Université de Lille

Blast

Basic Local Alignment Search Tool

Contenus basés sur les cours des enseignants en bioinformatique de l'université de Lille

Sylvain.legrand@univ-lille.fr

Introduction

Problématique

- On suppose que vous avez une **séquence nucléotidique ou protéique** obtenue à partir d'un échantillon biologique ou provenant d'une base de donnée

NCBI

Séquençage

>Votre séquence
ACGTCGTAGCATGCA
GCGATTAGCTACGAT
CAGCTAGCATCGATC
GATGGGCATTACGTA
CGTAGCTATGTCAGC
GATCGA...

Problématique

Vous voulez savoir si la séquence que vous avez obtenue est **déjà connue** ou est **similaire** à d'autres séquences dans les bases de données

Une recherche de **similarité de séquences** fournit souvent les premières informations sur une nouvelle séquence nucléotidique ou protéique

→ inférer **la fonction de la séquence** à partir de séquences similaires

Problématique

- On se donne :

  une séquence **requête** (query) *q*

  une **banque** de séquences *T* = {*t1*, ..,*tn*}

- Ce que l'on souhaite : trouver des **alignements significatifs** entre *q* et les *ti*

- Les algorithmes classiques (exemple : alignement local de Smith et Waterman) ne fonctionnent pas : prennent trop de temps, il faut trouver des parades

Blast, généralités

- Blast (définition NCBI) : The Basic Local Alignment Search Tool (BLAST) trouve des **régions de similarités locales** entre des séquences. Le programme compare des séquences nucléotidiques ou protéiques à des bases de données de séquences et calcule la **significativité statistique** des alignements

- Blast peut être utilisé pour inférer des **relations fonctionnelles et évolutives** entre les séquences et peut aussi aider à **identifier des membres d'une famille de gènes**

- Blast utilise des **heuristiques*** pour donner des résultats rapidement

*Une heuristique est une méthode de calcul qui fournit rapidement une solution réalisable, pas nécessairement optimale, elle peut manquer des résutats

Blast, généralités

- La plupart des protéines étant **modulaires** (composée de domaine(s) fonctionnel(s)), Blast est fait pour retrouver ces domaines entre des séquences différentes.

- L'algorithme permet aussi d'aligner un **mRNA sur de l'ADN génomique**.

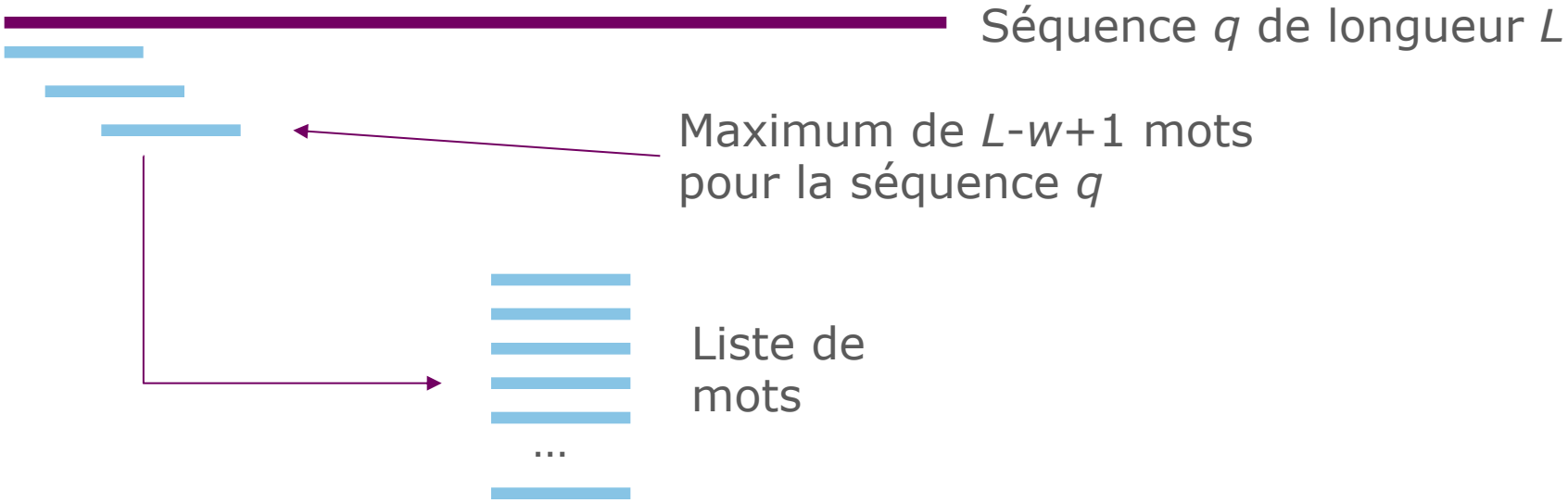
- En revanche s'il est attendu d'aligner 2 séquences sur leur pleine longueur (alignement global), il est possible que Blast ne retourne que les parties les plus conservées de cet alignement

Historique

- Première version produite par le NCBI en **1990** (Altschul et al. 1990)
- Cette version ne réalise que des **alignements sans gaps**, mais fournit une p-value qui permet à l'utilisateur de juger de la significativité des résultats
- Une version autorisant les **gap** (Blast2) apparaît en 1997 (Altschul et al. 1997) et inclue le **PSI Blast** (voir plus loin)
- En 2009, le NCBI sort une nouvelle version de Blast (**BLAST+**) (Camacho et al. 2009)
- **Actuellement BLAST+2.4.0** released (02/06/16)

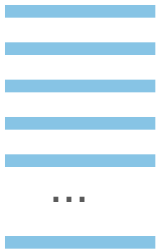
Algorithme

- **1ère étape** : définir à partir de la séquence requête *q* **une liste de mots** (graines) de taille définie *w* (taille par défaut de 11 pour l'ADN et de 3 pour les protéines)



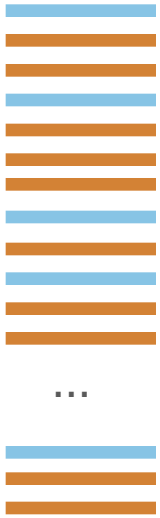
Algorithme

- Particularité pour les **protéines**



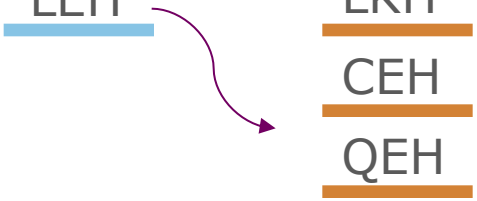
Liste de mots

Recherche de mots similaires selon une matrice de substitution (PAM ou BLOSUM) (seuil >= *T*)



Liste étendue aux mots similaires

Exemple :



Mots similaires

- Précision sur les **mots similaires**

- Pour chaque mot de taille w=3, Blast génère les mots voisins à l'aide d'une matrice BLOSUM62 avec un seuil de score T=11

- Mot de 3 acides aminés → 20^3 alignements possibles

LEH→ score= 17
LKH→ score= 13
CEH→ score= 12
seuil QEH→ score= 11
LMP→ score= 10
LFH→ score= 9
LER→ score= 9
SEH→ score=9
…

- Les mots voisins sont alignés avec LEH et le score d'alignement est calculé à partir de la matrice BLOSUM62.

- Ne sont gardés que les mots présentant un score supérieur ou égal au seuil T

Matrice de substitution

- Une matrice de substitution permet d'associer un score à chaque paire de résidus que l'on trouve dans un alignement

- Pour les séquences nucléotidiques, on utilise généralement des pénalités identiques pour toutes les substitutions

	A	C	G	T
A	1			
C	-3	1		
G	-3	-3	1	
T	-3	-3	-3	1

- Pour un alignement donné, le score est la somme des scores de chaque paire de résidus

```
          A  C  G  C  A  T  G  C  A  T  C
          A  G  G  C  A  T  C  G  A  T  T
Score:    1 -3  1  1  1  1 -3 -3  1  1  1 = -1
```

Matrice de substitution

- Pour les séquences protéiques, on utilise des matrices (BLOSUM et PAM) qui donnent des scores différents suivant les substitutions

- Les scores positifs indiquent des substitutions fréquentes (« acceptées »), c'est-à-dire des substitutions observées plus fréquemment que ce à quoi on s'attendrait au hasard

- Les valeurs négatives indiquent des mutations rares, que l'on observe moins fréquemment qu'au hasard. C'est un indice de contre sélection, suggérant que ces mutations sont défavorables à la fonction de la protéine

Matrice BLOSUM62

		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
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				
	B	-2	-1	3	4	-3	0	1	-1	0	-3	-4	0	-3	-3	-2	0	-1	-4	-3	-3	4			
	Z	-1	0	0	1	-3	3	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-3	-2	-2	1	4		
	X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1	-1	-1	
	*	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	1

	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	B	Z	X	*
Hydrophobic	A				C			G		I	L		M		P					V				
Aromatic									H					F				W	Y					
Polar			N			Q										S	T		Y					
Basic		R							H			K												
Acidic				D			E																	

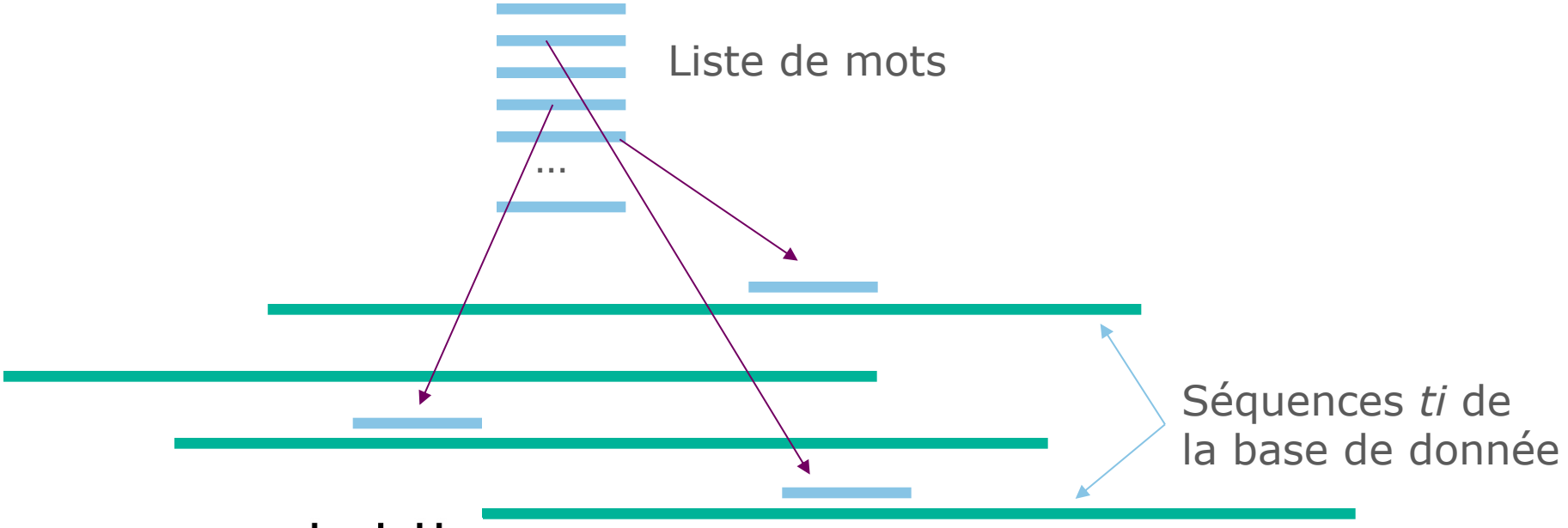
22

From J. van Helden, Université d'Aix-Marseille

- **2ère étape** : rechercher des alignements exacts entre les mots de la liste (ADN) ou de la liste étendue (protéines) et les séquences *ti* de la base de donnée

