtre pour les séquences de faibles complexités

Mask: Mask for lookup table only
 Mask lower case letters

BLAST

Search database nr using Blastp (protein-protein BLAST)
 Show results in a new window

CE QUE VOUS ALLEZ VOIR SUR UN SERVEUR BLAST OUTPUT NCBI

BLAST Basic Local Alignment Search Tool My NCBI [Sign In] [Register]

Home Recent Results Saved Strategies Help

NCBI/BLAST/blastp/Formatting Results - ECNBNP0F01R [Formatting options]

Job Title: Protein sequence(261 letters)

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

Request ID: ECNBNP0F01R
Status: Searching
Submitted at: Tue Sep 11 04:44:37 2007
Current time: Tue Sep 11 04:44:44 2007
Time since submission:

This page will be automatically updated in 10 seconds



CE QUE VOUS ALLEZ VOIR SUR UN SERVEUR BLAST OUTPUT NCBI

BLAST Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

NCBI/BLAST/blastp/Formatting Results - ECNBNP0F01R [Reformat these Results] [Edit and Resubmit] [

Job Title: Protein sequence(261 letters)

BLASTP 2.2.17 (Aug-26-2007)

Reference:
Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer, 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.

Reference:
Schäffer, Alejandro A., L. Aravind, Thomas L. Madden, Sergei Shavirin, John L. Spouge, Yuri I. Wolf, Eugene V. Koonin, and Stephen F. Altschul (2001), "Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements", Nucleic Acids Res. 29:2994-3005.

RID: ECNBNP0F01R

Database: All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects

5,459,000 sequences; 1,890,632,936 total letters

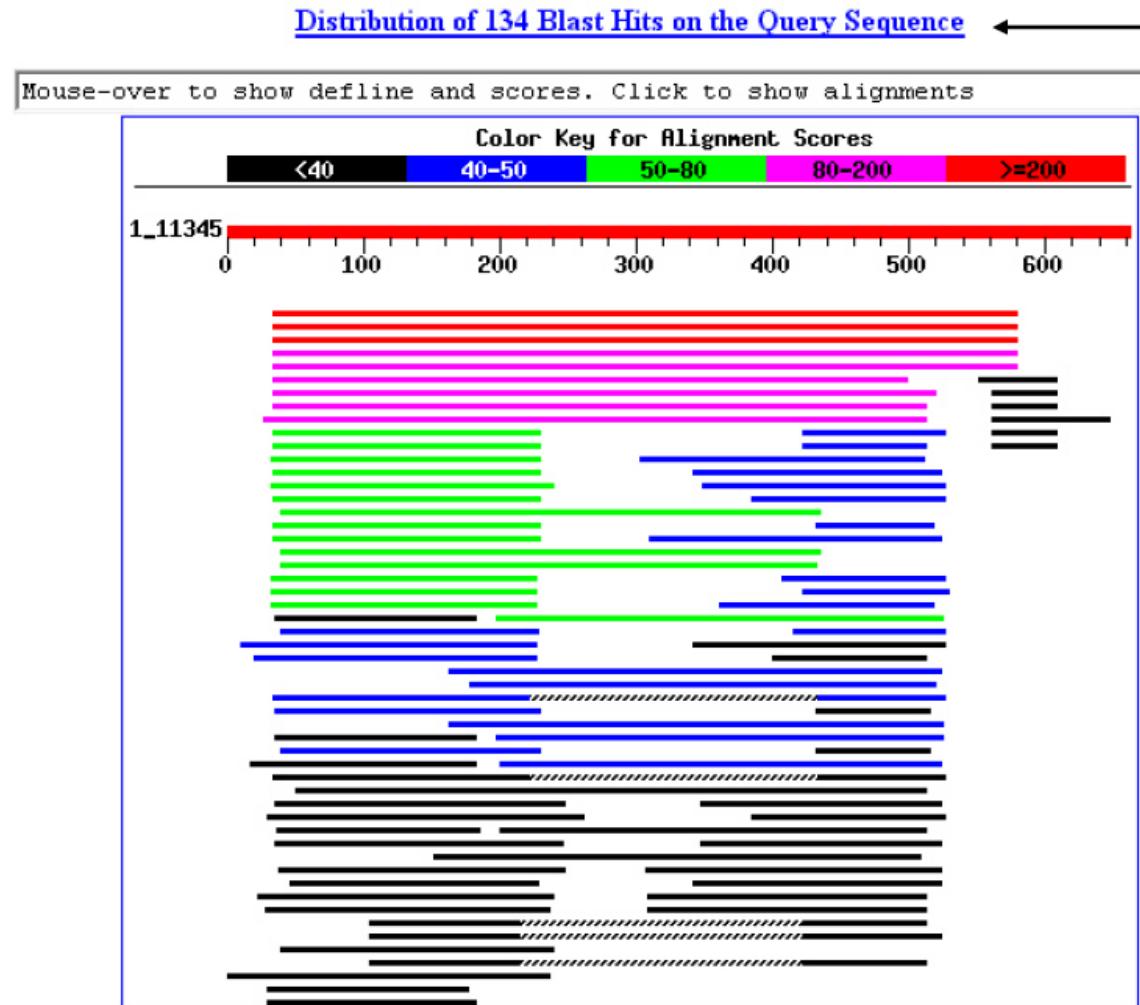
If you have any problems or questions with the results of this search please refer to the [BLAST FAQs](#)
[Taxonomy reports](#)

Query= Length=261

Banques de données choisies

Séquence requête

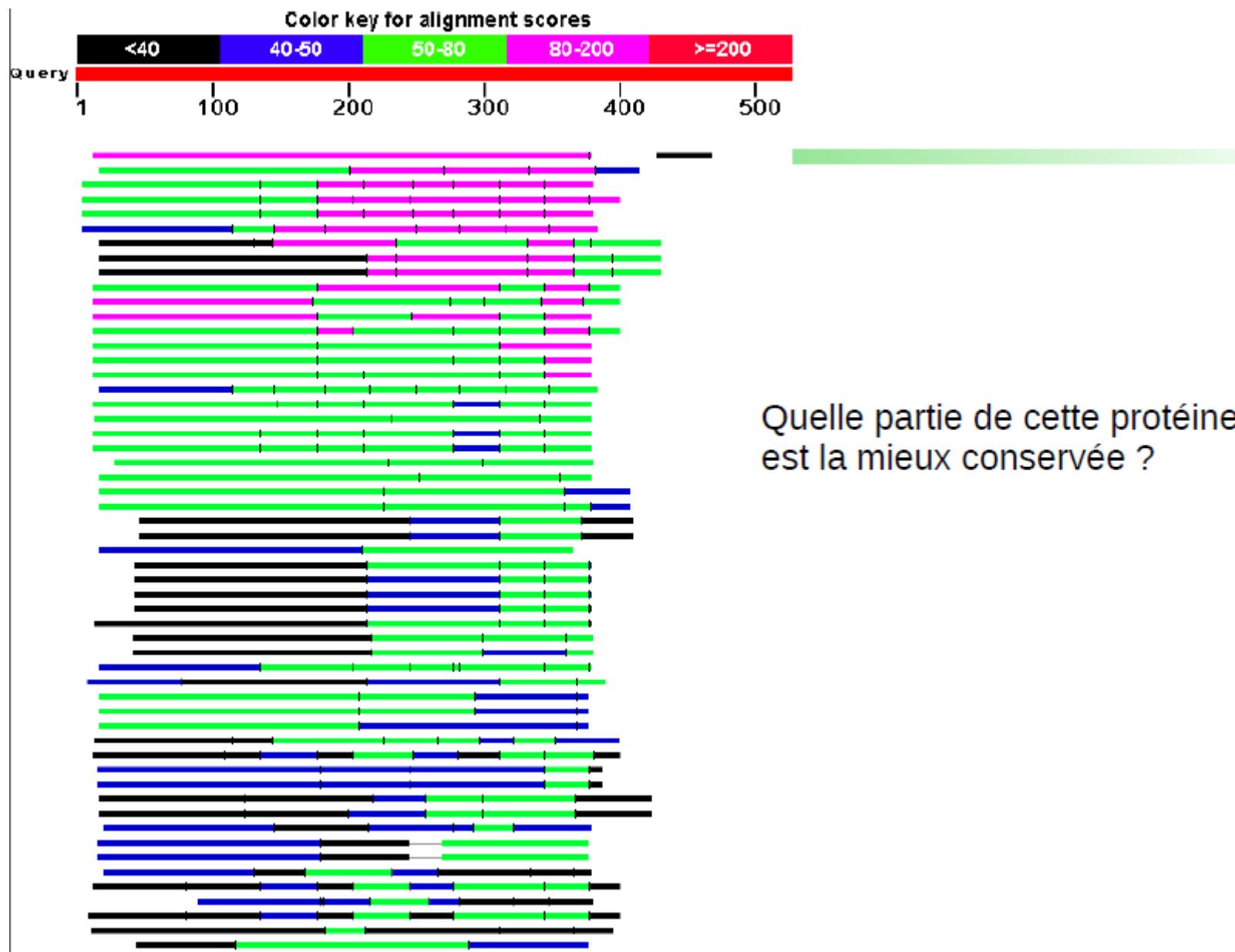
CE QUE VOUS ALLEZ VOIR SUR UN SERVEUR BLAST OUTPUT NCBI



Nombres de hits

Répartition des hits en fonction de leur score

Vision du recouvrement des différents HSP



Quelle partie de cette protéine
est la mieux conservée ?



CE QUE VOUS ALLEZ VOIR SUR UN SERVEUR BLAST OUTPUT NCBI

Une forte valeur de la E value indiquerait que le résultat pourrait être du au hasard

Sequences producing significant alignments:

[swissprot:CTRB_HUMAN](#) Chymotrypsinogen B precursor (EC 3.4.21.1).
[swissprot:CTR2_CANFA](#) Chymotrypsinogen 2 precursor (EC 3.4.21.1).
[swissprot:CTRB_RAT](#) Chymotrypsinogen B precursor (EC 3.4.21.1).
[swissprot:CTRB_BOVIN](#) Chymotrypsinogen B (EC 3.4.21.1).
[swissprot:CTRA_BOVIN](#) Chymotrypsinogen A (EC 3.4.21.1).
[swissprot:CTRA_GADMO](#) Chymotrypsin A precursor (EC 3.4.21.1).

