Blast

Basic Local Alignment Search Tool

Adapted from the courses of the Bonsai team,

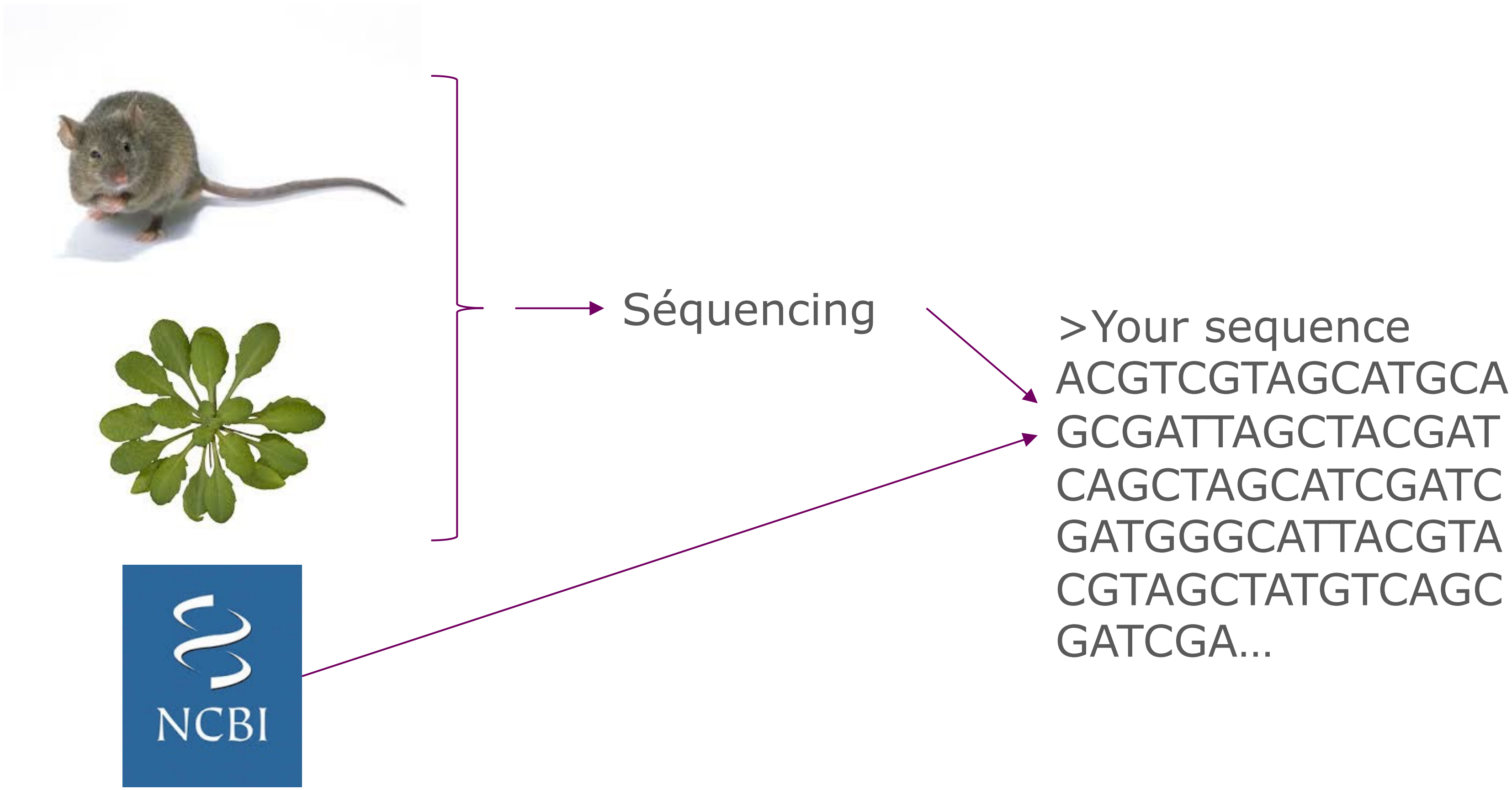
CRIStAL UMR 9189

Sylvain.legrand@univ-lille.fr

Introduction

challenge

- We assumed that you have a **nucleotide or protein sequence** obtained from a biological sample or from a database