- Ces alignements forment des *hits*

- Un *hit* est donc un mot « commun » de taille *w* (et de score supérieur à *T* dans le cas des protéines) entre les séquences *q* et *ti*



- **3ème étape** : chaque *hit* est étendu à gauche et à droite : L'extension est stoppée lorsque de le score du *hit* décroit de plus de *X* (*X-drop*)

- Schématiquement

Séquence *q*

Séquence *de la base de donnée*

- Chaque hit étendu forme un **LMSP**: Localy Maximal scoring Segment Pair

- Ne sont conservées que les LMPSP de score supérieur à un score seuil donné, les **HSP** : High scoring Segment Pair

- La HSP la plus significative est nommée **MSP** : Maximum scoring Segment Pair

Algorithme

- Précision sur le X-drop

  Query *q* : Y A N C Q E H K M G S

  Subject ti : D A P C Q E H K R G W P N D C

Xdrop=2
Score calculé selon BLOSUM62

Hit de départ

Y A N C Q E H K M G S
D A P C Q E H K R G W P N D C

5 10 18

Score cumulé

Extension à droite

Y A N C Q E H K M G S
D A P C Q E H K R G W P N D C

5 10 18 23 22 28 25

Score cumulé

Le score décroit de 3 >Xdrop → l'alignement est arrêté

Extension à gauche

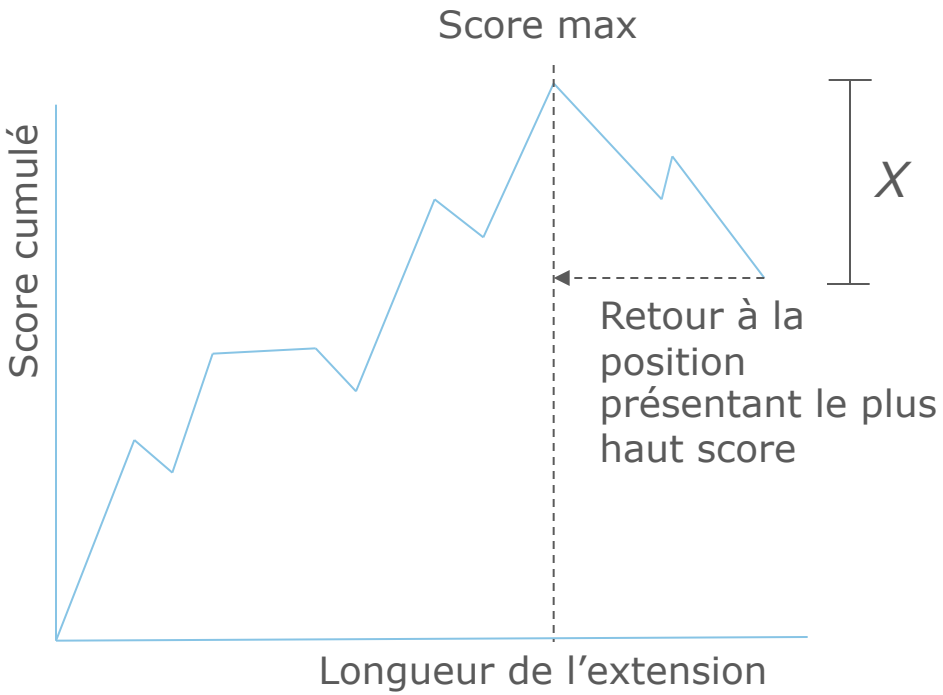
Y A N C Q E H K M G S
D A P C Q E H K R G W P N D C

26 29 25 27 18 13 8

Score cumulé

Algorithme

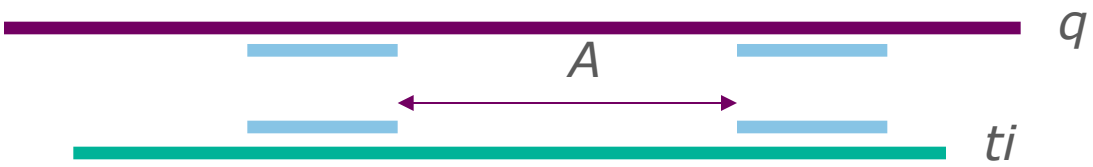
- Précision sur le X-drop



Gapped-Blast (BLAST2)

- Se baser sur 2 hits distants au maximum de A (BLASTP)

En contre partie pour garder une bonne sensibilité, T est abaissé de 13 à 11



- Etendre les hits en autorisant les gaps comme précédemment

- Cette méthode est plus rapide que la précédente

Significativité des alignements

Significativité des alignements

- Deux séquences peuvent toujours être alignées

- Il existe toujours un (au moins) alignement de meilleur score S entre deux séquences (un MSP)

- Problèmes :

- Ce score est-il suffisamment élevé pour prouver une homologie ?

- Peut-on trouver un MSP de meilleur score dans deux séquences aléatoires ?

Significativité des alignements

- Soit *S* le score obtenue par l'alignement de 2 séquences

- La **p-valeur** (p-value) mesure la ***Probabilité*** que 2 séquences aléatoires de même longueur et de même composition possèdent un MSP de score ≥ *S*

- La **E-valeur** (E-value) mesure l'***Esperance*** *E* du nombre *n* de MSPs de score ≥ *S* dans 2 séquences aléatoires de même longueur et de même composition

- Par exemple, si la E-valeur est égale à 10 pour une HSP de score S, cela veut dire que 10 HSPs avec un score ≥ S peuvent être trouvées par chance ! → Donc probablement votre alignement n'est pas significatif !

Calcul de la E-value

- Selon Karlin et Altschul, 1991

$$E = Kmne^{-\lambda s} \qquad p = 1 - e^{-E}$$

Avec *m* la taille de la séquence *q, n la taille de la banque de données, S le score de la HSP, K et l dépendent de la matrice de score, K peut être ajusté en fonction du coût des Gaps*

- Si S est le score d'un hit
- Le bit score (score normalisé) est : $S' = \frac{\lambda s - lnK}{ln2}$
- La E-value est alors : $E = mn2^{-S'}$

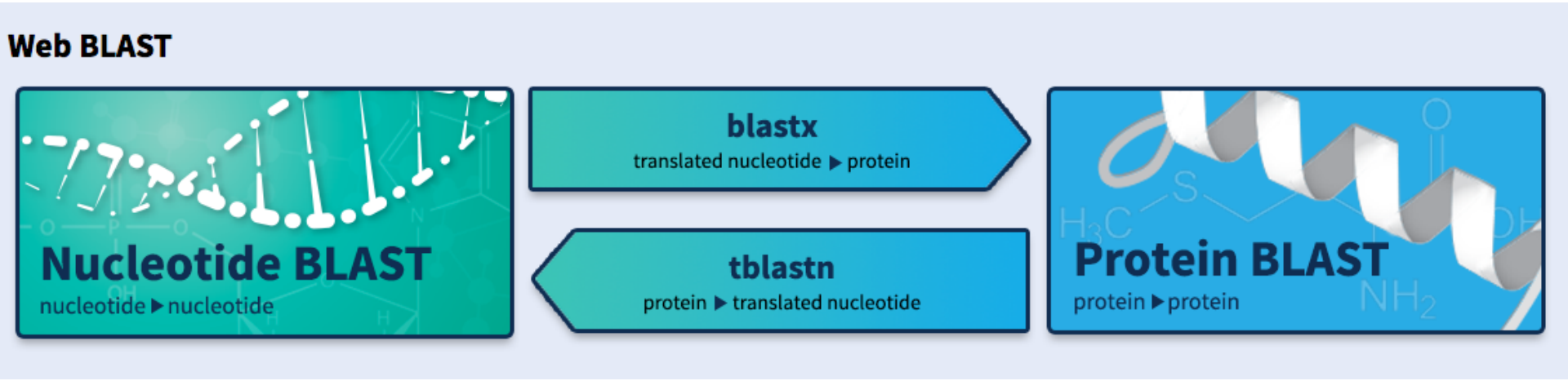
Variation de la E-value

- si la taille de la séquence query augmente : la E-value …
- Si la taille de la banque est divisée par deux : la E-value …
- Si le score augmente : la E-value …
- Quel bit-score pour obtenir une E-value de 0.05 pour une séquence de longueur 250 et une bd de longueur 50000000 ?
- Si on passe la E-value à 0.01, quel sera le bit-score ?

Variation de la E-value

- si la taille de la séquence query augmente : la E-value augmente
- Si la taille de la banque est divisée par deux : la E-value diminue
- Si le score augmente : la E-value diminue
- Quel bit-score pour obtenir une E-value de 0.05 pour une séquence de longueur 250 et une bd de longueur 50000000 ? 38 bits
- Si on passe la E-value à 0.01, quel sera le bit-score ? 40 bits

Lancer des Blast !