Score (bits)	E Value
433	e-121
386	e-107
383	e-106
348	4e-96
330	1e-90
286	2e-77

Le lien vers l'entrée de la base de données qui a été utilisée

Un score élevé, ou mieux une série de scores élevés, suggère une relation mais à vérifier en regardant l'alignement

CE QUE VOUS ALLEZ VOIR SUR UN SERVEUR BLAST OUTPUT NCBI

swissprot:CO2_HUMAN	Complement C2 precursor (EC 3.4.21.43) (C3/C...)	<u>55</u>	1e-07
swissprot:CO2_MOUSE	Complement C2 precursor (EC 3.4.21.43) (C3/C...)	<u>53</u>	3e-07
swissprot:ACH2_LONAC	Achelase II protease (EC 3.4.21.-).	<u>52</u>	1e-06
swissprot:GD_DROME	Serine protease gd precursor (EC 3.4.21.-) (G...)	<u>46</u>	4e-05
swissprot:ACRO_CAPHI	Acrosin (EC 3.4.21.10) (Fragment).	<u>39</u>	0.009
swissprot:CTRP_PENMO	Chymotrypsin (EC 3.4.21.1) (Fragment).	<u>36</u>	0.047
swissprot:VSPA_CERCE	Cerastotin (EC 3.4.21.-) (Fragments).	<u>35</u>	0.098
swissprot:EL2B_HORSE	Neutrophil elastase 2B (EC 3.4.21.-) (Prote...	<u>35</u>	0.13
swissprot:CERC_SCHMA	Cercarial protease precursor (EC 3.4.21.-) ...	<u>34</u>	0.26
swissprot:EL2A_HORSE	Neutrophil elastase 2A (EC 3.4.21.-) (Prote...	<u>33</u>	0.42
swissprot:HPT_RABBIT	Haptoglobin beta chain (Fragment).	<u>31</u>	1.4
swissprot:NMT1_ASPPA	NMT1 protein homolog.	<u>30</u>	4.8

Un score faible, avec une forte E value, suggère fortement que la similitude entre les séquences est le résultat du hasard

CE QUE VOUS ALLEZ VOIR SUR UN SERVEUR BLAST OUTPUT NCBI

```
>swissprot:VSP5_TRIMU Mucrofibrase 5 precursor (EC 3.4.21.-).
```

```
Length = 257
```

Score = 103 bits (280), Expect = 3e-22

Identities = 74/232 (31%), Positives = 110/232 (46%), Gaps = 10/232 (4%)

```
Query: 34 IVNGEEAVPGTWPWQVTLQDRSGFHFCGGSLISEDWVVTAAHCGVRTSEILIAGEFDQGS 93
       I+ G+E      P+ V +      + CGG+LI+E+WV+TAAHC      EI +      +
Sbjct: 25 IIGGDECNINEHPFLVLVYYDD--YQCGGTLINEEWVL TAAHCNGENMEIYLGGMHSKKVP 82
```

```
Query: 94 DEDNIQVLRIAKVFQPKYSILTVNNDITLLKLASPARYSQTISAVCLPSVDDDAGSLCA 153
       ++D + + K F + N DI L++L P R S I+ + LPS GS+C
Sbjct: 83 NKDRRRRVPKEKFFCDSSKNYTKWNKDIMLIRLNRPVRKSAHIAPLSLPSSPPSVGSVCR 142
```

```
Query: 154 TTGWGRTKYNANKSPDKLERAALPLLTNAECKRSW-GRRLTDVMICGA--ASGVSSCMGD 210
       GWG          PD      A + LL      C+ ++ G      T      +C      G      SC GD
Sbjct: 143 IMGWGTISPTKVTLPDVPRCANINLLDYEV CRAAYAGLPATSRTL CAGILEGGKDSCGGD 202
```

```
Query: 211 SGGPLVCQKD GAYTLVAIVSWASDTCS-ASSGGVYAKVT KIIIPWVQKILSSN 261
       SGGPL+C +G + IVSW D C+ G+Y V + W++ I++ N
```

```
Sbjct: 203 SGGPLIC--NGQFQ--GIVSWGGDPCAQPHEPGLYT NVFDHLDWIKGIIAGN 250
```

CE QUE VOUS ALLEZ VOIR SUR UN SERVEUR BLAST OUTPUT NCBI

```
>swissprot:CTRB_HUMAN Chymotrypsinogen B precursor (EC 3.4.21.1).  
Length = 263
```

Score = 103 bits (1222), Expect = e-121
Identities = 220/263 (83%). Positives = 252/263 (95%), Gaps = 2/263 (0%)

Query: 1 MAFIWLSCYALLGTTFGCGVNAIHPVLTGLSKIVNGEEAVPGTWPMQVTLQDRSGFHFC 60
MAF+WLLSC+ALLGTTFGCGV AIHPVLT+GLS+IVNGE+AVPG+WPMQV+LQD++GFHFC

Sbjct: 1 MAFIWLSCWALLGTTFGCGVPAIHPVLSGLSRIVNGEDAVGSWPWQVSLQDKTGFHFC 60

Query: 61 GGSLISEDWWVTAAHCGVRTSE ILIAGEFDQGSDEENIQVLRIAKVFKQPKYSILTVNND 120
GGSLISEDWWVTAAHCGVRTS++++AGEFDQGSDE+NIQVL+IAKVFK PK+SILTVNND

Sbjct: 61 GGSLISEDWWVTAAHCGVRTSDVVVAGEFDQGSDEENIQVLKIAKVFKNPKFSILTVNND 120

Query: 121 ITLLKLASPARYSQTISAVCLPSVDDD--AGSLCATTGWGRTKYNANKSPDKLERAALPL 178
ITLLKLA+PAR+SQT+SAVCLPS DDD AG+LCATTGWG+TKYNANK+PDKL++AALPL

Sbjct: 121 ITLLKLATPARFSQTVSAVCLPSADDFPAGTL CATTGWGKTKYNANKTPDKLQQAALPL 180

Query: 179 LTNAECKRSWGRRLTDVMICGAASGVSSCMGDSGGPLVCQKDGA+TLV AIVSWASDTCSA 238
L+NAECK+SWGRR+TDVMIC ASGVSSCMGDSGGPLVCQKDGA+TLV IVSW SDTCS

Sbjct: 181 LSNAECKKSWGRRITDVMICAGASGVSSCMGDSGGPLVCQKDGA+TLV GIVSWGSDTCST 240

Query: 239 SSGGVYAKVTKIIPWVQKILSSN 261
SS GVYA+VTK+IPWVQKIL++N

Sbjct: 241 SSPGVYARVTKLIPWVQKILAAN 263

CE QUE VOUS ALLEZ VOIR SUR UN SERVEUR BLAST OUTPUT NCBI

```
>swissprot:HPT_RABIT Haptoglobin beta chain (Fragment).
```

Length = 40

Longueurs des query

Score = 31.3 bits (74), Expect = 1.4

Identities = 15/41 (36%), Positives = 22/41 (53%), Gaps = 1/41 (2%)

Query: 34 IVNGEEAVPGTWPWQVTLQDRSGFHFCGGSLISEDWVVTAA 74

I+ G G++PWQ + R G +LISE W++T A

Sbjct: 1 IIGGSLDAKGSFPWQAKMVSRHNL-VTGATLISEQWLLTTA 40

```
>swissprot:NMT1_ISPPA NMT1 protein homolog.
```

Length = 342

Longueurs des alignements

Score = 29.6 bits (69), Expect = 4.8

Identities = 11/34 (32%), Positives = 22/34 (64%)

Query: 72 TAAHCGVRTSEILIAGEFDQGSDEDNIQVLRIAK 105

TA CG+ ++ +I G+ D G +N+Q++ +A+

Sbjct: 137 TAVRCGMNVTKAIIRGDIAGIGLENVQMVELAE 170

Attention aux pourcentages (ID et Pos) par rapport à la longueur de l'HSP !



CE QUE VOUS ALLEZ VOIR SUR UN SERVEUR BLAST OUTPUT EBI

FASTA

FASTA

FASTA is another commonly used sequence similarity search tool which uses heuristics for fast **local** alignment searching.

Protein Nucleotide Genomes Whole Genome Shotgun

SEARCH

SSEARCH is an optimal (as opposed to heuristics-based) **local** alignment search tool using the Smith-Waterman algorithm. Optimal searches guarantee you find the best alignment score for your given parameters.

Protein Nucleotide Genomes Whole Genome Shotgun

PSI-Search

PSI-Search combines the sensitivity of the Smith-Waterman search algorithm (SSEARCH) with the PSI-BLAST profile construction strategy to find distantly related protein sequences.

Protein

GGSEARCH

GGSEARCH performs optimal **global-global** alignment searches using the Needleman-Wunsch algorithm.

Protein Nucleotide

GLSEARCH

GLSEARCH performs an optimal sequence search using alignments that are **global** in the query but **local** in the database sequence. This can be useful when you want to match all of a short query sequence to part of a larger database sequence.

Protein Nucleotide

FASTM/S/F

These specialist programs allow searches of databases using a group of short peptides as the query.

Protein Nucleotide Proteomes

BLAST

NCBI BLAST

NCBI BLAST (blastall) is the most commonly used sequence similarity search tool. It uses heuristics to perform fast **local** alignment searches.

Protein Nucleotide Vectors

WU-BLAST

WU-BLAST is similar to NCBI BLAST but combines multiple parameter options into a simpler 'sensitivity' setting.

Protein Nucleotide

PSI-BLAST

PSI-BLAST allows users to construct and perform a BLAST search with a custom, position-specific, scoring matrix which can help find distant evolutionary relationships. PHI-BLAST functionality is also available to restrict results using patterns.

Protein

ENA Sequence Search

The EBI has a new search tool which is far faster than BLAST, with only a marginal loss in search sensitivity.

Try it out at [ENA Sequence Search](#).

CE QUE VOUS ALLEZ VOIR SUR UN SERVEUR BLAST OUTPUT EBI

WU-BLAST

Protein Nucleotide Web services Help & Documentation

Tools > Sequence Similarity Searching > WU-BLAST

Nucleotide Database Query

The emphasis of this tool is to find regions of sequence similarity quickly, with minimum loss of sensitivity.

STEP 1 - Select your databases

NUCLEOTIDE DATABASES

110 Databanks Selected X Cle

▼ EMBL-Bank

- EMBL Release
- EMBL Updates
- EMBL Coding Sequence
- Others
- IMGT

STEP 2 - Enter your input sequence

Enter or paste a **DNA/RNA** sequence in any supported format:

Upload a file:

STEP 3 - Set your parameters

PROGRAM
blastn

The default settings will fulfill the needs of most users and, for that reason, are not visible.

More options... (Click here, if you want to view or change the default settings.)



CE QUE VOUS ALLEZ VOIR SUR UN SERVEUR BLAST OUTPUT EBI

WU-BLAST

Protein | Nucleotide | Web services | Help & Documentation

Tools > Sequence Similarity Searching > WU-BLAST

Protein Database Query

The emphasis of this tool is to find regions of sequence similarity quickly, with minimum loss of sensitivity.

STEP 1 - Select your databases

PROTEIN DATABASES

1 Databank Selected Clear Selection

- UniProt Knowledgebase
- UniProtKB/Swiss-Prot
- UniProtKB/Swiss-Prot isoforms
- UniProtKB/TrEMBL
- UniProtKB Taxonomic Subsets
- UniProt Clusters

STEP 2 - Enter your input sequence

Enter or paste a PROTEIN sequence in any supported format:

Upload a file: Parcourir...

STEP 3 - Set your parameters

PROGRAM

The default settings will fulfill the needs of most users and, for that reason, are not visible.

More options... (Click here, if you want to view or change the default settings.)