Séquencing

>Your sequence
ACGTCGTAGCATGCA
GCGATTAGCTACGAT
CAGCTAGCATCGATC
GATGGGCATTACGTA
CGTAGCTATGTCAGC
GATCGA...

challenge

- You want to know if the sequence you have obtained is **already known or is similar** to other sequences in the databases

- A **sequence similarity searc**h often provides the first information about a new nucleotide or protein sequence

→ **inferring the function** from similar sequences

challenge

- We have:

  a **query** sequence *q*

  a **database** *T* = {*t1*, ..,*tn*}

- What we want: to find **significant alignments** between *q* and *ti*

- Classical algorithms (e.g. Smith and Waterman's local alignment) do not work: too time consuming, need to find **workarounds**

Blast, generalities

- Blast (NCBI definition) : The Basic Local Alignment Search Tool (BLAST) finds regions of **local similarity between sequences**. The program compares nucleotide or protein sequences to sequence databases and calculates the **statistical significance** of the alignments

- Blast can be used to infer **functional and evolutionary relationships** between sequences and can also help identify **members of a gene family**

- Blast uses **heuristics*** to deliver results quickly

*A heuristic is a computational method that quickly provides a feasible solution, not necessarily an optimal one, it may miss some results

Blast, generalities

- Since most proteins are **modular** (composed of functional domain(s)), Blast is made to find these domains between different sequences .

- The algorithm also allows the alignment between **mRNA and genomic sequences**

- However, if 2 sequences are expected to be aligned to their full length (global alignment), it is possible that Blast will only return the most conserved parts of this alignment

History

- First version released by NCBI in 1990 (Altschul et al. 1990)

- This version only performs **ungapped alignments**, but provides a p-value that allows the user to assess the significance of the results

- A version allowing **gaps** (Blast2) appeared in 1997 (Altschul et al. 1997) and included the **PSI Blast** (see below)

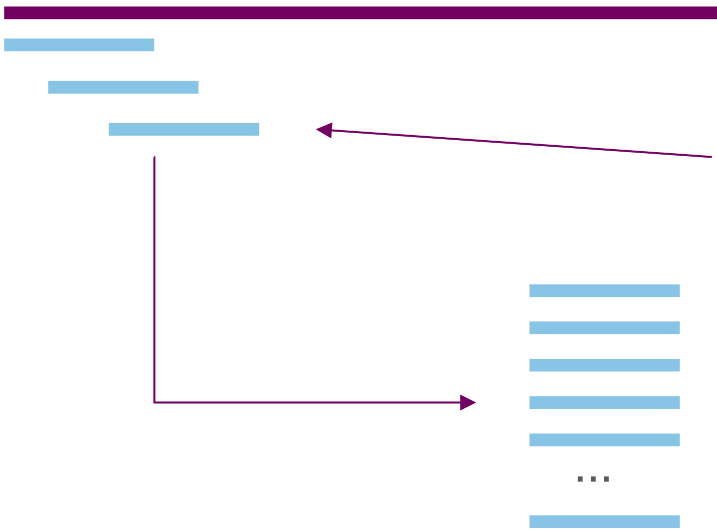
- In 2009, NCBI released a new version of Blast (**BLAST+**) (Camacho et al. 2009)

- **Currently BLAST+2.11.0 released (April 2021)**

Algorithm

Algorithm

- **<u>1st step</u>**: define from the query sequence *q* **a list of words** (seeds) of defined size *w* (default size of 11 for DNA and 3 for proteins)