Query \ Database	nucléique	protéique	nucléique traduit
nucléique	blastn	x	x
protéique	x	blastp	tblastn
nucléique traduit	x	blastx	tblastx

ftp://ftp.ncbi.nlm.nih.gov/pub/factsheets/HowTo_BLASTGuide.pdf

Home page

Specialized searches

SmartBLAST

Find proteins highly similar to your query

Primer-BLAST

Design primers specific to your PCR template

Global Align

Compare two sequences across their entire span (Needleman-Wunsch)

CD-search

Find conserved domains in your sequence

GEO

Find matches to gene expression profiles

IgBLAST

Search immunoglobulins and T cell receptor sequences

VecScreen

Search sequences for vector contamination

CDART

Find sequences with similar conserved domain architecture

Targeted Loci

Search markers for phylogenetic analysis

Multiple Alignment

Align sequences using domain and protein constraints

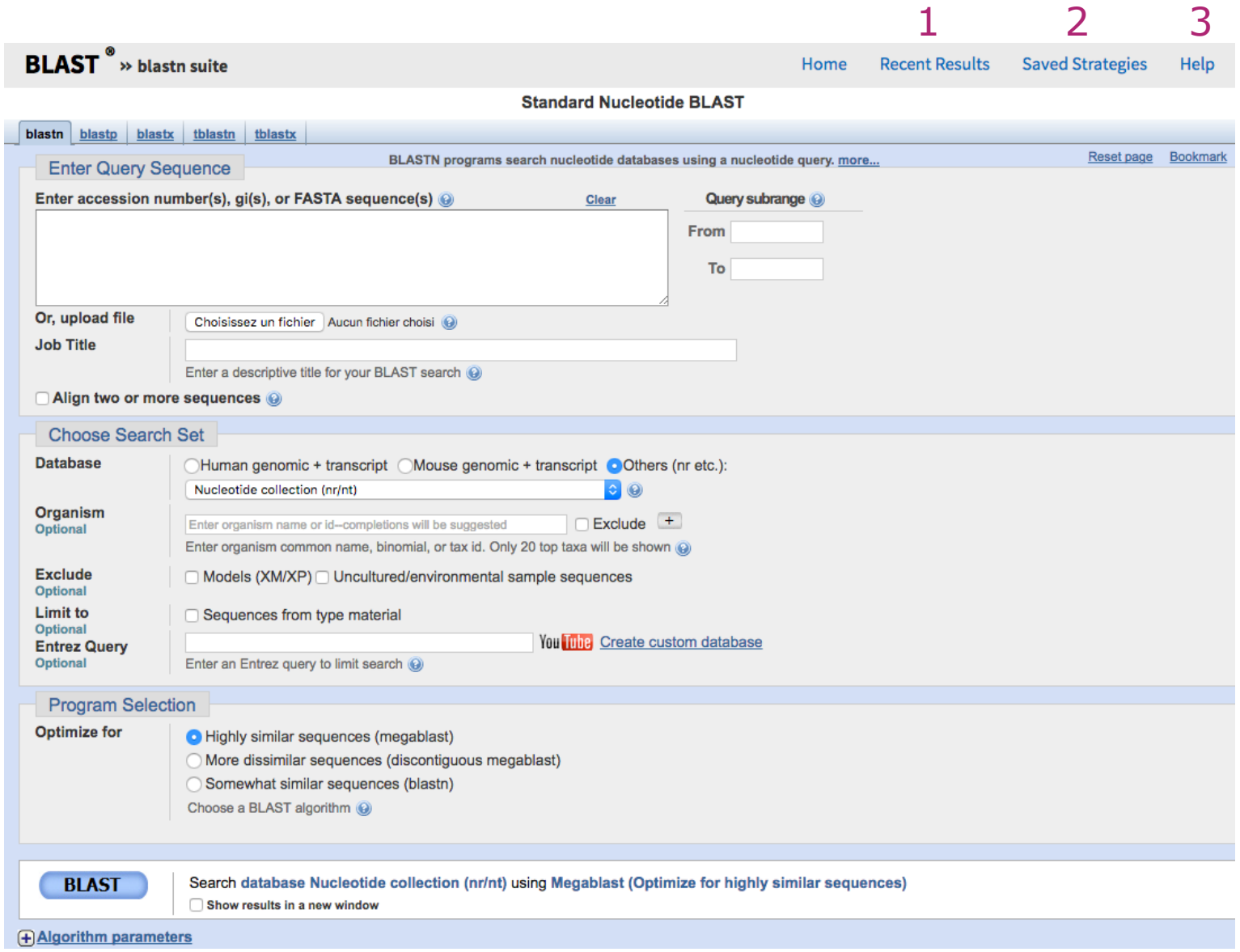
BioAssay

Search protein or nucleotide targets in PubChem BioAssay

MOLE-BLAST

Establish taxonomy for uncultured or enviromental sequences

Formulaire


1
2
3
BLAST ® » blastn suite
Home
Recent Results
Saved Strategies
Help
Standard Nucleotide BLAST
4
blastn
blastp
blastx
tblastn
tblastx
Enter Query Sequence
BLASTN programs search nucleotide databases using a nucleotide query. more...
Reset page
Bookmark
Enter accession number(s), gi(s), or FASTA sequence(s)
Clear
Query subrange
5
From
To
Or, upload file
Choisissez un fichier
Aucun fichier choisi
Job Title
Enter a descriptive title for your BLAST search
Align two or more sequences
Choose Search Set
Database
Human genomic + transcript
Mouse genomic + transcript
Others (nr etc.):
Nucleotide collection (nr/nt)
6
Organism
Optional
Enter organism name or id--completions will be suggested
Exclude
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown
Exclude
Optional
Models (XM/XP)
Uncultured/environmental sample sequences
Limit to
Optional
Sequences from type material
Entrez Query
Optional
Create custom database
Enter an Entrez query to limit search
Program Selection
Optimize for
7
Highly similar sequences (megablast)
More dissimilar sequences (discontiguous megablast)
Somewhat similar sequences (blastn)
Choose a BLAST algorithm
BLAST
Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar sequences)
Show results in a new window
8
Algorithm parameters

Formulaire

- **1 Recent Results :** résultats de vos recherches des dernières 36h. Si vous êtes enregistrés sur MyNCBI, vous pouvez accéder à vos résultats de n'importe qu'elle machine. Dans le cas contraire ne sont conservés que les recherches de la session active du navigateur

- **2 Saved Strategies :** vous permet de sauver les paramètres d'une recherche Blast afin de relancer une recherche avec les même paramètres ultérieurement (connexion à MyNCBI requise)

- **3 Help :** documentations, liens, et tutoriels

- **4 :** type de Blast choisi

5 Enter Query Sequence

Copiez/collez ou uploadez votre ou vos séquences requêtes. Vous pouvez également définir une plage de recherche dans vos séquences. Vous pouvez donner un titre à votre recherche

La fonction « Align two or more sequences » vous permet de comparer des séquences entre-elles sans passer par une base de donnée

6 Choose Search Set

Sélectionnez votre base de donnée. Vous pouvez limiter votre recherche à certains organismes ou exclure des organismes. Vous pouvez exclure les séquences produites à partir des projets d'annotation de génomes ou provenant d'organisme non cultivés/élevés. Vous pouvez limiter votre recherche aux spécimens et souches modèles

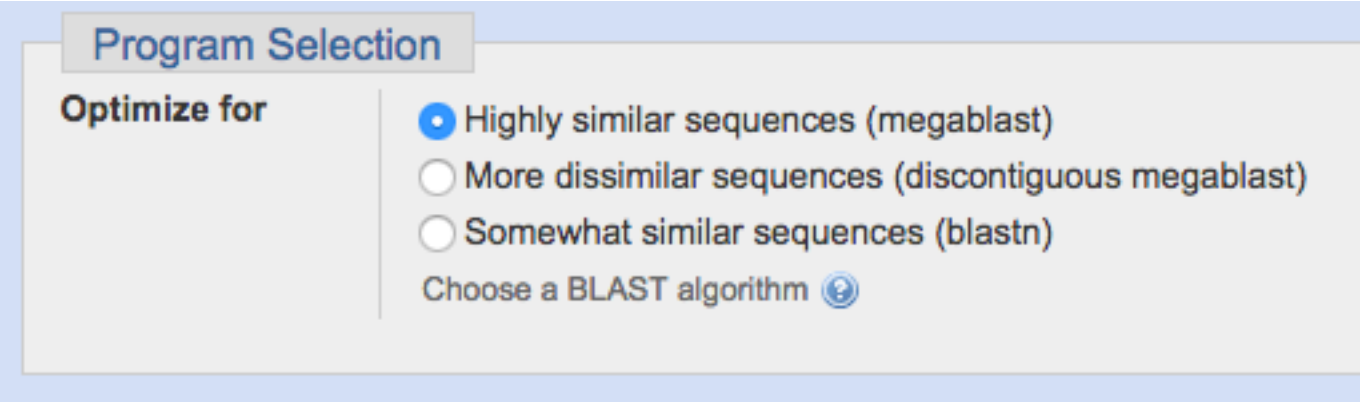
7 Program Selection

Vous permet d'optimiser votre recherche pour différents scénarios (exemple , recherches intra ou inter espèces)

8 Algorithm parameters

C’est l’endroit modifier les paramètres de l’agoritme de BLAST qui a été sélectionné (voir section dédiée)

Blast nucléotidique



- **Megablast**:

- Un Blast plus rapide lorsqu'on recherche une grande similarité

- Mise en œuvre : utiliser des mots de taille plus grande (28 contre 11)

- A réserver lorsque l'on recherche des séquences très proches ou lorsque l'on veut savoir si notre séquence est dans la banque

- **Discontigous megablast**:

- Utiliser une graine espacée plutôt qu'un mot exact (graine contiguë)

- Utile pour des comparaisons inter-espèces

- Exemple de graine contiguë : 1 1 1 1 1 : un mot exact (sans mismatch) de 5 nucléotides

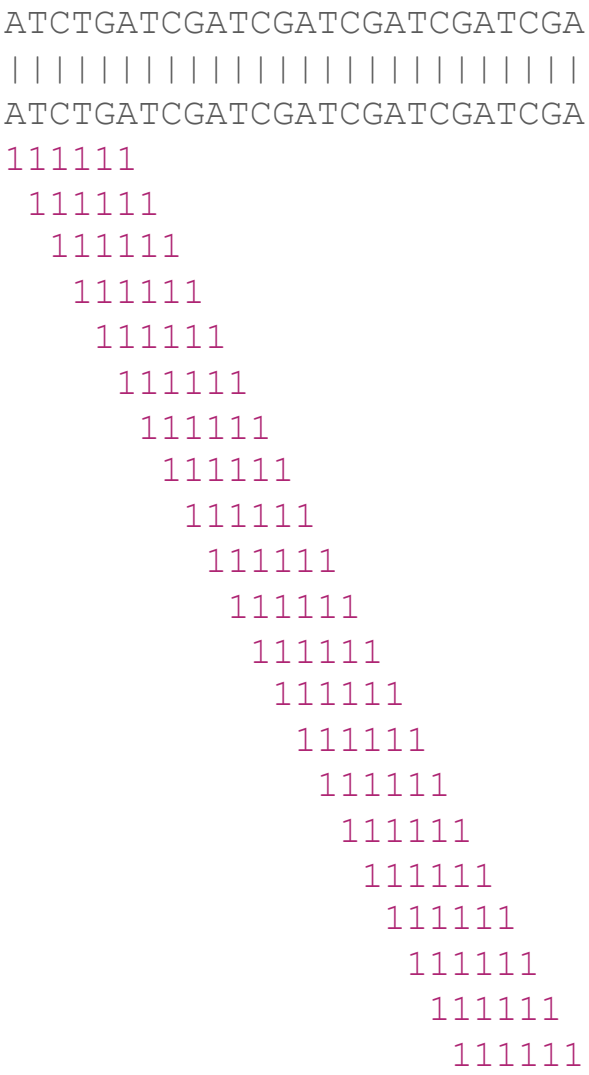
- Exemple de graine espacée : 1 0 1 1 0 1 1 : un mot de 7 nucléotides, les positions 2 et 5 peuvent représenter des mésappariements