STEP 4 - Submit your job

Be notified by email (Tick this box if you want to be notified by email when the results are available)

Submit

CE QUE VOUS ALLEZ VOIR SUR UN SERVEUR BLAST OUTPUT EBI

STEP 3 - Set your parameters

PROGRAM	blastp								
MATRIX	blosum62	EXP THR.	10 (default)	FILTER	seg	VIEW FILTER	no	SENSITIVITY	normal
SCORES	50 (default)	ALIGNMENTS	50 (default)	SORT	pvalue	STATS	sump	topcomboN	1
ALIGNMENT VIEWS	pairwise								

CE QUE VOUS ALLEZ VOIR SUR UN SERVEUR BLAST OUTPUT EBI

Summary Table | Tool Output | Visual Output | Functional Predictions | Submission Details

Selection:

Select All | Invert | Clear

Annotations:

Show | Hide

Alignments:

Show | Hide

Entries:

Download in format

Align.	DB:ID	Source	Length	Score	Identities	Positives	E0
<input checked="" type="checkbox"/>	SP:MIS_HUMAN	Muellerian-inhibiting factor OS=Homo sapiens GN=AMH PE=1 SV=3 Cross-references and related information in: ► Gene expression ► Small molecules ► Nucleotide sequences ► Genomes ► Ontologies ► Molecular interactions ► Protein families ► Literature ► Reactions, pathways & diseases ► Protein sequences	560	1466	77.0	77.0	3.1E-217
<input checked="" type="checkbox"/>	TR:G3S0U3_GORGO	Uncharacterized protein OS=Gorilla gorilla gorilla GN=AMH PE=3 SV=1 Cross-references and related information in: ► Genomes ► Ontologies ► Protein families ► Protein sequences	560	1443	76.0	76.0	2.2E-214
<input checked="" type="checkbox"/>	TR:G3S5V4_GORGO	Uncharacterized protein OS=Gorilla gorilla gorilla GN=AMH PE=3 SV=1 Cross-references and related information in: ► Genomes ► Ontologies ► Protein families ► Protein sequences	545	1419	76.0	76.0	7.6E-212
<input checked="" type="checkbox"/>	TR:H2NWW7_PONAB	Uncharacterized protein OS=Pongo abelii GN=AMH PE=3 SV=1 Cross-references and related information in: ► Gene expression ► Nucleotide sequences ► Genomes ► Ontologies ► Protein families ► Protein sequences	561	1309	74.0	75.0	6.9E-198
<input checked="" type="checkbox"/>	TR:F7FPU1_CALJA	Uncharacterized protein OS=Callithrix jacchus GN=LOC100406441 PE=3 SV=1 Cross-references and related information in: ► Nucleotide sequences ► Genomes ► Ontologies ► Protein families ► Protein sequences	557	1261	71.0	75.0	9.4E-192
<input checked="" type="checkbox"/>	TR:H2QEWA1_PANTR	Uncharacterized protein OS=Pan troglodytes GN=ENSG00000104899 PE=3 SV=1 Cross-references and related information in: ► Nucleotide sequences ► Genomes ► Ontologies ► Protein families ► Literature ► Protein sequences	510	1445	76.0	76.0	5.4E-187
<input checked="" type="checkbox"/>	TR:F7FPV1_CALJA	Uncharacterized protein OS=Callithrix jacchus GN=LOC100406441 PE=3 SV=1 Cross-references and related information in: ► Nucleotide sequences ► Genomes ► Ontologies ► Protein families ► Protein sequences	523	1247	72.0	75.0	2.2E-181

CE QUE VOUS ALLEZ VOIR SUR UN SERVEUR BLAST OUTPUT EBI

Align. ♦	DB:ID	Source	Length ♦	Score ♦	Identities ♦	Positives ♦	E0 ♦
<input checked="" type="checkbox"/> 1	SP:MIS_HUMAN	Muellerian-inhibiting factor OS=Homo sapiens GN=AMH PE=1 SV=3 <i>Cross-references and related information in:</i> ► Gene expression ► Small molecules ► Nucleotide sequences ► Genomes ► Ontologies ► Molecular interactions ► Protein families ► Literature ► Reactions, pathways & diseases ► Protein sequences	560	1466	77.0	77.0	3.1E-217
		>SP:MIS_HUMAN P03971 Muellerian-inhibiting factor OS=Homo sapiens GN=AMH PE=1 SV=3 Length = 560 Score = 1466 (521.1 bits), Expect = 3.1e-217, Sum P(2) = 3.1e-217, Group = 1 Identities = 290/373 (77%), Positives = 290/373 (77%) Query: 1 MRDXooooooooooooooXEA LRAEPAVGTSGLIFREDLDWPPGSPQEPLCLVALG 60 MRD EA Sbjct: 1 MRDLPLTSLALVLSALCALLTEALRAEPAVGTSGLIFREDLDWPPGSPQEPLCLVALG 60 Query: 61 GDSNCSSSPRLRVVGALSAYEQAFGLGAVQARWGPRLATFGVCNTGDRQAALPSLRRLCA 120 GDSNCSSSPRLRVVGALSAYEQAFGLGAVQARWGPRLATFGVCNTGDRQAALPSLRRLCA Sbjct: 61 GDSNCSSSPRLRVVGALSAYEQAFGLGAVQARWGPRLATFGVCNTGDRQAALPSLRRLCA 120 Query: 121 WLRDPGGQRQLVVLHLEEVWTWEPTPSLRFQoooooooooooooXVLYPGPGPEVTVTRAG 180 WLRDPGGQRQLVVLHLEEVWTWEPTPSLRFQ VLYPGPGPEVTVTRAG Sbjct: 121 WLRDPGGQRQLVVLHLEEVWTWEPTPSLRFQEP PPPGAGCPPEALLVLYPGPGPEVTVTRAG 180 Query: 181 LPGAQSCLPSRDTTRYLVLAVDRPAGAWRGSCLA LTQPRGEDSRLSTARLQALLFGDDHR 240 LPGAQSCLPSRDTTRYLVLAVDRPAGAWRGSCLA LTQPRGEDSRLSTARLQALLFGDDHR Sbjct: 181 LPGAQSCLPSRDTTRYLVLAVDRPAGAWRGSCLA LTQPRGEDSRLSTARLQALLFGDDHR 240 Query: 241 CFTRMTPALLLPRSEPA PLPAHGQLDTVVFPPPRPSAEELESPPSADPFLETLTXXOOO 300 CFTRMTPALLLPRSEPA PLPAHGQLDTVVFPPPRPSAEELESPPSADPFLETLT Sbjct: 241 CFTRMTPALLLPRSEPA PLPAHGQLDTVVFPPPRPSAEELESPPSADPFLETLT TRLVRA 300 Query: 301 XoooooooooooooooooXGFPQGLVNLSDPAAoooooooooooooDP 360 GFPQGLVNLSDPAA DP Sbjct: 301 LRVPPPARASAPR LALDPDALAGFPQGLVNLSDPAA LERL LDGEEPLL LLRPTAATTGDP 360 Query: 361 APLHDPTSAPWAT 373 APLHDPTSAPWAT Sbjct: 361 APLHDPTSAPWAT 373 Score = 674 (242.3 bits), Expect = 3.1e-217, Sum P(2) = 3.1e-217, Group = 1 Identities = 125/125 (100%), Positives = 125/125 (100%) Query: 436 VEWRGRDP RCGPGRQA QRSAGATA ADGPCALREL SVDLRAERSV LIPETYQANN CQGVCGWP 495 VEWRGRDP RCGPGRQA QRSAGATA ADGPCALREL SVDLRAERSV LIPETYQANN CQGVCGWP Sbjct: 436 VEWRGRDP RCGPGRQA QRSAGATA ADGPCALREL SVDLRAERSV LIPETYQANN CQGVCGWP 495 Query: 496 QSDRNP RYGNHVV LLKM QVRGA ALAR PPCC VPTAY AGKLL ISL SEER ISAH HVPNM VAT 555 QSDRNP RYGNHVV LLKM QVRGA ALAR PPCC VPTAY AGKLL ISL SEER ISAH HVPNM VAT Sbjct: 496 QSDRNP RYGNHVV LLKM QVRGA ALAR PPCC VPTAY AGKLL ISL SEER ISAH HVPNM VAT 555 Query: 556 ECGCR 560 ECGCR Sbjct: 556 ECGCR 560					

La séquence query et la séquence subject sont identiques pourquoi n'obtenons-nous pas 100% d'identité pleine longueur?



Results for job wublast-l20130516-120901-0250-57471242-pg

[Summary Table](#)
[Tool Output](#)
[Visual Output](#)
[Functional Predictions](#)
[Submission De](#)
[Download](#)
[Download in XML format](#)
[Send to DbClustal](#)
[Send to MView](#)

BLASTP 2.0MP-WashU [04-May-2006] [linux26-x86_64-I32LPP64 2006-05-10T17:22:28]

Copyright (C) 1996-2006 Washington University, Saint Louis, Missouri USA.
All Rights Reserved.

Reference: Gish, W. (1996-2006) <http://blast.wustl.edu>

Query= AMH human
(561 letters)

Database: uniprotkb/uniprotkb_swissprot; uniprotkb/uniprotkb_swissprotsv;
uniprotkb/uniprotkb_trembl
34,569,655 sequences; 11,137,001,357 total letters.
Searching....10....20....30....40....50....60....70....80....90....100% done

Sequences producing High-scoring Segment Pairs:	Smallest Sum		
	High Score	Probability P(N)	N
SP:MIS_HUMAN P03971 Muellerian-inhibiting factor OS=Homo ...	1466	3.1e-217	2
TR:G3SOU3_GORGO G3SOU3 Uncharacterized protein OS=Gorilla...	1443	2.2e-214	2
TR:G3S5V4_GORGO G3S5V4 Uncharacterized protein OS=Gorilla...	1419	7.6e-212	2
TR:H2NWW7_PONAB H2NWW7 Uncharacterized protein OS=Pongo a...	1309	6.9e-198	2
TR:F7FPV1_CALJA F7FPV1 Uncharacterized protein OS=Callith...	1261	9.4e-192	2
TR:H2QEW1_PANTR H2QEW1 Uncharacterized protein OS=Pan tro...	1445	5.4e-187	2
TR:F7FPV1_CALJA F7FPV1 Uncharacterized protein OS=Callith...	1247	2.2e-181	2
TR:F2YMM5_HORSE F2YMM5 Anti-Mullerian hormone OS=Equus ca...	1072	7.2e-170	2
TR:HOXPFL1_OTOGA HOXPFL1 Uncharacterized protein (Fragment)...	1017	1.2e-165	2
TR:G5BBC4_METGA G5BBC4 Muellerian-inhibiting factor OS=He...	1013	1.2e-158	2
TR:I3MTK1_SPTTR I3MTK1 Uncharacterized protein OS=Spermop...	966	1.2e-158	2
TR:HOW8G2_CAVPO HOW8G2 Uncharacterized protein OS=Cavia p...	1003	6.3e-158	2
TR:G1P2E6_MYOOLI G1P2E6 Uncharacterized protein OS=Myotis ...	958	5.7e-157	2
SP:MIS_PIG P79295 Muellerian-inhibiting factor OS=Sus scr...	949	1.3e-155	2
TR:Q5EC55_MOUSE Q5EC55 Anti-Mullerian hormone OS=Mus musc...	944	6.0e-151	2
SP:MIS_BOVIN P03972 Muellerian-inhibiting factor OS=Bos t...	897	6.0e-151	2
TR:F1M2L1_BOVIN F1M2L1 Muellerian-inhibiting factor OS=Bo...	896	7.6e-151	2
TR:A7E2R1_MOUSE A7E2R1 Amb protein (Fragment) OS=Mus musc...	944	2.0e-150	2
SP:MIS_RAT P49000 Muellerian-inhibiting factor OS=Rattus ...	934	2.6e-148	2
TR:IOBWL3_BUBBU IOBWL3 Anti-Mullerian hormone OS=Bubalus ...	897	2.6e-148	2
SP:MIS_MOUSE P27106 Muellerian-inhibiting factor OS=Mus m...	911	4.4e-146	2