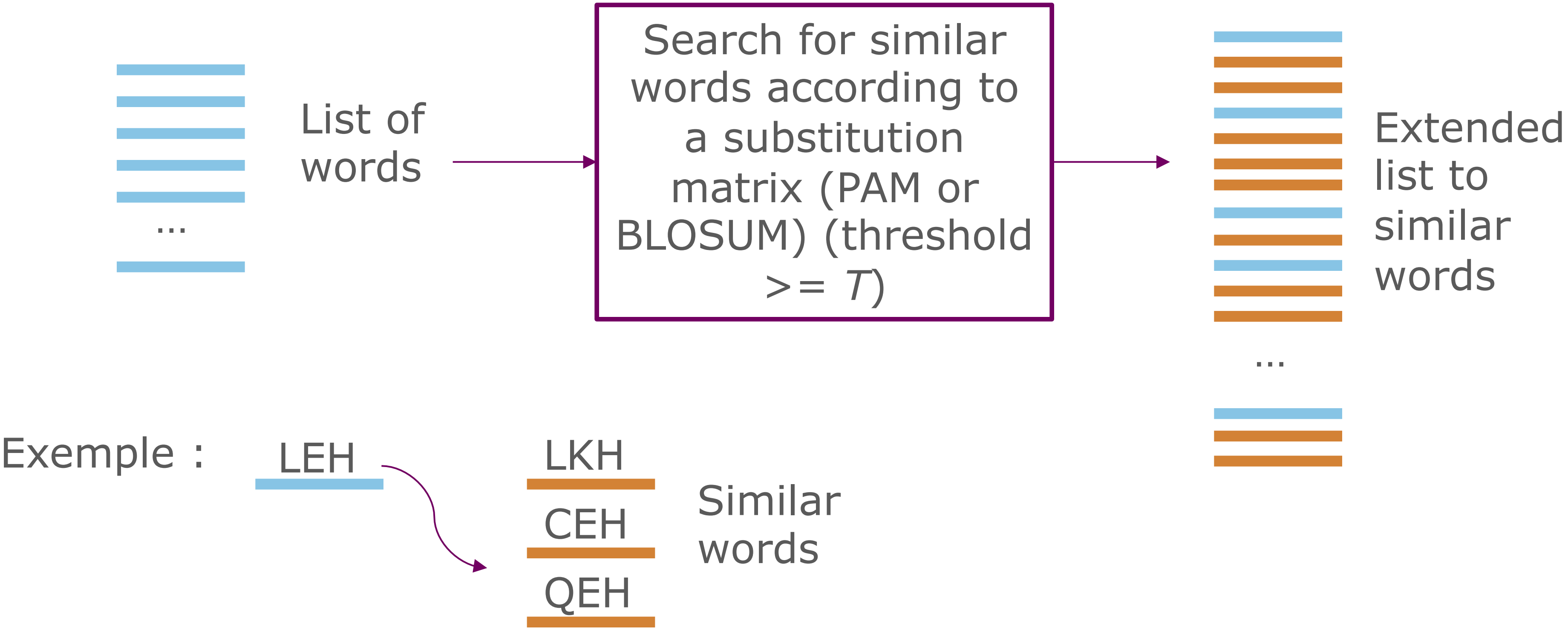
Sequence *q* of length *L*

Maximum of *L*-*w*+1 words for sequence *q*

list of words

Algorithm

- Particularity for **proteins**



List of words

…

Search for similar words according to a substitution matrix (PAM or BLOSUM) (threshold >= *T*)

Extended list to similar words

…

Exemple :

LEH

LKH

CEH

QEH

Similar words

Algorithm

- Clarification on **similar words**

- For each word of size *w*=3, Blast generates the neighbouring words using a BLOSUM62 matrix with a score threshold *T*=11

- Words with 3 amino acids: 20^3 possible alignments !

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

- The neighbouring words are aligned with LEH and the alignment score is calculated from the BLOSUM matrix62

- Only words with a score ≥ to the threshold *T* are conserved

Substitution matrix

- A substitution matrix is used to **associate a score to each pair of residues** in an alignment

- For **nucleotide** sequences, **identical penalti**es are generally used for all substitutions