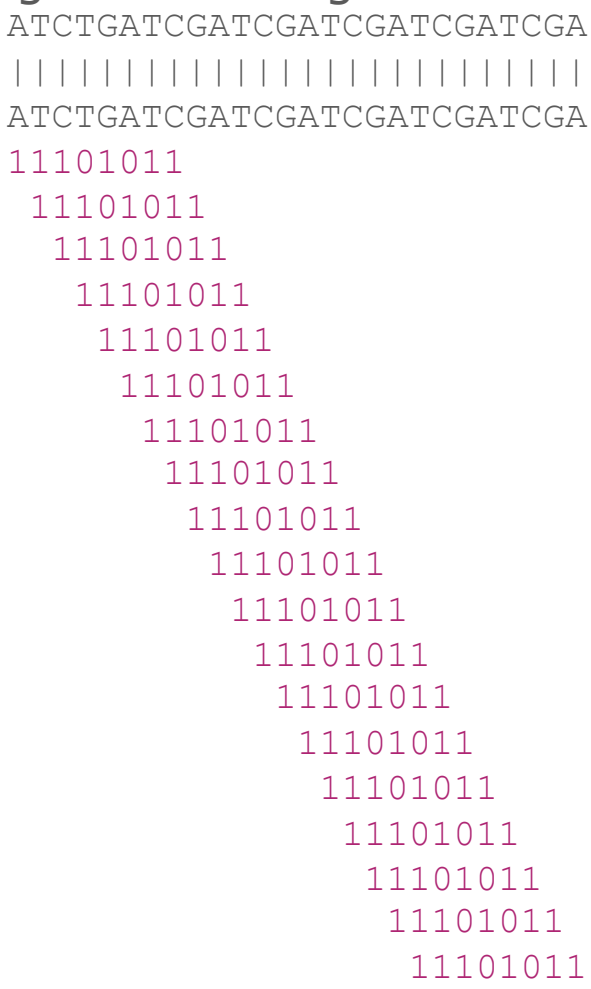
Graines espacées *vs* graines contiguës

- Soit une séquence *q* de longueur *l*=26
- Soit un graine (mot) de taille 6
- On peut donc définir un maximum de 26-6+1=21 graines
- La séquence *ti est identique à q : donc toutes les graines peuvent s'aligner sur ti*

```
ATCTGATCGATCGATCGATCGATCGA : q
||||||||||||||||||||||||||
ATCTGATCGATCGATCGATCGATCGA : ti
111111
 111111
  111111
   111111
    111111
     111111
      111111
       111111
        111111
         111111
          111111
           111111
            111111
             111111
              111111
               111111
                111111
                 111111
                  111111
                   111111
                    111111
```

Graines espacées *vs* graines contiguës





- Les graines espacées se comportent ici de la même manière que les graines contiguës

Graines espacées *vs* graines contiguës

- Introduisons un mismatch entre *q* et *ti* …

```
ATCTGATCGATCGATCGATCGA
|||||.||||||||||||||||
ATCTGCTCGATCGATCGATCGA
111111
 111111
  111111
   111111
    111111
     111111
      111111
       111111
        111111
         111111
          111111
           111111
            111111
             111111
              111111
               111111
                111111
                 111111
                  111111
                   111111
                    111111
                     111111
```

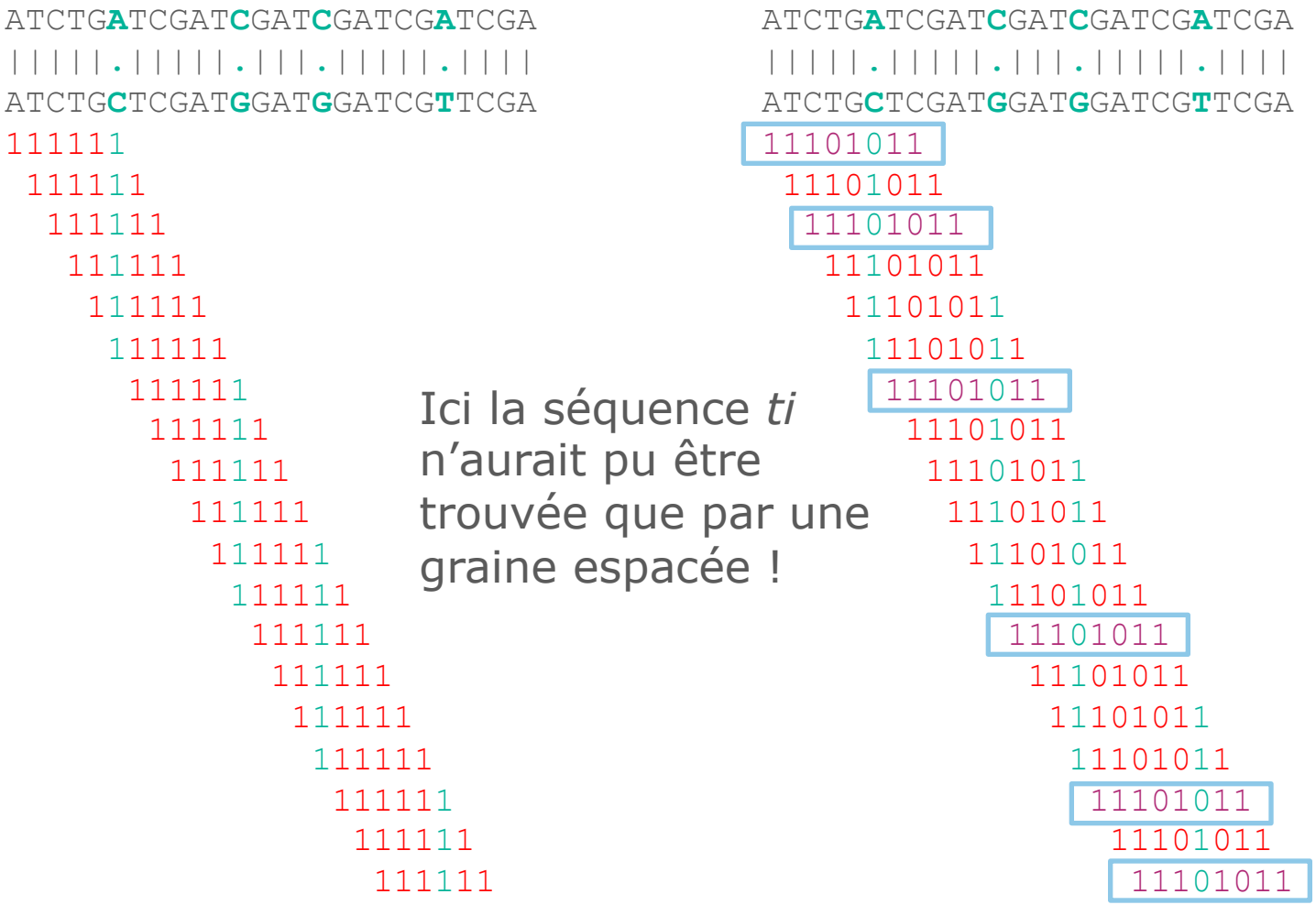
```
ATCTGATCGATCGATCGATCGA
|||||.||||||||||||||||
ATCTGCTCGATGGATGGATCGTTCGA
11101011
 11101011
  11101011
   11101011
    11101011
     11101011
      11101011
       11101011
        11101011
         11101011
          11101011
           11101011
            11101011
             11101011
              11101011
               11101011
                11101011
                 11101011
                  11101011
```

Certaines graines (en rouge) sont perdue

Les graines espacée peuvent être permissives

Graines espacées *vs* graines contiguës

- Introduisons davantage de mismatchs entre *q* et *ti* … …



Ici la séquence *ti* n'aurait pu être trouvée que par une graine espacée !

Blast protéique

- **PSI-BLAST :**

 - Recherche initiale avec blastp

 - Construction d'un alignement multiple puis d'un profil à partir des meilleurs hits → matrice de score position-spécifique (PSSM)

 - Nouvelle recherche avec le profil

- **PHI-BLAST :**

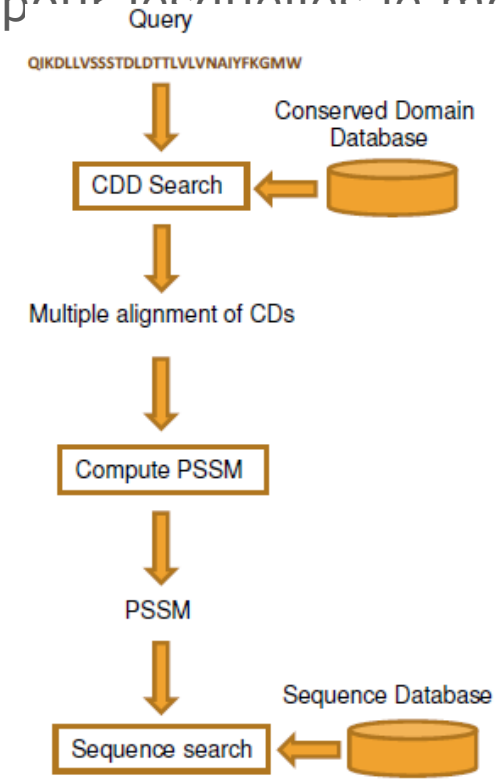
  - Entrée : une séquence protéique et un motif (expression régulière)

  - Restriction de la banque aux séquences pour lesquelles le motif est retrouvé

    - DELTA-BLAST

  - utilisation de PSSM construites à partir d'une base de donnée de domaines conservés (NCBI CDD: conserved domain database)