Sequences producing High-scoring Segment Pairs:	High Score	Probability P(N)	N
SP:MIS_HUMAN P03971 Muellerian-inhibiting factor OS=Homo ...	1466	3.1e-217	2
TR:G3SOU3_GORGO G3SOU3 Uncharacterized protein OS=Gorilla...	1443	2.2e-214	2
TR:G3S5V4_GORGO G3S5V4 Uncharacterized protein OS=Gorilla...	1419	7.6e-212	2
TR:H2NWW7_PONAB H2NWW7 Uncharacterized protein OS=Pongo a...	1309	6.9e-198	2
TR:F7FPV1_CALJA F7FPV1 Uncharacterized protein OS=Callith...	1261	9.4e-192	2
TR:H2QEW1_PANTR H2QEW1 Uncharacterized protein OS=Pan tro...	1445	5.4e-187	2
TR:F7FPV1_CALJA F7FPV1 Uncharacterized protein OS=Callith...	1247	2.2e-181	2
TR:F2YMM5_HORSE F2YMM5 Anti-Mullerian hormone OS=Equus ca...	1072	7.2e-170	2
TR:HOXPFL1_OTOGA HOXPFL1 Uncharacterized protein (Fragment)...	1017	1.2e-165	2
TR:G5BBC4_METGA G5BBC4 Muellerian-inhibiting factor OS=He...	1013	1.2e-158	2
TR:I3MTK1_SPTTR I3MTK1 Uncharacterized protein OS=Spermop...	966	1.2e-158	2
TR:HOW8G2_CAVPO HOW8G2 Uncharacterized protein OS=Cavia p...	1003	6.3e-158	2
TR:G1P2E6_MYOOLI G1P2E6 Uncharacterized protein OS=Myotis ...	958	5.7e-157	2
SP:MIS_PIG P79295 Muellerian-inhibiting factor OS=Sus scr...	949	1.3e-155	2
TR:Q5EC55_MOUSE Q5EC55 Anti-Mullerian hormone OS=Mus musc...	944	6.0e-151	2
SP:MIS_BOVIN P03972 Muellerian-inhibiting factor OS=Bos t...	897	6.0e-151	2
TR:F1M2L1_BOVIN F1M2L1 Muellerian-inhibiting factor OS=Bo...	896	7.6e-151	2
TR:A7E2R1_MOUSE A7E2R1 Amb protein (Fragment) OS=Mus musc...	944	2.0e-150	2
SP:MIS_RAT P49000 Muellerian-inhibiting factor OS=Rattus ...	934	2.6e-148	2
TR:IOBWL3_BUBBU IOBWL3 Anti-Mullerian hormone OS=Bubalus ...	897	2.6e-148	2
SP:MIS_MOUSE P27106 Muellerian-inhibiting factor OS=Mus m...	911	4.4e-146	2

Results for job wublast-l20130516-120901-0250-57471242-pg

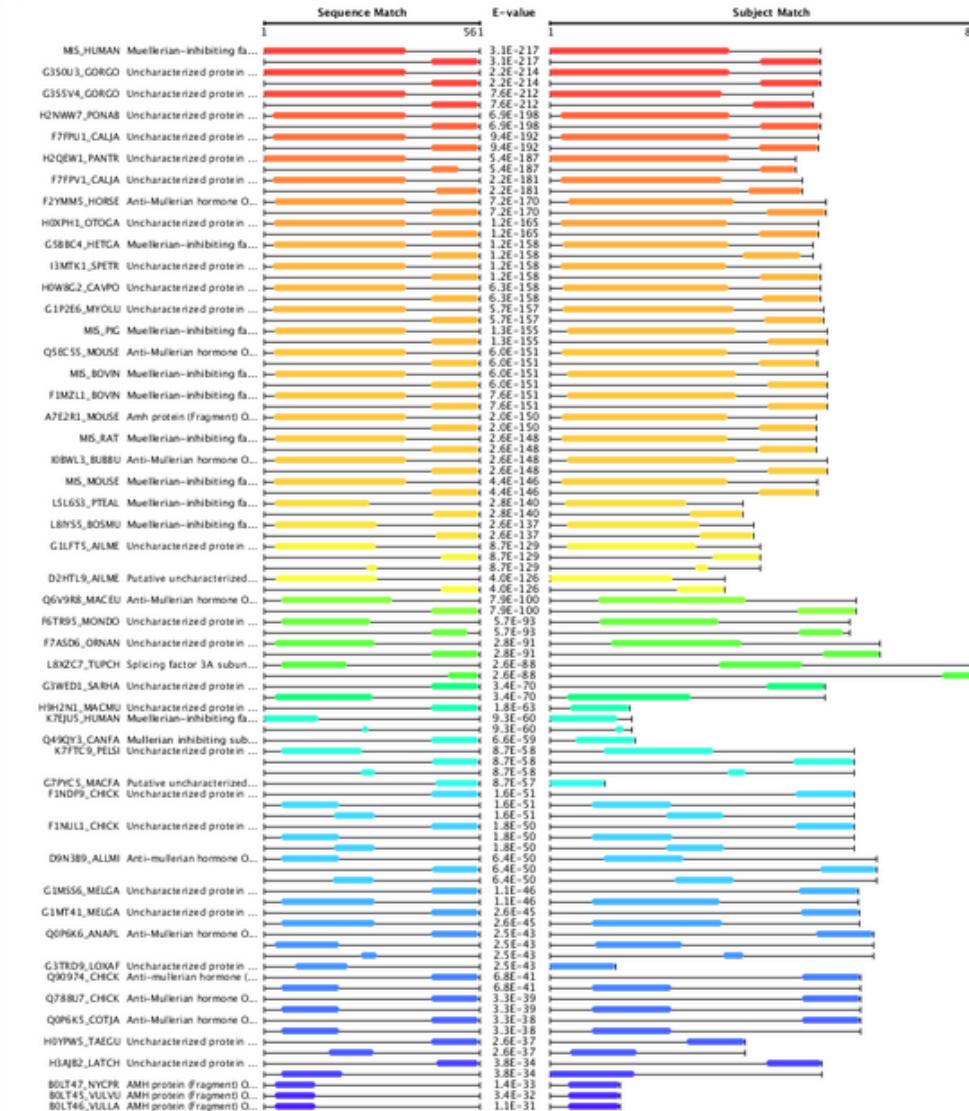
[Summary Table](#) [Tool Output](#) [Visual Output](#) [Functional Predictions](#) [Submission Details](#)

Color scale:

 fixed dynamic [Update](#)[Download in SVG format](#)

BLASTP (version: 2.0MP-WashU [04-May-2006] [linux26-x64-i32LPF64 2006-05-10T17:22:28])
 Database: uniprotkb
 Sequence: AMH human
 Length: 561

Launched Thu, May 16, 2013 at 12:09:01
 Finished Thu, May 16, 2013 at 12:11:25



E-value
 3.1E-217 1.847E-124 4.507E-78 7.041E-55 1.1E-31

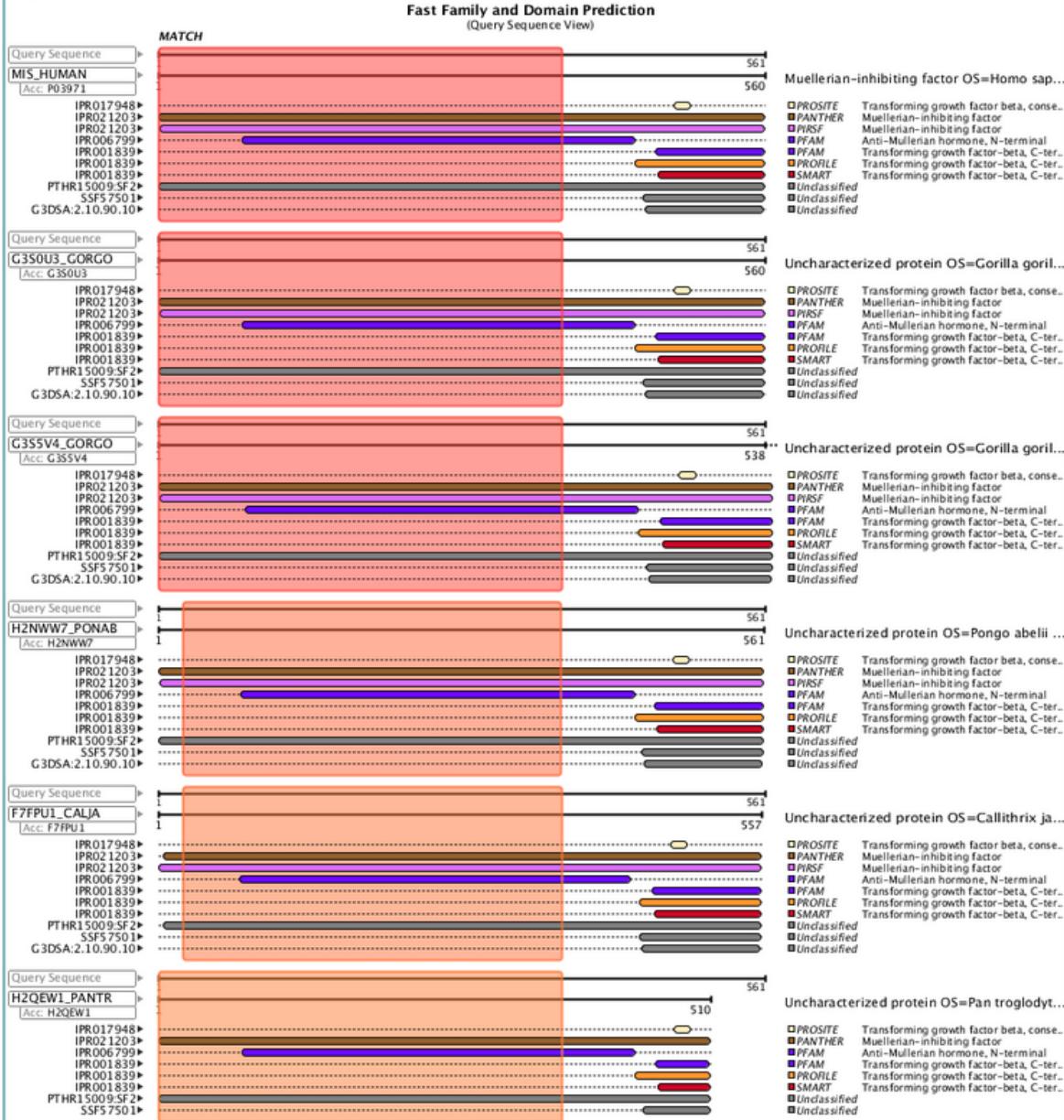
Color scale:

 fixed dynamic[Update](#)[Download in SVG format](#)[Switch to Subject Sequence View](#)

BLASTP (version: 2.0MP-WashU [04-May-2006] [linux26-x64-i32LPF64 2006-05-10T17:22:28])
 Database: uniprotkb
 Sequence: AMH human
 Length: 561