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

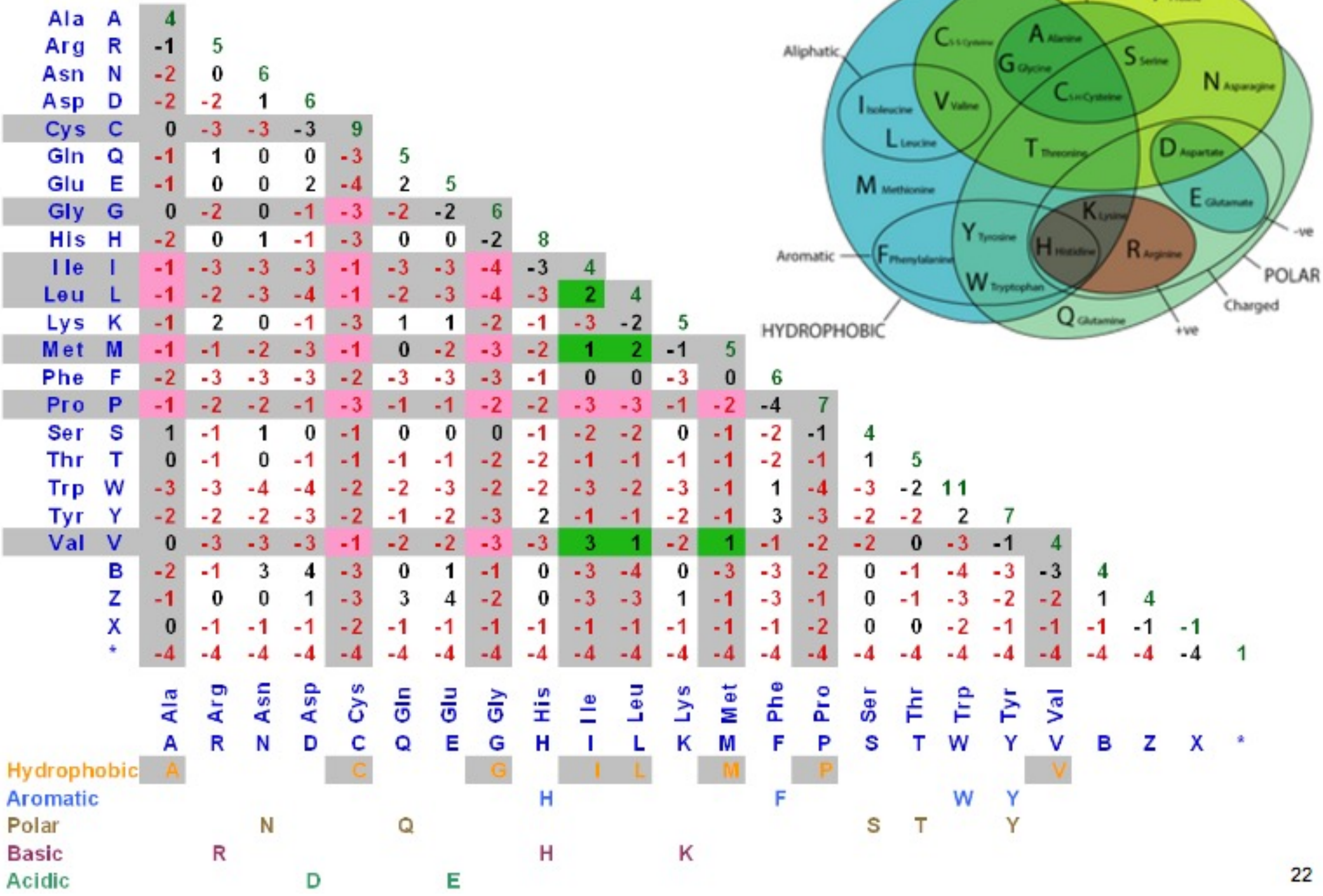
- For a given alignment, **the score is the sum** of the scores of each pair of residues

```
        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
```