  - Plus rapide que PSI-BLAST, plus sensible également

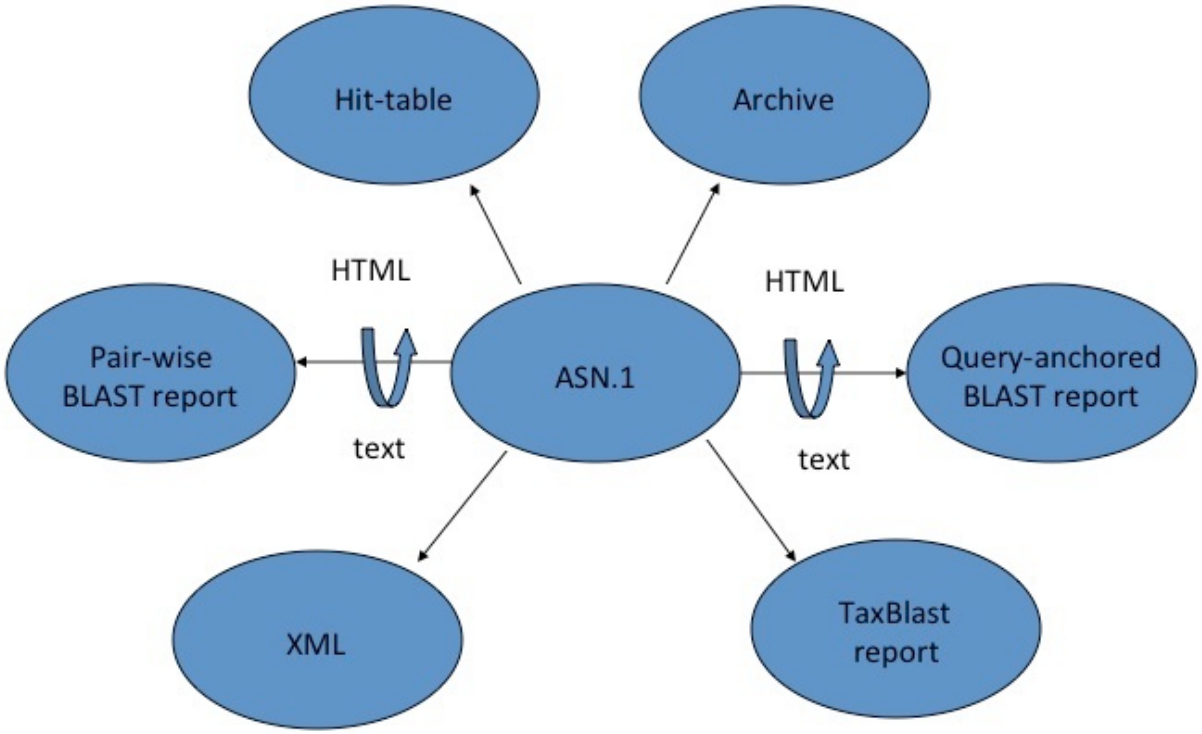


Boratyn et al 2012

Résultats de Blast

Résultats

- Des résultats structurés : une sortie fléxible

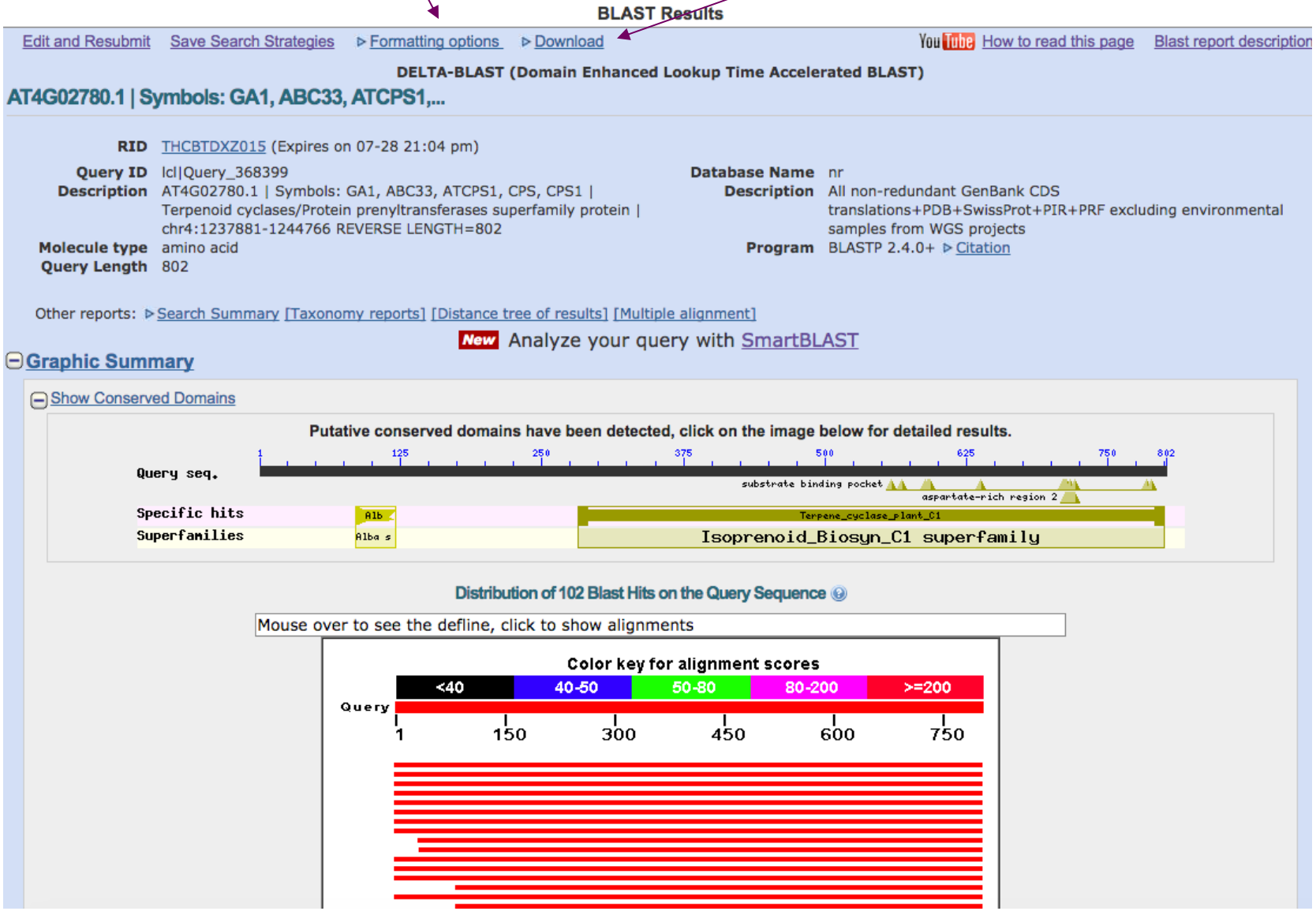


Madden, 2013

Résultats

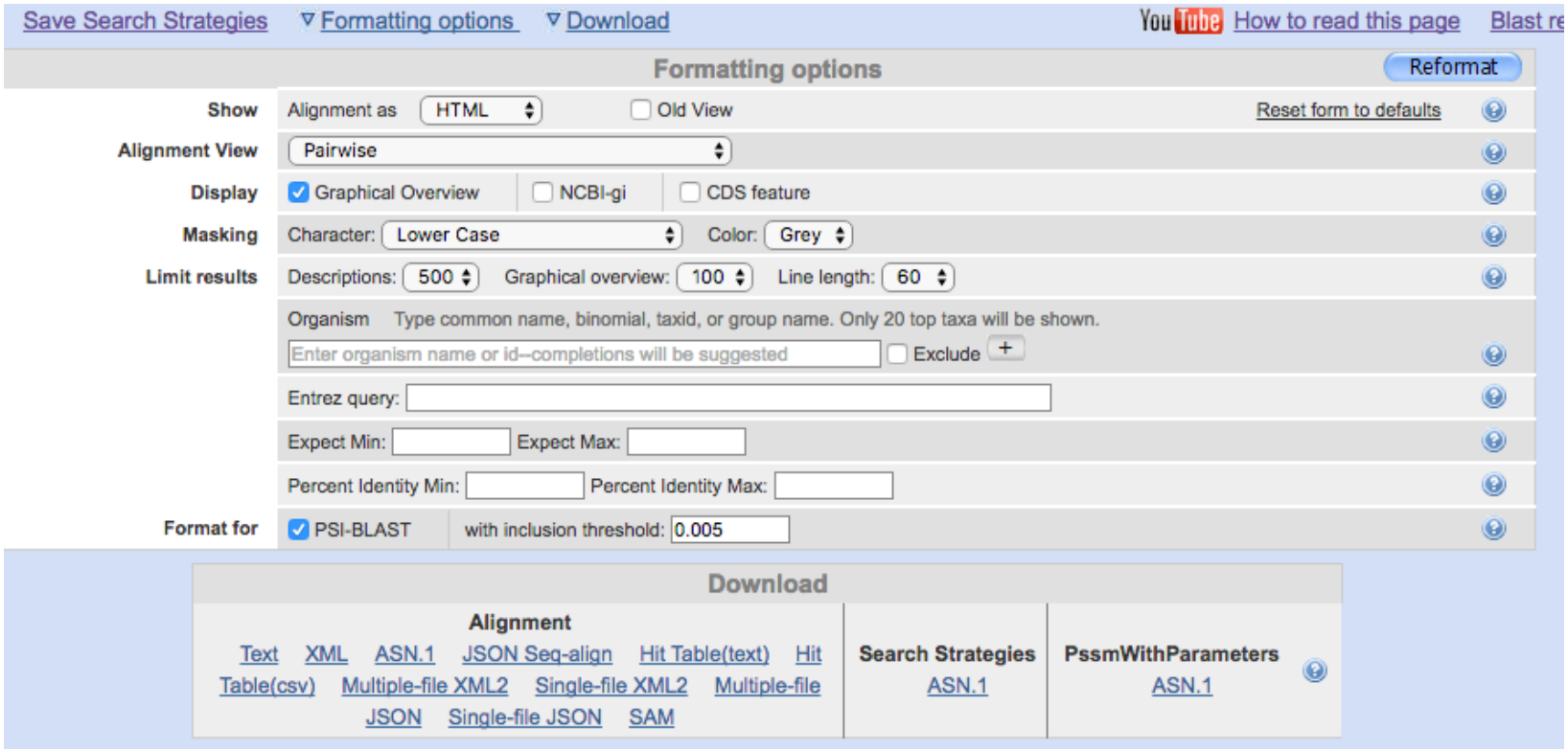
Formatage des résultats

Téléchargements



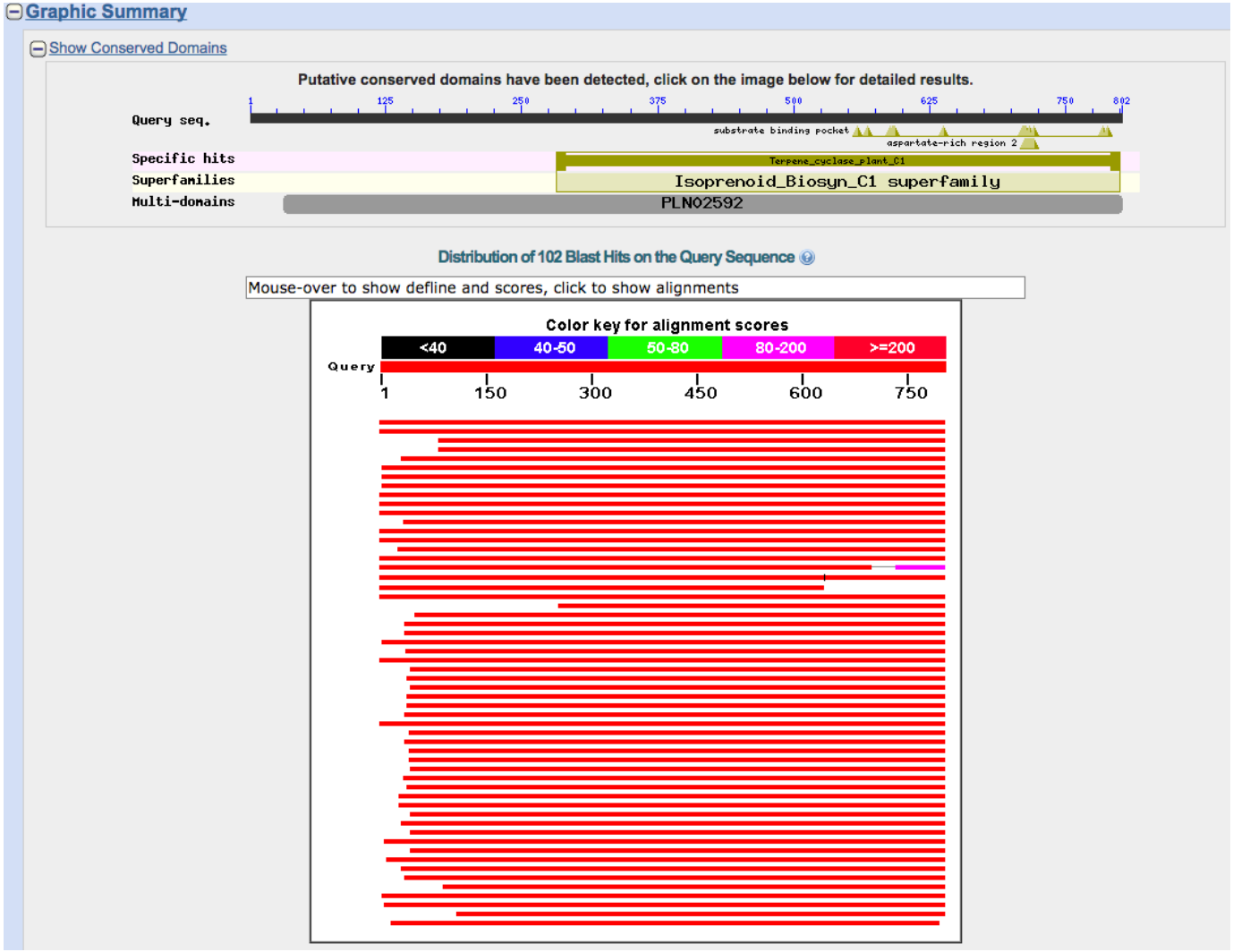
Résultats

- Différentes possibilités de format et d'export…



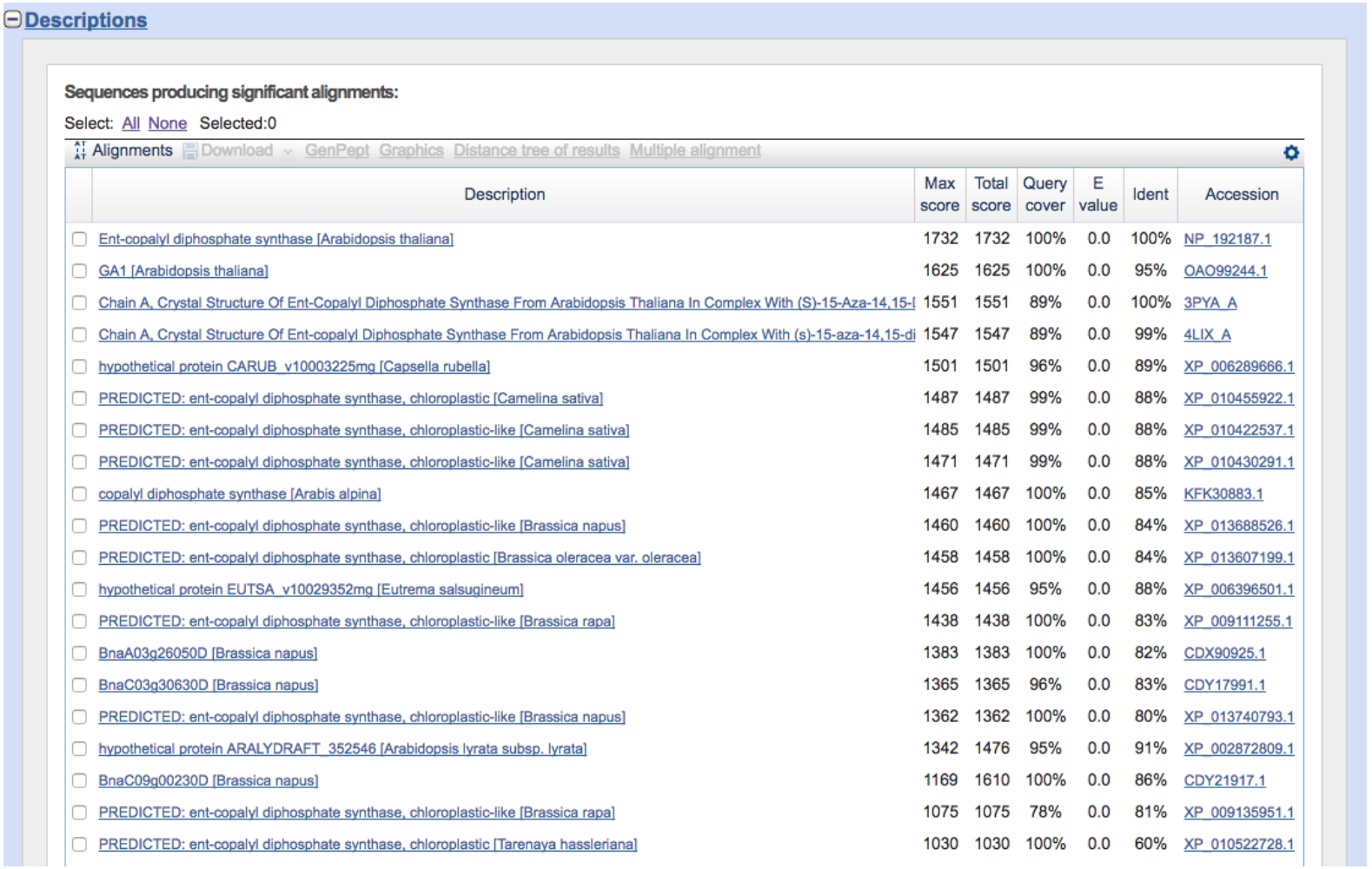
Résultats

- Graphic summary



Résultats

Descriptions



Descriptions

Sequences producing significant alignments:

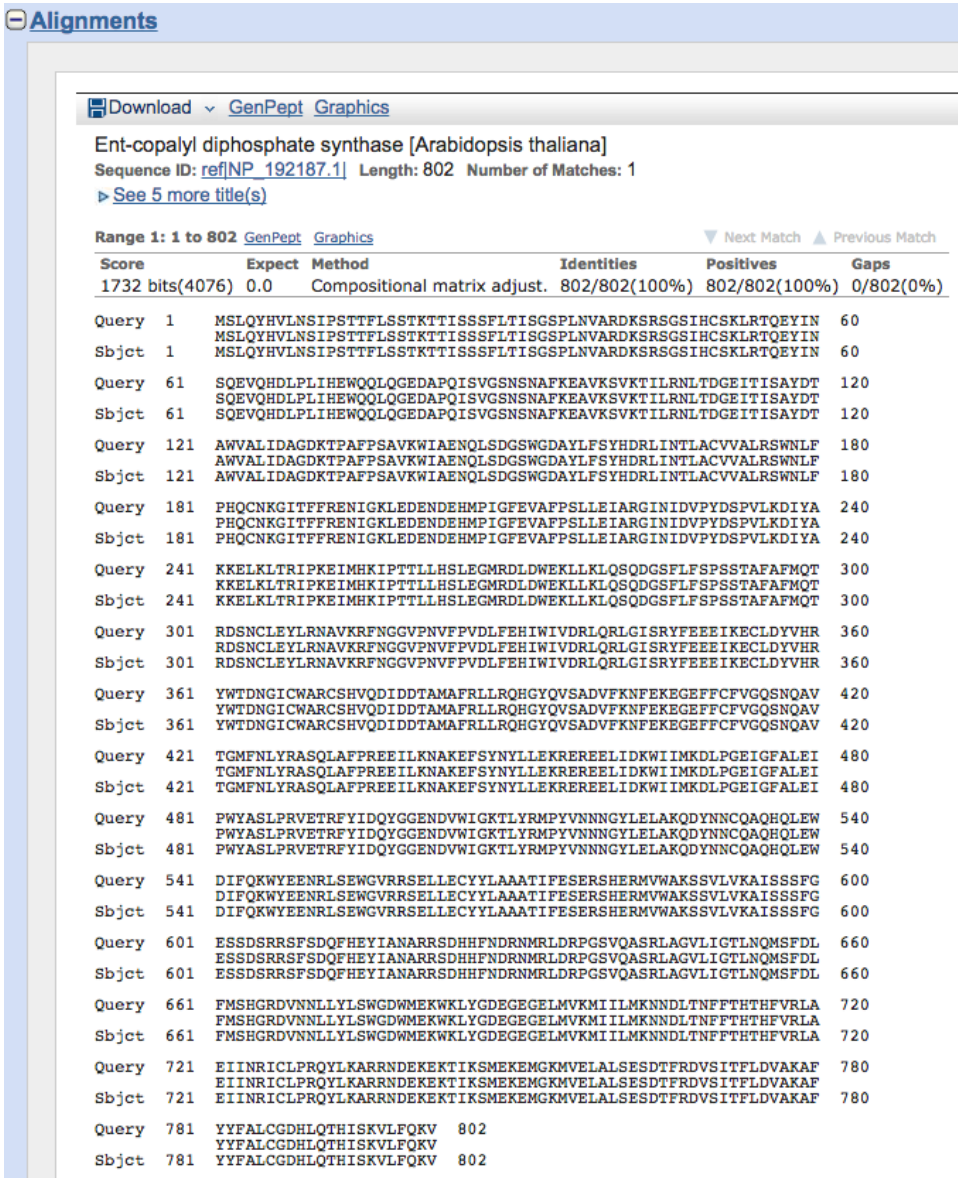
Select: All None Selected:0

Alignments Download GenPept Graphics Distance tree of results Multiple alignment

Description	Max score	Total score	Query cover	E value	Ident	Accession
Ent-copalyl diphosphate synthase [Arabidopsis thaliana]	1732	1732	100%	0.0	100%	NP_192187.1
GA1 [Arabidopsis thaliana]	1625	1625	100%	0.0	95%	OAO99244.1
Chain A, Crystal Structure Of Ent-Copalyl Diphosphate Synthase From Arabidopsis Thaliana In Complex With (S)-15-Aza-14,15-[	1551	1551	89%	0.0	100%	3PYA_A
Chain A, Crystal Structure Of Ent-copalyl Diphosphate Synthase From Arabidopsis Thaliana In Complex With (s)-15-aza-14,15-di	1547	1547	89%	0.0	99%	4LIX_A
hypothetical protein CARUB_v10003225mg [Capsella rubella]	1501	1501	96%	0.0	89%	XP_006289666.1
PREDICTED: ent-copalyl diphosphate synthase, chloroplastic [Camelina sativa]	1487	1487	99%	0.0	88%	XP_010455922.1