Launched Thu, May 16, 2013 at 12:09:01
 Finished Thu, May 16, 2013 at 12:11:25

Protein features:

 GENE3D PANTHER PFAM PIRSF PRINTS PRODOM PROFILE PROSITE SMART SSF TIGERFAMs Unclassified[Update](#)

CE QUE VOUS ALLEZ VOIR SUR UN SERVEUR BLAST OUTPUT EBI

Results for job wublast-I20130516-120901-0250-57471242-pg

[Summary Table](#) [Tool Output](#) [Visual Output](#) [Functional Predictions](#) [Submission Details](#)

Program

BLASTP

Version

2.0MP-WashU [04-May-2006] [linux26-x64-I32LPF64 2006-05-10T17:22:28]

Database

uniprotkb

Title

Launched Date

Thu, May 16, 2013 at 12:09:01

End Date

Thu, May 16, 2013 at 12:11:25

Input Sequence

wublast-I20130516-120901-0250-57471242-pg.input

Output Result

wublast-I20130516-120901-0250-57471242-pg.output

Command

```
/nfs/public/ro/es/appbin/linux-x86_64/wu-blast-wrapper/wu-blast_wrapper.pl blastp "uniprotkb" wublast-I20130516-120901-0250-57471242-pg.sequence E=10 B=50 V=50 -mformat=1 -matrix blosum62 -sump -topcomboN 1 -filter seg -sort_by_pvalue
```

Input Parameters

Program

blastp

Expectation value threshold

10

Alignments

50

Scores

50

Align views

1

D'AUTRES SERVEURS BLAST

The image displays three search interfaces from the Ensembl website:

- EnsemblPlants**: A search interface for plant genomes. It includes a search bar for "All species" and a list of popular genomes: *Arabidopsis thaliana* (TAIR10), *Oryza sativa* (MSU6), and *Zea mays* (AGPv2).
- EnsemblBacteria**: A search interface for bacterial genomes. It includes a search bar for "All collections" and a list of popular genomes: *Escherichia coli K12* (EB1_e_coli_k12), *Bacillus subtilis* (EB2_b_subtilis), and *Mycobacterium tuberculosis H37Rv* (EB1_m_tuberculosis_h37rv).
- BlastView**: An integrated platform for sequence similarity searches. It features a search interface with tabs for "new", "SETUP", "CONFIG", "RESULTS", and "DISPLAY". The "SETUP" tab is active, showing fields for "Important Notice" (mentioning Blat as default DNA search), "Enter the Query Sequence" (with options for pasting FASTA or plain text sequences, uploading files, entering sequence IDs, or retrieving existing ticket IDs), "Select the databases to search against" (with dropdown menus for species like *Gasterosteus aculeatus*, *Gorilla gorilla*, and *Homo sapiens*, and database types like "LATESTGP" and "PEP_ALL"), "Select the Search Tool" (with radio buttons for BLASTN, BLAT, and TBLASTX, currently set to BLAT), and "Search sensitivity" (set to "Near-exact matches"). A dropdown menu on the right lists search modes: "Near-exact matches" (selected), "Exact matches", "Allow some local mismatch", "Distant homologies", and "No optimisation".

BLAST EN LIGNE DE COMMANDE

BLAST travaille sur des bases de données **pré-indexées**, c'est-à-dire traitées de manière à respecter une structure de données précise. L'indexation d'une base de données se fait au moyen de la commande **formatdb** (remplacée par makeblastdb pour les dernières versions de Blast, fin 2012). Sur le site du NCBI, il est néanmoins possible de télécharger un certain nombre de bases de données pré-indexées.

Deux étapes:

1. Formater la banque de données
2. Exécuter le blast



BLAST EN LIGNE DE COMMANDE

Construction d'une banque BLAST

La première étape lors de l'utilisation de BLAST consiste à construire une base de séquences utilisable par BLAST. Le programme **formatdb**, fourni dans le package **blast2** (ou makeblastdb en fonction de la version du package blast utilisé), permet de créer une telle base de données, à partir de séquences stockées au format FASTA.

Quelques options de formatdb:

- i nom du fichier en entrée (au format FASTA)
- p type des séquence (par défaut T pour protéique, sinon F pour nucléique)
- n nom de la base de données résultat (par défaut nom du fichier en entrée)
- o option pour parser les SeqID et créer des indexes (par défaut F, T si les séquences sont au format NCBI).

Pour avoir plus de précisions sur les options, vous pouvez taper formatdb -help ou man formatdb dans un terminal ou regarder le site web suivant :

http://www.ncbi.nlm.nih.gov/IEB/ToolBox/C_DOC/lxr/source/doc/blast/formatdb.html

Par exemple, supposons que l'on souhaite formater le fichier ecoli.nt contenant un grand ensemble de séquences d'E. coli. La commande nécessaire est :

formatdb -i ecoli.nt -p F -n maBdBlastEcoli



BLAST EN LIGNE DE COMMANDE

Exécution d'un BLAST

Il est ensuite possible de faire des requêtes sur la base ainsi créée. Pour cela, il est possible d'utiliser le programme blastall fourni dans le package *blast2*.

Quelques options :

- p type de BLAST (blastn, blastp, blastx, tblastn, tblastx)
- d noms des banques BLAST à utiliser (sans les extensions de fichier)
- i fichier en entrée au format fasta
- o fichier de sortie (par défaut stdout)
- e seuil de la E-value (par défaut=10.0).



BLAST EN LIGNE DE COMMANDE

Pour avoir plus de précisions sur les options, vous pouvez taper blastall ou man blastall dans le terminal ou regarder les sites web suivants : <http://www.ncbi.nlm.nih.gov/staff/tao/URLAPI/blastall/> ou <http://www.ncbi.nlm.nih.gov/books/NBK1763/>.

Par exemple, supposons que notre séquence inconnue soit contenue dans le fichier test.txt, la commande permettant de la comparer aux séquences de la banque maBdBlastEcoli est :

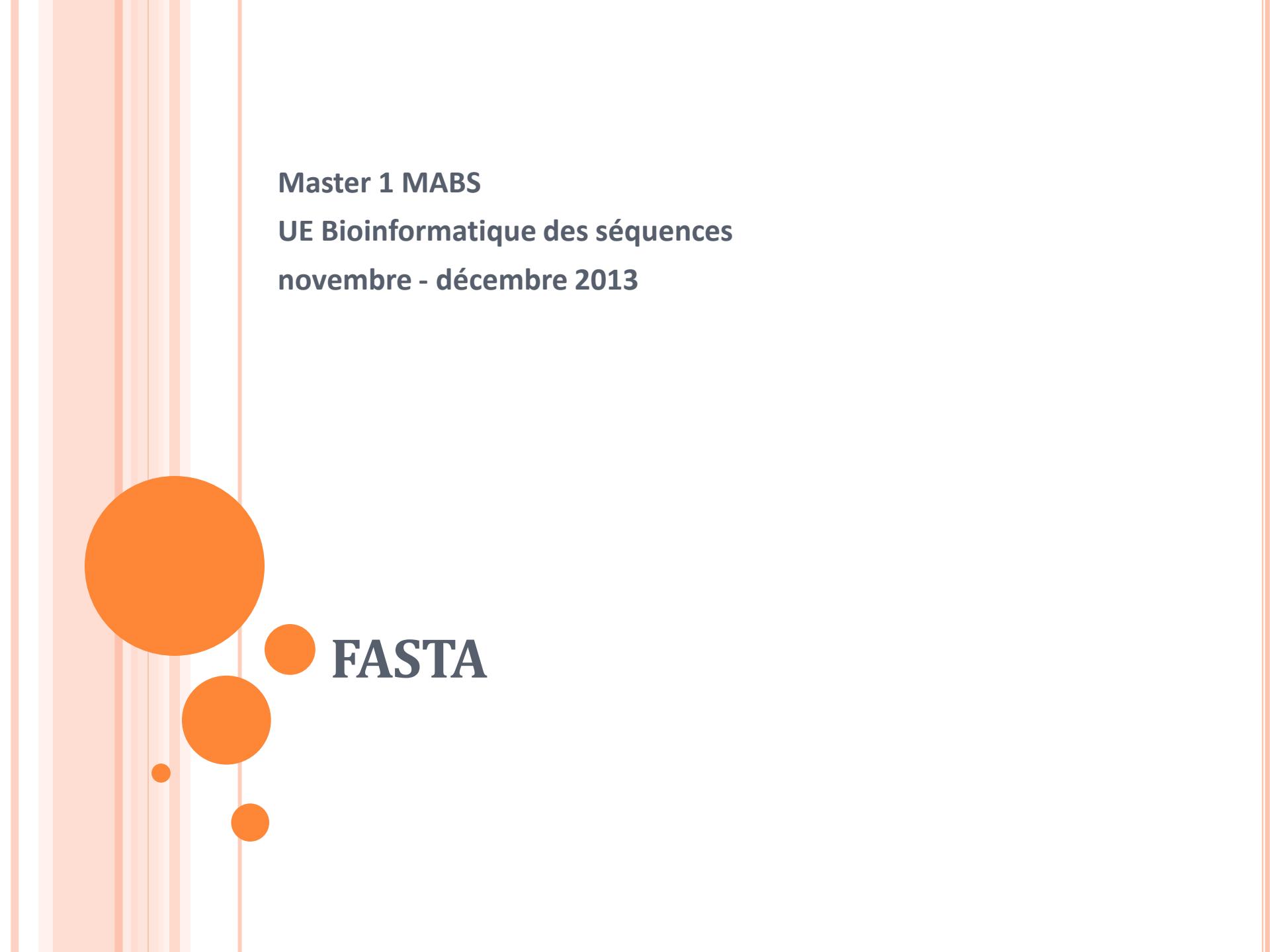
blastall -p blastn -d maBdBlastEcoli -i test.txt -o test.out

Supposons maintenant que nous recherchons les homologues d'une séquence nucléotidique contenue dans le fichier K03455seqref.fas (format FASTA), avec un seuil égal à 0.15 (E-value < 0.15), dans la base GenBank (supposons celle-ci installée en local et nommée nr). Pour cela, on souhaite afficher aux environs de 5000 résultats (-v), sans aucun alignements (-b 0). La commande nécessaire est :

blastall -p blastn -d nr -i K03455seqref.fas -e 0.15 -v 5000 -b 0

Notons que blastall ne permet pas de choisir l'espèce (l'interface Web du NCBI le permet, au travers d'une requête Entrez).