Substitution matrix

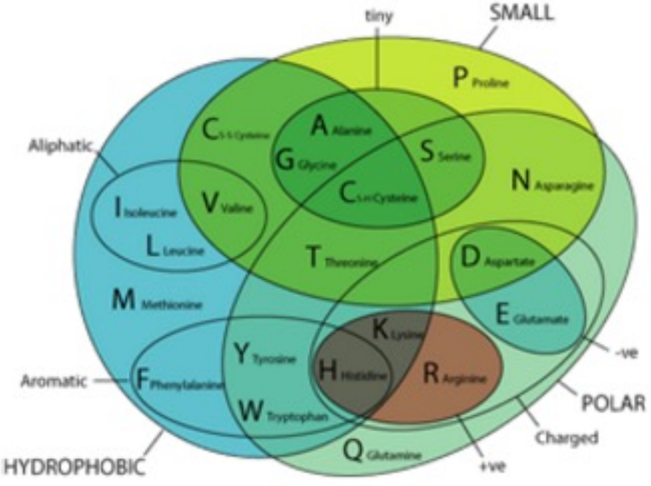
- For protein sequences, **BLOSUM or PAM** matrices are used. They provide different scores depending on the substitutions
- **Positive scores** indicate frequent ("accepted") substitutions, i.e. substitutions observed more frequently than would be expected by chance
- **Negative values** indicate rare mutations, which are observed less frequently than at random. This is an indication of negative selection, suggesting that these mutations are unfavourable to the function of the protein

BLOSUM62 matrix



	Ala A	Arg R	Asn N	Asp D	Cys C	Gln Q	Glu E	Gly G	His H	Ile I	Leu L	Lys K	Met M	Phe F	Pro P	Ser S	Thr T	Trp W	Tyr Y	Val V	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
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



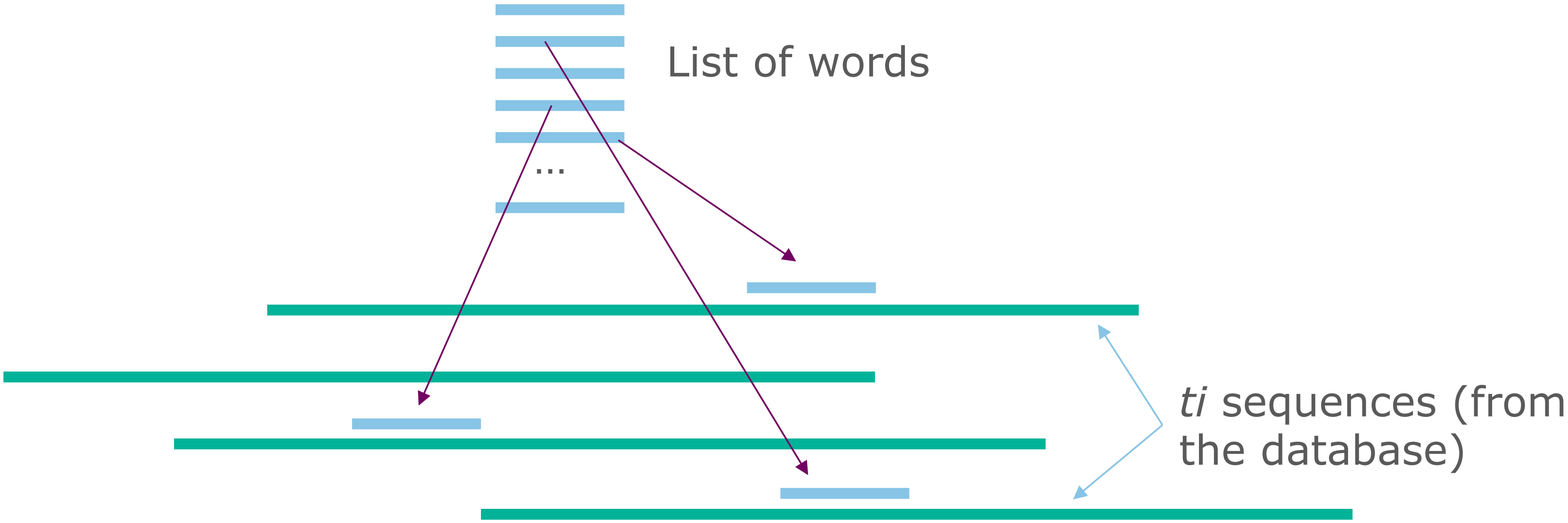
From J. van Helden, Université d’Aix-Marseille

Algorithm

- **Step 2**: Search for exact matches between the words in the list (DNA) or the extended list (proteins) and the *ti* sequences in the database