PREDICTED: ent-copalyl diphosphate synthase, chloroplastic-like [Camelina sativa]	1485	1485	99%	0.0	88%	XP_010422537.1
PREDICTED: ent-copalyl diphosphate synthase, chloroplastic-like [Camelina sativa]	1471	1471	99%	0.0	88%	XP_010430291.1
copalyl diphosphate synthase [Arabis alpina]	1467	1467	100%	0.0	85%	KFK30883.1
PREDICTED: ent-copalyl diphosphate synthase, chloroplastic-like [Brassica napus]	1460	1460	100%	0.0	84%	XP_013688526.1
PREDICTED: ent-copalyl diphosphate synthase, chloroplastic [Brassica oleracea var. oleracea]	1458	1458	100%	0.0	84%	XP_013607199.1
hypothetical protein EUTSA_v10029352mg [Eutrema salsugineum]	1456	1456	95%	0.0	88%	XP_006396501.1
PREDICTED: ent-copalyl diphosphate synthase, chloroplastic-like [Brassica rapa]	1438	1438	100%	0.0	83%	XP_009111255.1
BnaA03g26050D [Brassica napus]	1383	1383	100%	0.0	82%	CDX90925.1
BnaC03g30630D [Brassica napus]	1365	1365	96%	0.0	83%	CDY17991.1
PREDICTED: ent-copalyl diphosphate synthase, chloroplastic-like [Brassica napus]	1362	1362	100%	0.0	80%	XP_013740793.1
hypothetical protein ARALYDRAFT_352546 [Arabidopsis lyrata subsp. lyrata]	1342	1476	95%	0.0	91%	XP_002872809.1
BnaC09g00230D [Brassica napus]	1169	1610	100%	0.0	86%	CDY21917.1
PREDICTED: ent-copalyl diphosphate synthase, chloroplastic-like [Brassica rapa]	1075	1075	78%	0.0	81%	XP_009135951.1
PREDICTED: ent-copalyl diphosphate synthase, chloroplastic [Tarenaya hassleriana]	1030	1030	100%	0.0	60%	XP_010522728.1