Master 1 MABS
UE Bioinformatique des séquences
novembre - décembre 2013

FASTA

FASTA

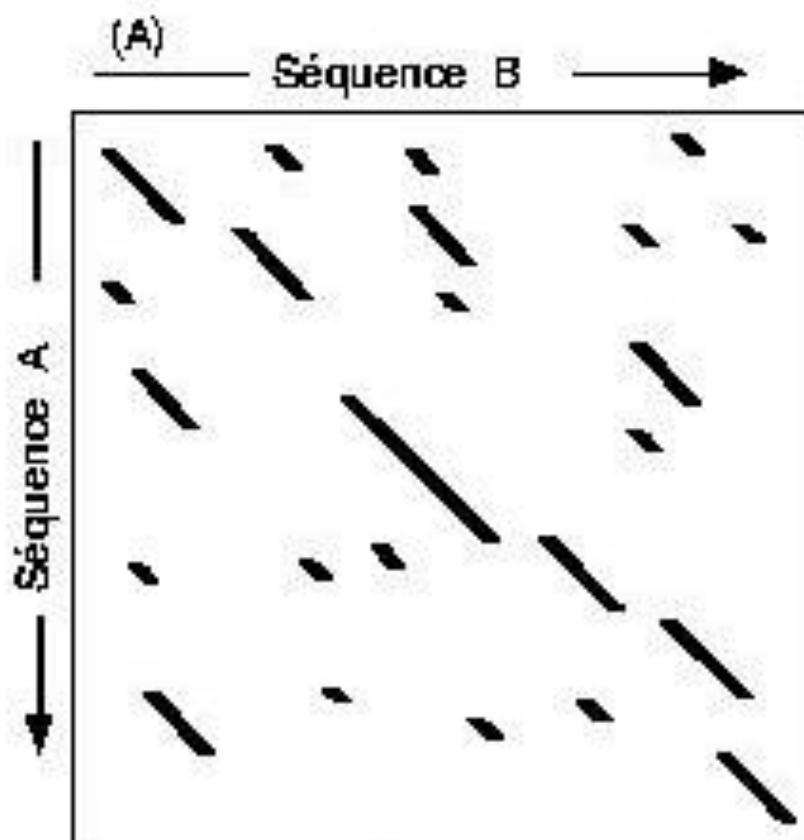
- Fast Accurate Search Tool Alignment
- Un autre algorithme pour rechercher des alignements locaux
- S'utilise de manière similaire à BLAST
- Plus ancien
- Le format du logiciel FASTA est très utilisé pour représenter les séquences
- Ex : BLAST utilise ce format
- http://fasta.bioch.virginia.edu/fasta_www2/fasta_list2.shtml
- <http://www.ebi.ac.uk/Tools/ssss/fasta/nucleotide.html>



Algorithme de FASTA (1)

(Pearson et Lipman, 1988)

Dérivé de la logique du DotPlot



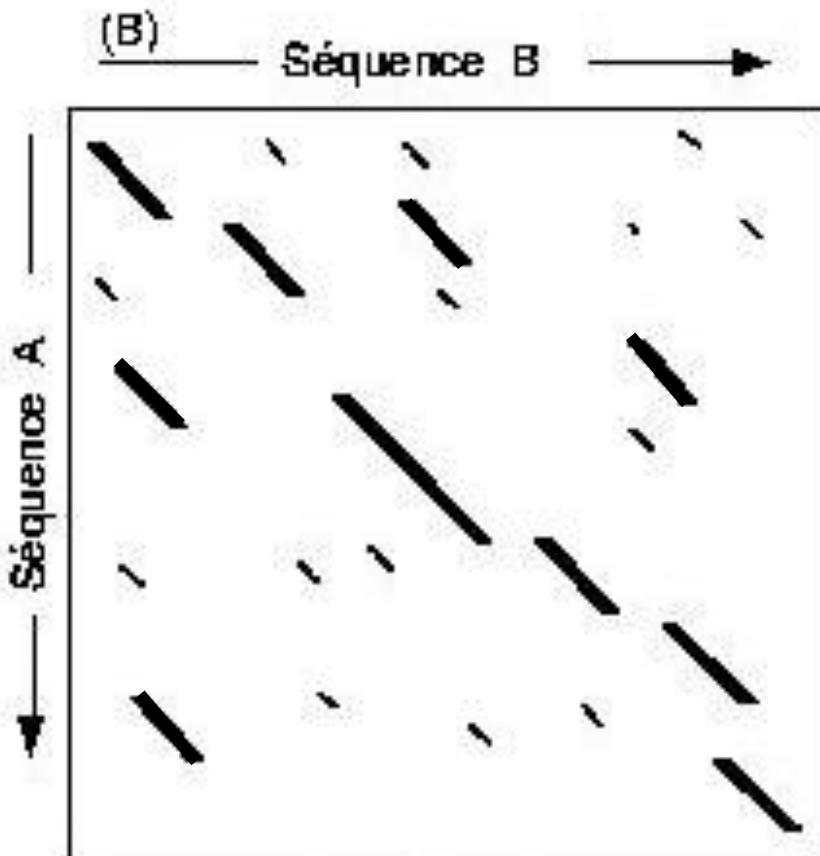
Recouvrement de mots de longueur
minimum k

Etape 1 : repérage des
régions les plus denses en
identités partagées

$k=4$ ou 6 pour ADN
 $k=1$ ou 2 pour a.a.



Algorithme de FASTA (2)

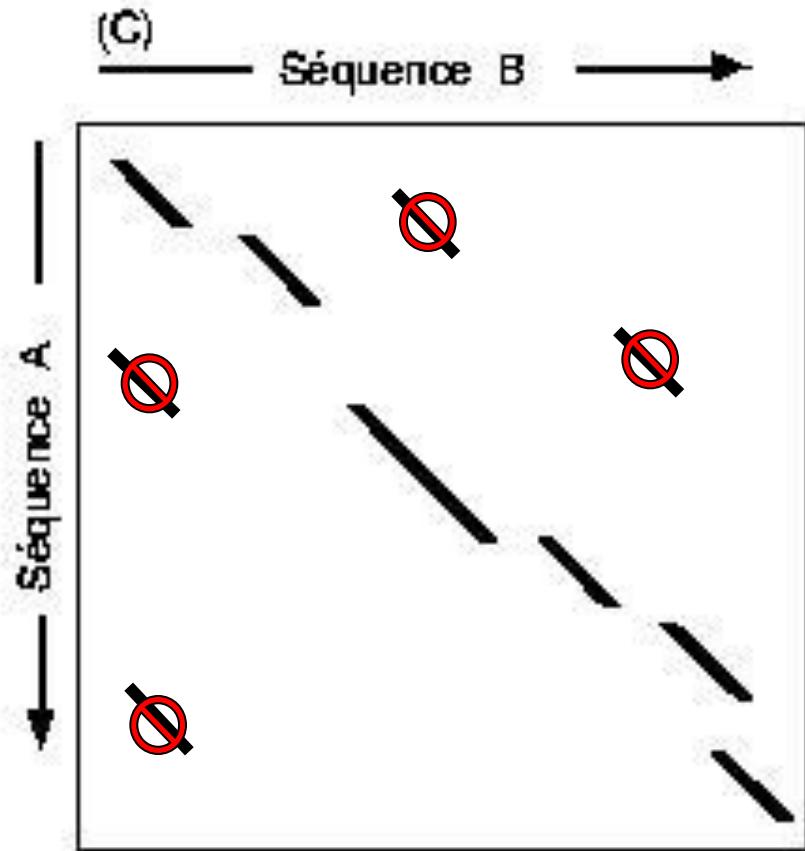


Score avec la matrice PAM
conserve les scores élevés. (init 1)

Etape 2 : calcul à l'aide d'une matrice de scores d'un score pour les 10 meilleures régions d'identité
= recherche de similitudes sans insertion-deletion

init1 est attribué à la région ayant le plus fort score parmi les 10 analysées

Algorithme de FASTA (3)



Utilisation de la méthode de "joining threshold" pour éliminer les segments peu probables de l'alignement parmis les segments de scores élevés.

Etape 3 : jonction de 2 régions si elles possèdent chacune un score supérieur à un score seuil.

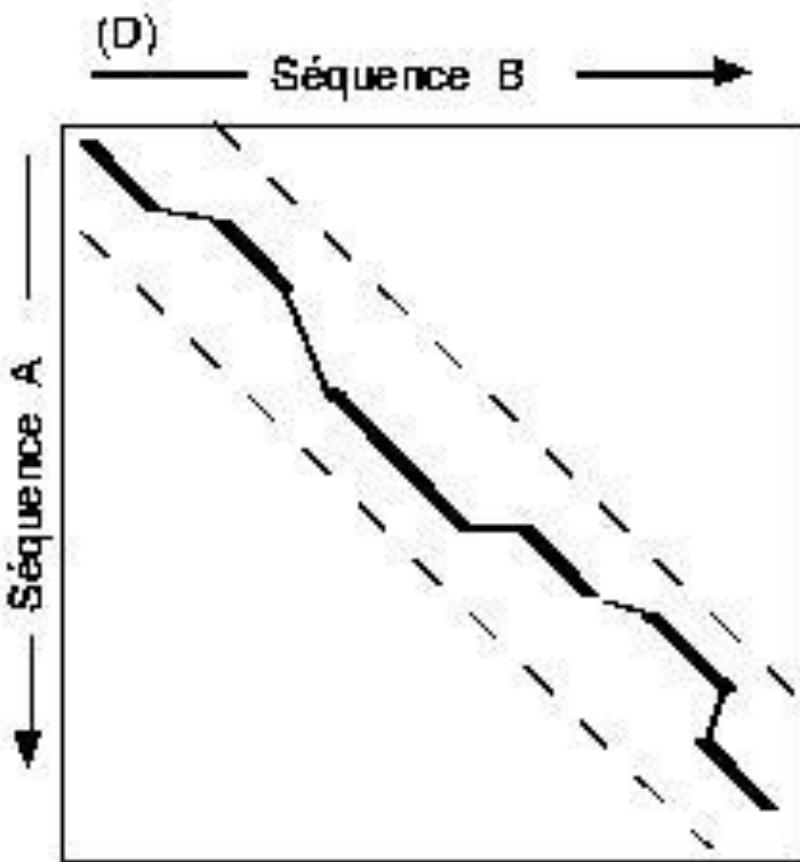
Seuil = score moyen attendu pour des régions non apparentées.

La somme de leur score moins la pénalité de jonction d. è. \geq score init1

Initn= nouveau meilleur score



Algorithme de FASTA (4)



Étape 4 :
alignement optimal
(programmation dynamique)
score **opt**

Utilisation d'une programmation
dynamique pour optimiser
l'alignement dans une étroite
bande entourant les scores
élevés.(opt)



En résultat de FASTA :