- These alignments are *hits*

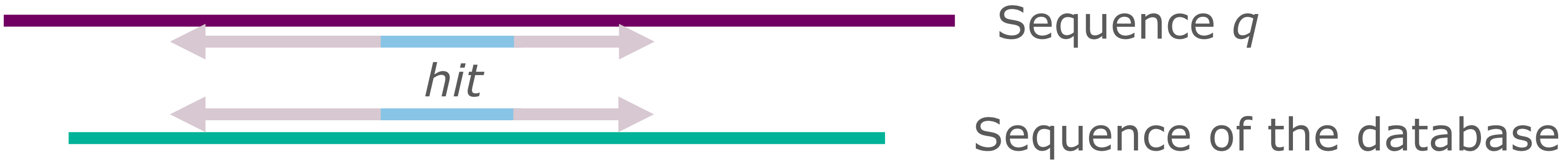
- A *hit* is therefore a "common" word of size *w* (and of score greater than *T* in the case of proteins) between the sequences *q* and *ti*



Algorithm

- **3rd step:** each hit is extended to the left and to the right: the extension is stopped when the *hit* score decreases by more than *X* (*X-drop*)

- Schematically



- Each extended hit forms an **LMSP**: Localy Maximal scoring Segment Pair

- Blast conserved only LMPSPs with a score higher than a given threshold score: the **HSPs**: High scoring Segment Pairs

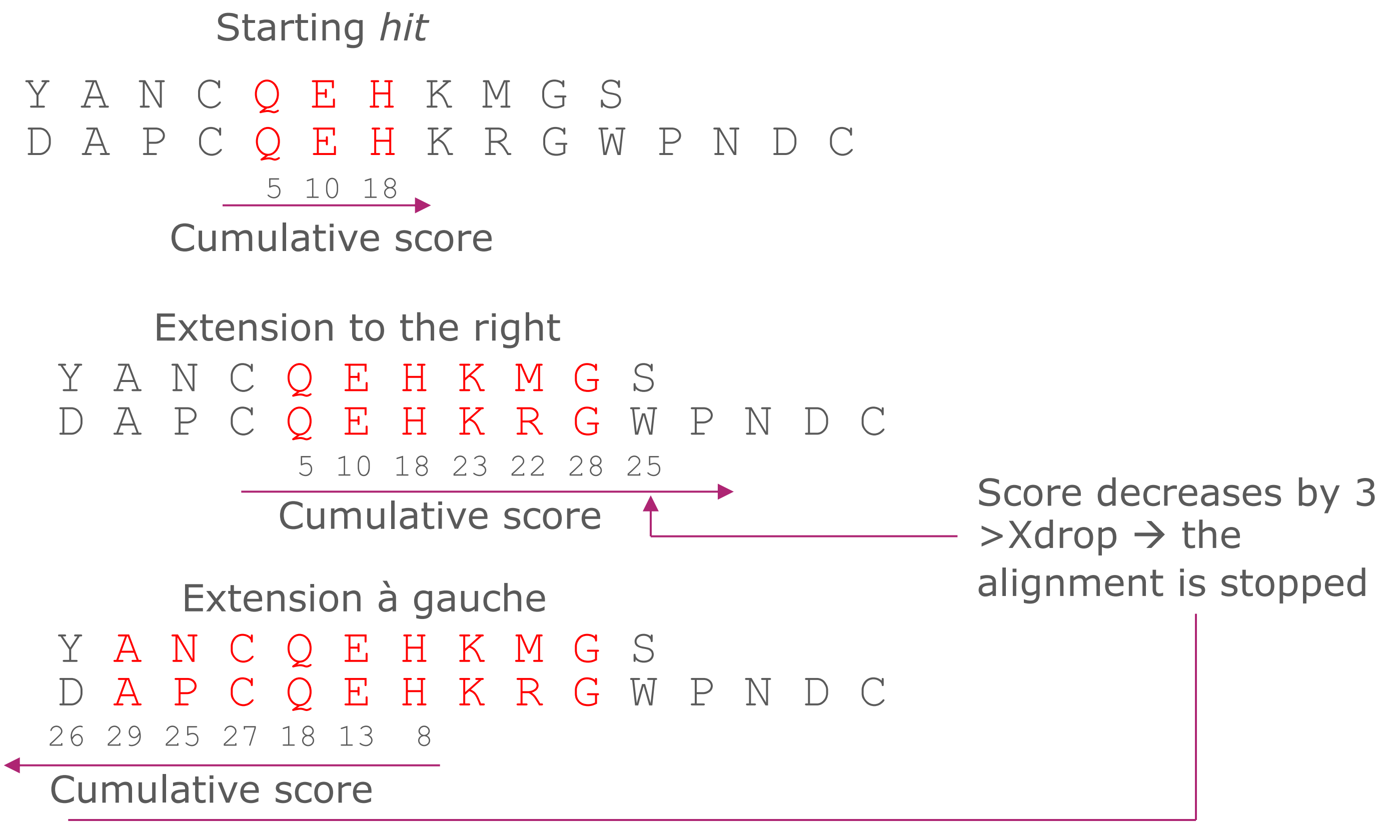
- The most significant HSP is called **MSP**: Maximum scoring Segment Pair