Résultats

- Alignements



Alignments

Download GenPept Graphics

Ent-copalyl diphosphate synthase [Arabidopsis thaliana]

Sequence ID: ref|NP_192187.1| Length: 802 Number of Matches: 1

See 5 more title(s)

Range 1: 1 to 802 GenPept Graphics Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps
1732 bits(4076)	0.0	Compositional matrix adjust.	802/802(100%)	802/802(100%)	0/802(0%)

```
Query  1    MSLQYHVLNSIPSTTFLSSTKTTISSSFLTISGSPLNVARDKSRSGSIHCSKLRTQEYIN  60
            MSLQYHVLNSIPSTTFLSSTKTTISSSFLTISGSPLNVARDKSRSGSIHCSKLRTQEYIN
Sbjct  1    MSLQYHVLNSIPSTTFLSSTKTTISSSFLTISGSPLNVARDKSRSGSIHCSKLRTQEYIN  60

Query  61   SQEVQHDLPLIHEWQQLQGEDAPQISVGSNSNAFKEAVKSVKTILRNLTDGEITISAYDT  120
            SQEVQHDLPLIHEWQQLQGEDAPQISVGSNSNAFKEAVKSVKTILRNLTDGEITISAYDT
Sbjct  61   SQEVQHDLPLIHEWQQLQGEDAPQISVGSNSNAFKEAVKSVKTILRNLTDGEITISAYDT  120

Query  121  AWVALIDAGDKTPAFPSAVKWIAENQLSDGSWGDAYLFSYHDRLINTLACVVALRSWNLF  180
            AWVALIDAGDKTPAFPSAVKWIAENQLSDGSWGDAYLFSYHDRLINTLACVVALRSWNLF
Sbjct  121  AWVALIDAGDKTPAFPSAVKWIAENQLSDGSWGDAYLFSYHDRLINTLACVVALRSWNLF  180

Query  181  PHQCNKGITFFRENIGKLEDENDEHMPIGFEVAFPSLLEIARGINIDVPYDSPVLKDIYA  240
            PHQCNKGITFFRENIGKLEDENDEHMPIGFEVAFPSLLEIARGINIDVPYDSPVLKDIYA
Sbjct  181  PHQCNKGITFFRENIGKLEDENDEHMPIGFEVAFPSLLEIARGINIDVPYDSPVLKDIYA  240

Query  241  KKELKLTRIPKEIMHKIPTTLLHSLEGMRDLDWEKLLKLQSQDGSFLFSPSSTAFAFMQT  300
            KKELKLTRIPKEIMHKIPTTLLHSLEGMRDLDWEKLLKLQSQDGSFLFSPSSTAFAFMQT
Sbjct  241  KKELKLTRIPKEIMHKIPTTLLHSLEGMRDLDWEKLLKLQSQDGSFLFSPSSTAFAFMQT  300

Query  301  RDSNCLEYLRNAVKRFNGGVPNVFPVDLFEHIWIVDRLQRLGISRYFEEEIKECLDYVHR  360
            RDSNCLEYLRNAVKRFNGGVPNVFPVDLFEHIWIVDRLQRLGISRYFEEEIKECLDYVHR
Sbjct  301  RDSNCLEYLRNAVKRFNGGVPNVFPVDLFEHIWIVDRLQRLGISRYFEEEIKECLDYVHR  360

Query  361  YWTDNGICWARCSHVQDIDDTAMAFRLLRQHGYQVSADVFKNFEKEGEFFCFVGQSNQAV  420
            YWTDNGICWARCSHVQDIDDTAMAFRLLRQHGYQVSADVFKNFEKEGEFFCFVGQSNQAV
Sbjct  361  YWTDNGICWARCSHVQDIDDTAMAFRLLRQHGYQVSADVFKNFEKEGEFFCFVGQSNQAV  420

Query  421  TGMFNLYRASQLAFPREEILKNAKEFSYNYLLEKREREELIDKWIIMKDLPGEIGFALEI  480
            TGMFNLYRASQLAFPREEILKNAKEFSYNYLLEKREREELIDKWIIMKDLPGEIGFALEI
Sbjct  421  TGMFNLYRASQLAFPREEILKNAKEFSYNYLLEKREREELIDKWIIMKDLPGEIGFALEI  480

Query  481  PWYASLPRVETRFYIDQYGGENDVWIGKTLYRMPYVNNNGYLELAKQDYNNCQAQHQLEW  540
            PWYASLPRVETRFYIDQYGGENDVWIGKTLYRMPYVNNNGYLELAKQDYNNCQAQHQLEW
Sbjct  481  PWYASLPRVETRFYIDQYGGENDVWIGKTLYRMPYVNNNGYLELAKQDYNNCQAQHQLEW  540

Query  541  DIFQKWYEENRLSEWGVRRSELLECYYLAAATIFESERSHERMVWAKSSVLVKAISSSFG  600
            DIFQKWYEENRLSEWGVRRSELLECYYLAAATIFESERSHERMVWAKSSVLVKAISSSFG
Sbjct  541  DIFQKWYEENRLSEWGVRRSELLECYYLAAATIFESERSHERMVWAKSSVLVKAISSSFG  600

Query  601  ESSDSRRSFSDQFHEYIANARRSDHHFNDRNMRLDRPGSVQASRLAGVLIGTLNQMSFDL  660
            ESSDSRRSFSDQFHEYIANARRSDHHFNDRNMRLDRPGSVQASRLAGVLIGTLNQMSFDL
Sbjct  601  ESSDSRRSFSDQFHEYIANARRSDHHFNDRNMRLDRPGSVQASRLAGVLIGTLNQMSFDL  660

Query  661  FMSHGRDVNNLLYLSWGDWMEKWKLYGDEGEGELMVKMIILMKNNDLTNFFTHTHFVRLA  720
            FMSHGRDVNNLLYLSWGDWMEKWKLYGDEGEGELMVKMIILMKNNDLTNFFTHTHFVRLA
Sbjct  661  FMSHGRDVNNLLYLSWGDWMEKWKLYGDEGEGELMVKMIILMKNNDLTNFFTHTHFVRLA  720

Query  721  EIINRICLPRQYLKARRNDEKEKTIKSMEKEMGKMVELALSESDTFRDVSITFLDVAKAF  780
            EIINRICLPRQYLKARRNDEKEKTIKSMEKEMGKMVELALSESDTFRDVSITFLDVAKAF
Sbjct  721  EIINRICLPRQYLKARRNDEKEKTIKSMEKEMGKMVELALSESDTFRDVSITFLDVAKAF  780

Query  781  YYFALCGDHLQTHISKVLFQKV  802
            YYFALCGDHLQTHISKVLFQKV
Sbjct  781  YYFALCGDHLQTHISKVLFQKV  802
```

Résultats

- Alignements

HSPs

Download GenPept Graphics Sort by: E value

hypothetical protein ARALYDRAFT_352546 [Arabidopsis lyrata subsp. lyrata]

Sequence ID: ref|XP_002872809.1| Length: 742 Number of Matches: 2

See 1 more title(s)

Range 1: 1 to 697 GenPept Graphics — Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps
1342 bits(3158)	0.0	Compositional matrix adjust.	640/700(91%)	648/700(92%)	5/700(0%)

```
Query  1    MSLQYHVLNSIPSTTFLSSTKTTISSSFLTISGSPLNVARDKSRSGSIHCSKLRTQEYIN  60
            MSLQYH LNSI ST FLSSTKTT SSSFLTISGSPLNVARDK RSGSI CSKLRTQEY
Sbjct  1    MSLQYHALNSIQSTNFLSSTKTTLSSSFLTISGSPLNVARDKPRSGSIYCSKLRTQEYTT  60

Query  61   SQEVQHDLPLIHEWQQLQGEDAPQISVGSNSNAFKEAVKSVKTILRNLTDGEITISAYDT  120
            SQEVQHDLPLI+ WQQLQ EDAPQIS+GSN NA  EAVKSVK ILRNLTDGEITISAYDT
Sbjct  61   SQEVQHDLPLIQ-WQQLQREDAPQISIGSNNNAIEEAVKSVKSILRNLTDGEITISAYDT  119

Query  121  AWVALIDAGDKTPAFPSAVKWIAENQLSDGSWGDAYLFSYHDRLINTLACVVALRSWNLF  180
            AWVALIDAGDKTPAFPSAVKWIAENQLSDGSWGDAYLFSYHDRLINTLACVVALRSWNLF
Sbjct  120  AWVALIDAGDKTPAFPSAVKWIAENQLSDGSWGDAYLFSYHDRLINTLACVVALRSWNLF  179

Query  181  PHQCNKGITFFRENIGKLEDENDEHMPIGFEVAFPSLLEIARGINIDVPYDSPVLKDIYA  240
            PHQC KGITFFRENIGKLEDENDEHMPIGFEVAFPSLLEIAR INIDVPYDSPVLKDIYA
Sbjct  180  PHQCHKGITFFRENIGKLEDENDEHMPIGFEVAFPSLLEIAREINIDVPYDSPVLKDIYA  239

Query  241  KKELKLTRIPKEIMHKIPTTLLHSLEGMRDLDWEKLLKLQSQDGSFLFSPSSTAFAFMQT  300
            KKELKLTRIPKEIMHKIPTTLLHSLEGMRDLDWEKLLKLQSQDGSFLFSPSSTAFAFMQT
Sbjct  240  KKELKLTRIPKEIMHKIPTTLLHSLEGMRDLDWEKLLKLQSQDGSFLFSPSSTAFAFMQT  299

Query  301  RDSNCLEYLRNAVKRFNGGVPNVFPVDLFEHIWIVDRLQRLGISRYFEEEIKECLDYVHR  360
            RDSNCL YLRNAVKRFNGGVPNVFPVDLFEHIWIVDRLQRLGISRYFEEEIKECLDYVHR
Sbjct  300  RDSNCLRYLRNAVKRFNGGVPNVFPVDLFEHIWIVDRLQRLGISRYFEEEIKECLDYVHR  359

Query  361  YWTDNGICWARCSHVQDIDDTAMAFRLLRQHGYQVSADVFKNFEKEGEFFCFVGQSNQAV  420
            YWTD GICWARCSHVQDIDDTAMAFRLLR HGYQVSADVFKNFEKEGEFFCFVGQSNQA
Sbjct  360  YWTDKGICWARCSHVQDIDDTAMAFRLLRLHGYQVSADVFKNFEKEGEFFCFVGQSNQAA  419

Query  421  TGMFNLYRASQLAFPREEILKNAKEFSYNYLLEKREREELIDKWIIMKDLPGEIGFALEI  480
            TGMFNLYRASQLAFPRE+ILKNAKEFS  YL   KRER+ELIDKWIIMKDLPGEIGFALEI
Sbjct  420  TGMFNLYRASQLAFPREDILKNAKEFSNKYLQGKRERDELIDKWIIMKDLPGEIGFALEI  479

Query  481  PWYASLPRVETRFYIDQYGGENDVWIGKTLYRMPYVNNNGYLELAKQDYNNCQAQHQLEW  540
            PWYASLPRVETRFYIDQYGGENDVWIGKTLYRMPYVNNNGYLELAKQDYNNCQA HQLEW
Sbjct  480  PWYASLPRVETRFYIDQYGGENDVWIGKTLYRMPYVNNNGYLELAKQDYNNCQALHQLEW  539

Query  541  DIFQKWYEENRLSEWGVRRSELLECYYLAAATIFESERSHERMVWAKSSVLVKAISSSFG  600
            D FQKWYEENRL EWGVRRSELLECY+LAAATIFESERSHER VWAKSSVLVKAI SSFG
Sbjct  540  DTFQKWYEENRLNEWGVRRSELLECYFLAAATIFESERSHERIVWAKSSVLVKAI-SSFG  598

Query  601  ESSDSRRSFSDQFHEYIANARRSDHHFNDRNMRLDRPGSVQASRLAGVLIGTLNQMSFDL  660
             SSDSRRSFS+QFH YIANARRSDHHFN R MRLDRPGSVQASRL G+LIGTLNQMSFDL
Sbjct  599  KSSDSRRSFSEQFHKYIANARRSDHHFNGRSMRLDRPGSVQASRLVGILIGTLNQMSFDL  658

Query  661  FMSHGRDVNNLLYLS--WGDWMEKWKLYGDEGEGELMVKM  698
            FMSHGRDV NLLY S     D  EK      E E E MV +
Sbjct  659  FMSHGRDVYNLLYQSARRNDEKEK-TIRSMETEMEKMVEL  697
```