- les 3 scores observés : init1, initn, opt
- le Z-score : score maximal attendu normalisé entre les deux séquences (opt attendu)

$$\text{Z-score} = (\text{score opt}-m) / e$$

m : moyenne des scores aléatoires

e : écart type des scores aléatoires

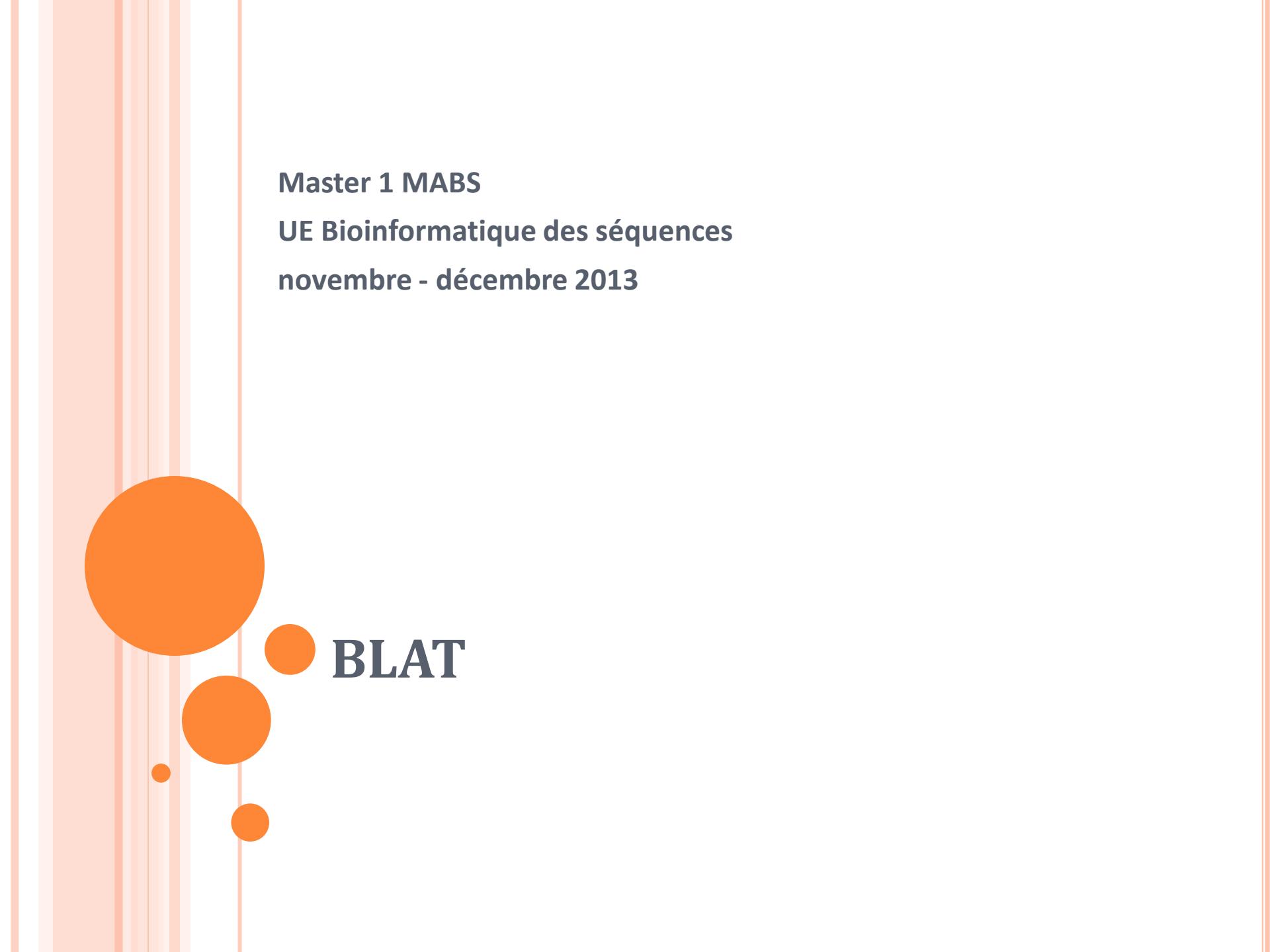
- la E-value : nombre de comparaisons donnant un score opt \geq Z-score

Un alignement est d'autant meilleur (significatif) que son score opt est supérieur au Z-score et que sa E-value est faible



<u>Programme</u>	<u>Séquence Query</u>	<u>Banque de séquences</u>
fasta	nucléique protéique	nucléique protéique
fastx	nucléique traduite	protéique
tfastx	protéique	nucléique traduite
tfasta	protéique	nucléique traduite





Master 1 MABS

UE Bioinformatique des séquences

novembre - décembre 2013

BLAT

BLAT

- The BLAST-Like Alignment Tool: similarity search in databanks.

But:

- sur l'ADN, BLAT est conçu pour trouver des séquences **rapidement** de 95% et plus de similarité, sur une longueur au moins égale à 25 bases.
- Sur les protéines, il trouve des séquence d'au moins 80% de similarité sur une longueur au moins égale à 20 acides aminés.
- C'est un blast très efficace, très rapide pour la recherche d'EST sur les génomes eucaryotes
- <http://genome.ucsc.edu>



BLAT: LES ÉTAPES

- **Etape 1 : *Search Stage***
 - Identification de régions homologues
 - Utilisation d'un Index sur la séquence Query et sur la base de donnée, puis filtre
- **Etape 2 : *Alignment Stage***
 - Alignement des régions identifiées
- **Etape 3 : *Stitching and Filling In***
 - rassemblement et remplissage entre régions
 - Programmation dynamique pour rassembler les alignements de régions en alignement global



BLAT: ZOOM SUR L'ÉTAPE 1

Etape 1 : *Search Stage*

1.a. Identification des régions homologues => index

Database : non-overlapping



Query : overlapping



BLAT: ZOOM SUR L'ÉTAPE 1

Etape 1 : *Search Stage*

1.a. Identification des régions homologues => index

Database : non-overlapping

cacaatttatcacgaccgc

3-mers: cac aat tat cac gac cgc

Index: aat 3 gac 12

 cac 0,9 tat 6

 cgc 15

Query : overlapping

aattctcac

3-mers: aat att ttc tct ctc tca cac

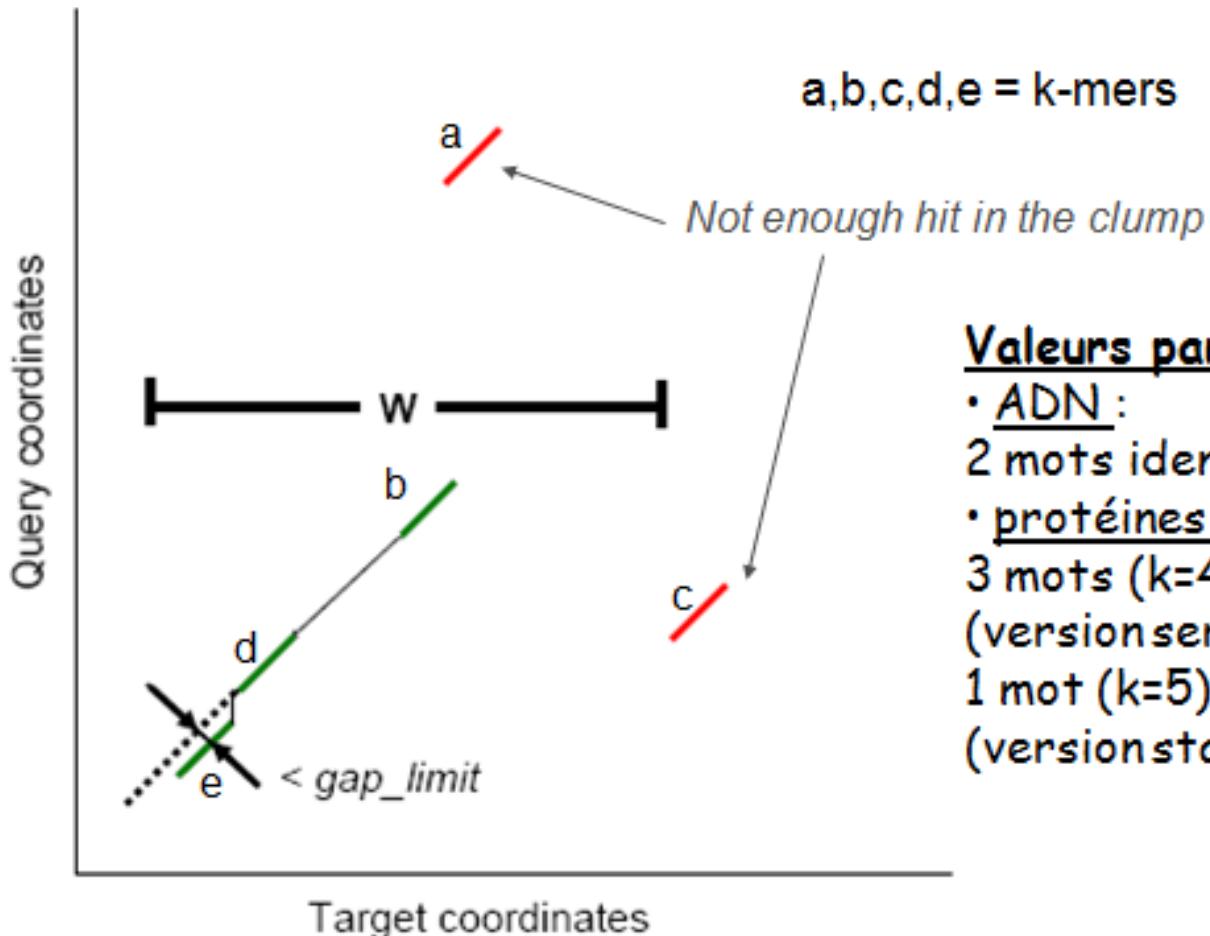
0 1 2 3 4 5 6

⇒ Liste de positions de hit des k-mers

⇒ Cluster de hits pour chercher régions homologues

BLAT: ZOOM SUR L'ÉTAPE 1

○ 1.b. Identification des régions homologues => filtre



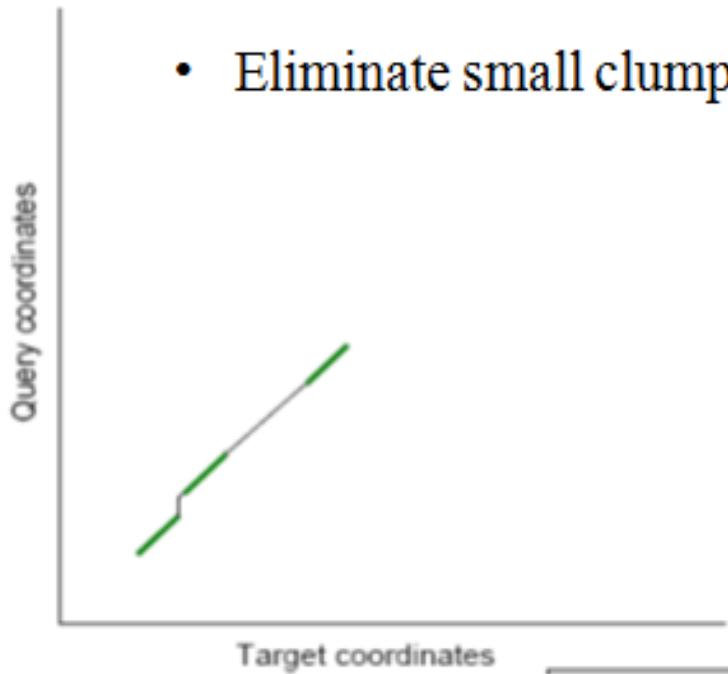
Valeurs par défaut :

- ADN :
2 mots identiques ($k=11$)
- protéines :
3 mots ($k=4$) identiques
(version serveur)
1 mot ($k=5$) identique
(version stand-alone)