Algorithm

- Clarification on 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 calculated
from BLOSUM62



Starting *hit*

```
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
```

Cumulative score

Extension to the right

```
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
```

Cumulative score

Score decreases by 3 >Xdrop → the alignment is stopped

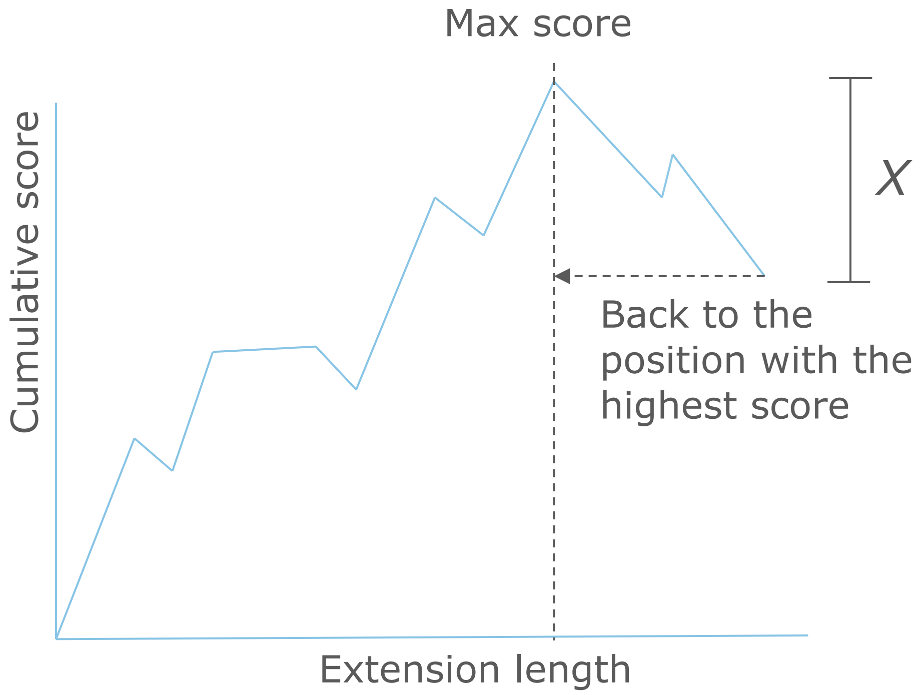
Extension à gauche

```
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
```

Cumulative score

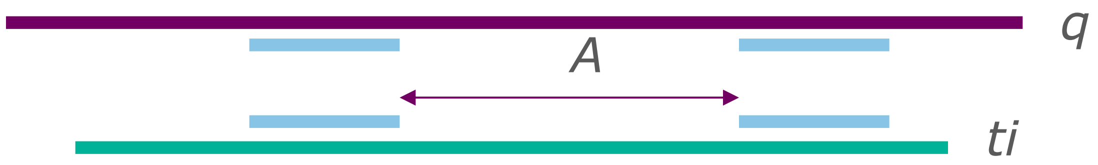
Algorithm

- Clarification on X-drop



Gapped-Blast (BLAST2)

- Based on 2 hits with a maximum distance of *A* (BLASTP). To keep a good sensitivity, *T* is lowered from 13 to 11



- Extend the hits by allowing gaps

- This method is **faster** than the previous one

Significance of alignments

Significance of alignments

- **Two sequences** can **always be aligned**

- There is always one (at least) **best *S*-score** alignment between two sequences (an MSP)

- **Issues**

- Is this score high enough to prove homology?