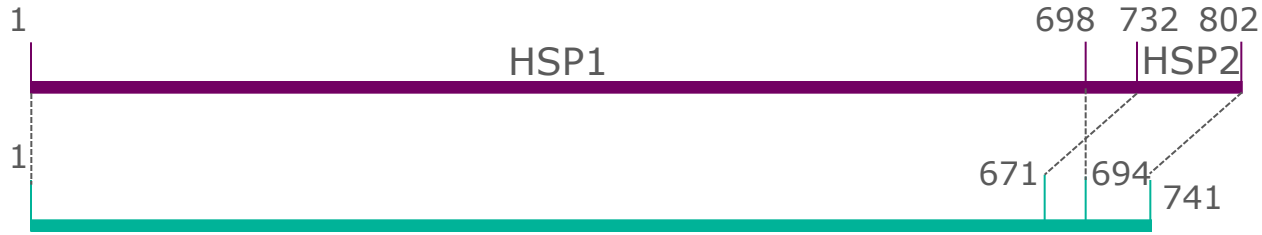
Range 2: 671 to 741 GenPept Graphics — Next Match Previous Match First Match

Score	Expect	Method	Identities	Positives	Gaps
133 bits(307)	1e-27	Compositional matrix adjust.	63/71(89%)	63/71(88%)	0/71(0%)

```
Query  732  YLKARRNDEKEKTIKSMEKEMGKMVELALSESDTFRDVSITFLDVAKAFYYFALCGDHLQ  791
            Y  ARRNDEKEKTI SME EM KMVELALSESDTFR VSITFLDVAKAFYY A CGDHLQ
Sbjct  671  YQSARRNDEKEKTIRSMETEMEKMVELALSESDTFRVVSITFLDVAKAFYYSASCGDHLQ  730

Query  792  THISKVLFQKV  802
            THISKVLFQKV
Sbjct  731  THISKVLFQKV  741
```

Blast vs Alignement global

Synthèse graphique de l'alignement Blast



Alignement global (Needle)

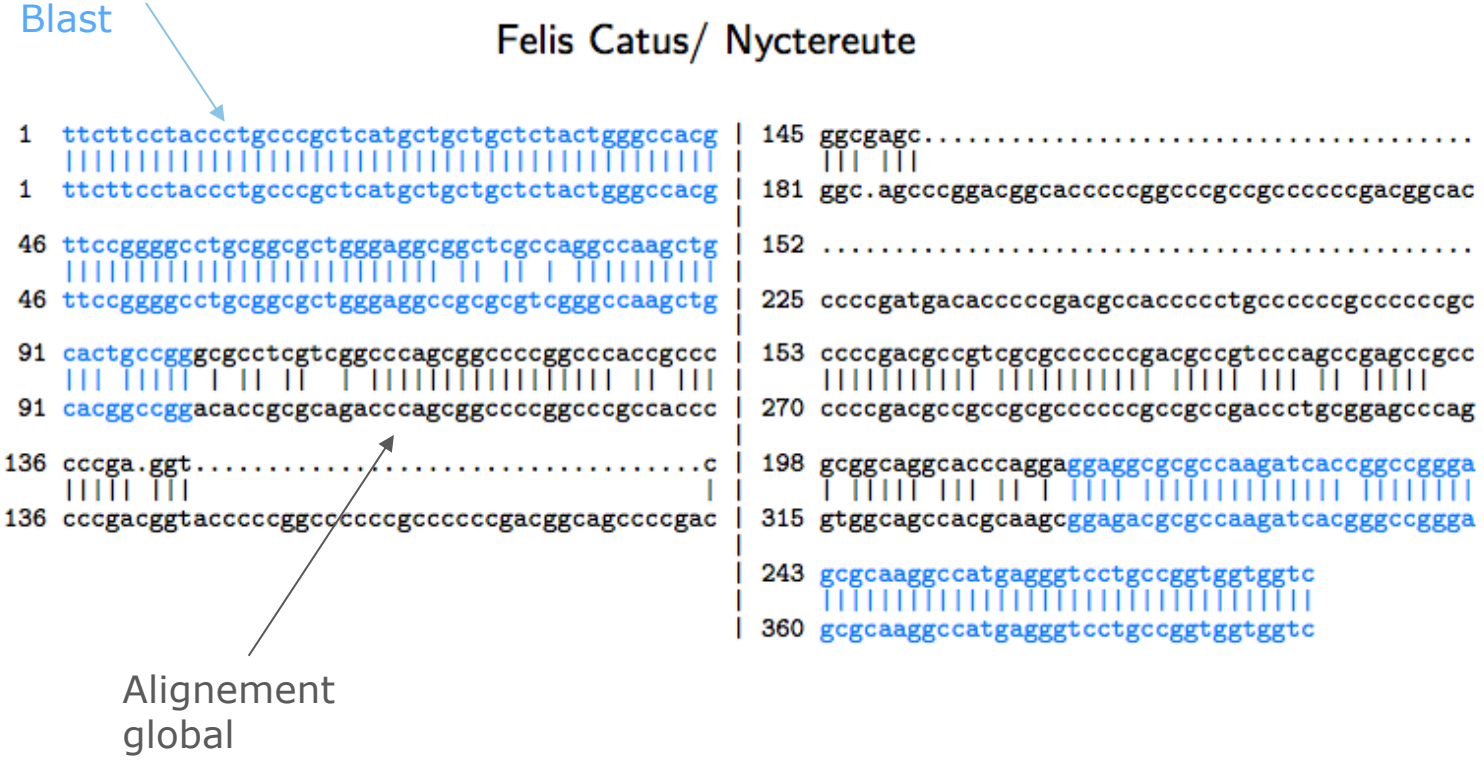
```
KSA_ARATH       501 ENDVWIGKTLYRMPYVNNNGYLELAKQDYNNCQAQHQLEWDIFQKWYEEN   550
                    ||||||||||||||||||||||||||||||||||.|||||||.||||||||
XP_002872809.   500 ENDVWIGKTLYRMPYVNNNGYLELAKQDYNNCQALHQLEWDTFQKWYEEN   549

KSA_ARATH       551 RLSEWGVRRSELLECYYLAAATIFESERSHERMVWAKSSVLVKAISSSFG   600
                    ||:|||||||||||:|||||||||||||:|||||||||||| ||||
XP_002872809.   550 RLNEWGVRRSELLECYFLAAATIFESERSHERIVWAKSSVLVKAI-SSFG   598

KSA_ARATH       601 ESSDSRRSFSDQFHEYIANARRSDHHFNDRNMRLDRPGSVQASRLAGVLI   650
                    :|||||||||:|||:|||||||||||||.|:||||||||||||||.|:||
XP_002872809.   599 KSSDSRRSFSEQFHKYIANARRSDHHFNGRSMRLDRPGSVQASRLVGILI   648

KSA_ARATH       651 GTLNQMSFDLFMSHGRDVNNLLYLSWGDWMEKWKLYGDEGEGELMVKMII   700
                    ||||||||||||||||||.||||.|
XP_002872809.   649 GTLNQMSFDLFMSHGRDVYNLLYQS------------------------   673

KSA_ARATH       701 LMKNNDLTNFFTHTHFVRLAEIINRICLPRQYLKARRNDEKEKTIKSMEK   750
                                                  ||||||||||||:|||.
XP_002872809.   674 ---------------------------------ARRNDEKEKTIRSMET   689

KSA_ARATH       751 EMGKMVELALSESDTFRDVSITFLDVAKAFYYFALCGDHLQTHISKVLFQ   800
                    ||.|||||||||||||.|||||||||||||||.|.|||||||||||||||
XP_002872809.   690 EMEKMVELALSESDTFRVVSITFLDVAKAFYYSASCGDHLQTHISKVLFQ   739

KSA_ARATH       801 KV-    802
                    ||
XP_002872809.   740 KVL    742
```

Fin HSP1

Début HSP2

Blast vs Alignement global

Alignement donné par Blast



Alignement global

Sylvain Legrand
Maître de Conférences
UMR CNRS 8198 EVO-ECO-PALEO
Evolution, Ecologie et Paléontologie
Université de Lille - Faculté des Sciences et Technologies
Bât SN2, bureau 208 - 59655 Villeneuve d'Ascq

sylvain.legrand@univ-lille.fr | http://eep.univ-lille.fr/
Tél. +33 (0)3 20 43 40 16