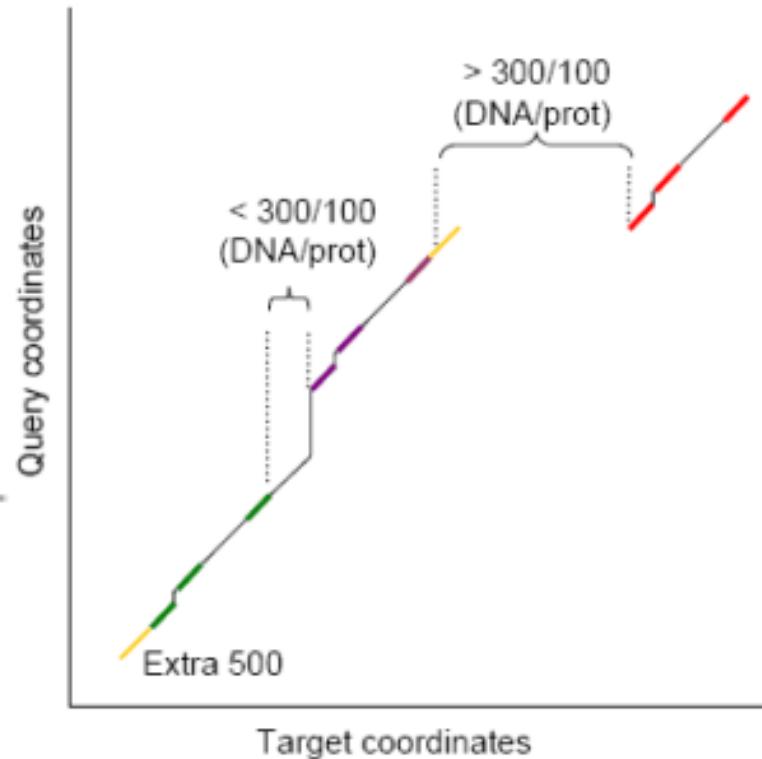
BLAT: ZOOM SUR L'ÉTAPE 1

○ 1.b. Identification des régions homologues => filtre

- Eliminate small clumps



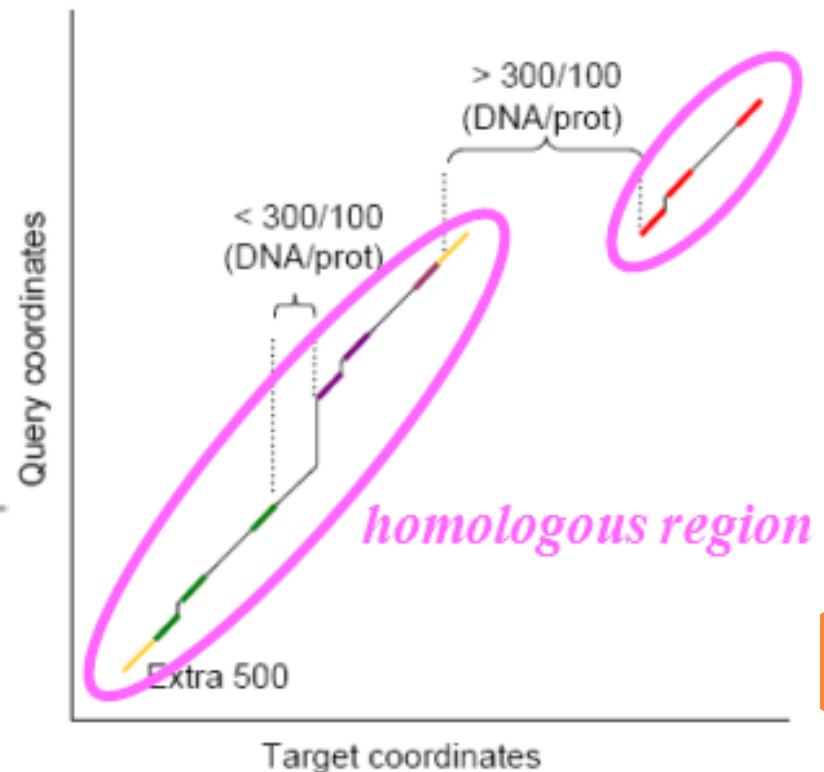
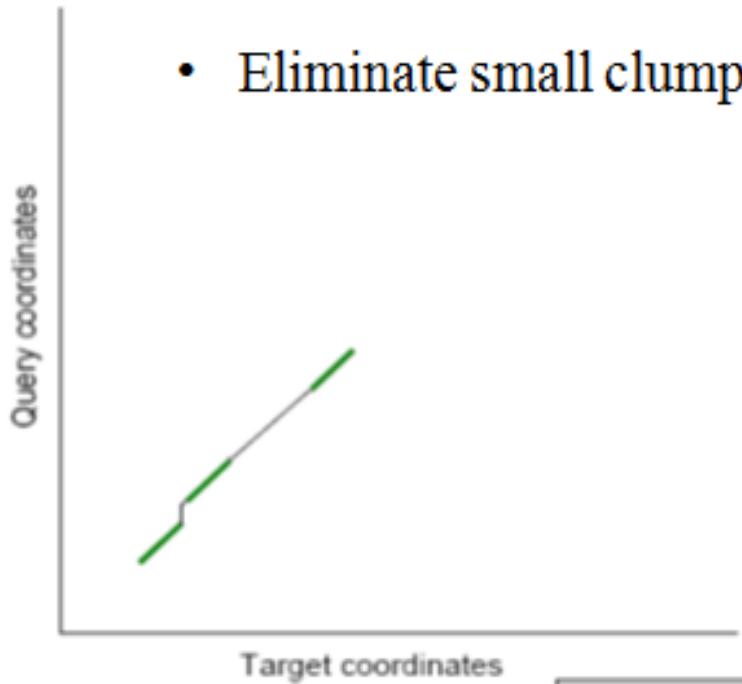
- Clump 'clumps'



BLAT: ZOOM SUR L'ÉTAPE 1

○ 1.b. Identification des régions homologues => filtre

- Eliminate small clumps



- Clump 'clumps'

FASTA/ BLAST/ BLAT

- **FASTA** + sensible pour les séquences nucléiques
mais + long !

1 seul alignement par paire de séquences

=> peut rater des zones de similarités

Problème dans le cas de réarrangements de domaines

- **BLAST** + rapide

retourne des alignements locaux

basé sur un modèle statistique

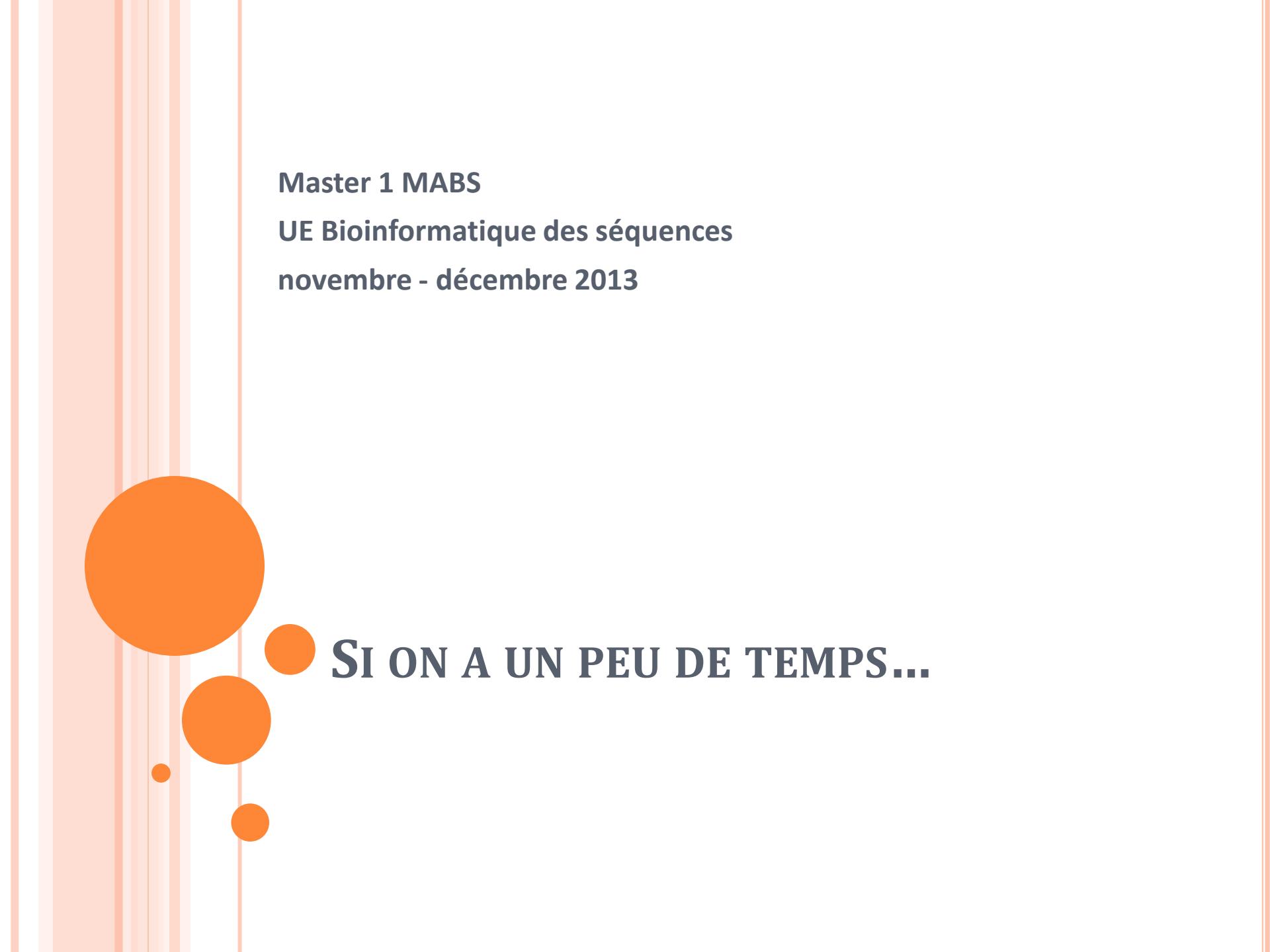
disponible sur la majorité des sites de génomes

- **BLAT** ++ rapide, mais – sensible

Alignement ARNm/génomique

disponible sur ENSEMBL





Master 1 MABS

UE Bioinformatique des séquences

novembre - décembre 2013

SI ON A UN PEU DE TEMPS...

EXERCICE

- **Quel(s) BLAST utiliser dans les situations suivantes :**
- Une molécule thérapeutique marquée a permis d'extraire la protéine cible de ce médicament, qui a ensuite été purifiée puis séquencée. On souhaite à présent savoir de quelle protéine il s'agit ?
- Un gène a été isolé chez E. Coli puis séquencé. On souhaite déterminer quelle(s) est(sont) la(les) protéine(s) codée(s) par ce gène ?
- L'enzyme de conversion de l'angiotensine est une protéase bien conservée au cours de l'évolution. Le gène de cette protéine chez l'homme a été récupéré dans la base EMBL, et on souhaite rechercher des gènes homologues chez d'autres espèces éloignées ?

EXERCICE

- **Quel(s) BLAST utiliser dans les situations suivantes :**
- Une molécule thérapeutique marquée a permis d'extraire la protéine cible de ce médicament, qui a ensuite été purifiée puis séquencée. On souhaite à présent savoir de quelle protéine il s'agit ? => **Blast p**
- Un gène a été isolé chez E. Coli puis séquencé. On souhaite déterminer quelle(s) est(sont) la(les) protéine(s) codée(s) par ce gène ? => **Blast x**
- L'enzyme de conversion de l'angiotensine est une protéase bien conservée au cours de l'évolution. Le gène de cette protéine chez l'homme a été récupéré dans la base EMBL, et on souhaite rechercher des gènes homologues chez d'autres espèces éloignées ? => **t Blast x**