- Can we find a MSP with a better score in two random sequences?

Significance of alignments

- *S* is the score obtained by the alignment of 2 sequences

- The **p-value** measures the **probability** that 2 random sequences of the same length and composition have an MSP of score ≥ S

- The **E-value** measures the esperance *E* of the number *n* of MSPs of score ≥ *S* in 2 random sequences of the same length and composition

→*For example, if the E-value is equal to 10 for a HSP with score S, it means that 10 HSPs with score ≥ S can be found by chance! So probably your alignment is not significant!*

Calculation of the E-value

- According to Karlin and Altschul, 1991

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

With *m* the size of the sequence *q*, *n* the size of the database, *S* the score of the HSP, *K* and *lambda* depend on the score matrix, *K* can be adjusted according to the cost of the gaps

- If S is the score for a hit
- The bit-score (normalized score) is: $S' = \dfrac{\lambda s - lnK}{ln2}$
- The E-value is then: $E = mn2^{-S'}$

Variation in E-value

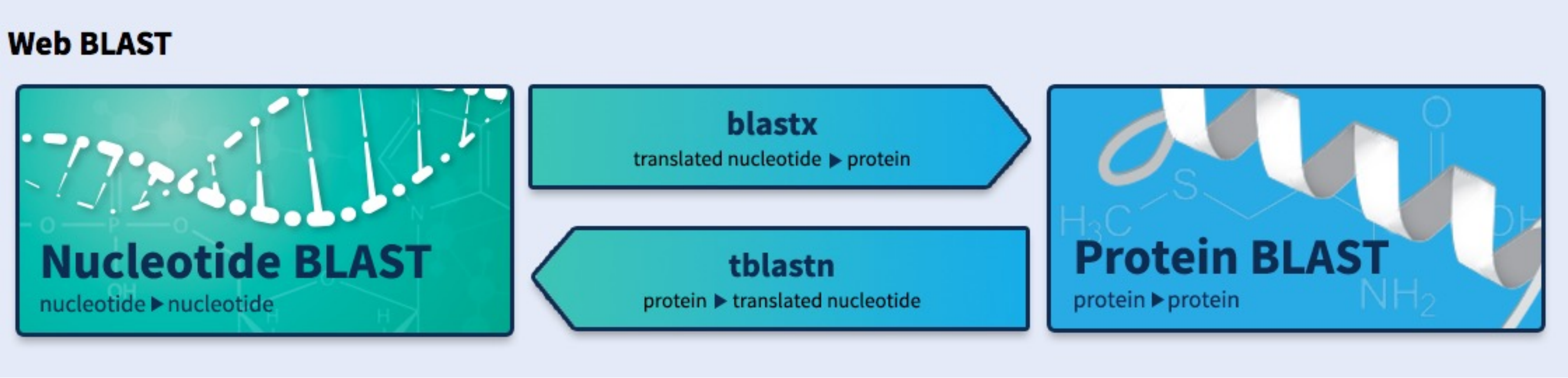
- if the size of the query sequence increases: the E-value ...
- If the size of the database is divided by two: the E-value ...
- If the score increases: the E-value ...
- What bit-score to obtain an E-value of 0.05 for a sequence of length 250 and a bd of length 50000000 ?
- If we increase the E-value to 0.01, what will be the bit-score?

Variation in E-value

- if the size of the query sequence increases: the E-value increases
- If the size of the database is divided by two: the E-value decreases
- If the score increases: the E-value decreases
- What bit-score to obtain an E-value of 0.05 for a sequence of length 250 and a bd of length 50000000 ? 38 bits
- If we increase the E-value to 0.01, what will be the bit-score? 40 bits

Run Blast!

Home page



Web BLAST

Nucleotide BLAST
nucleotide ▶ nucleotide

blastx
translated nucleotide ▶ protein

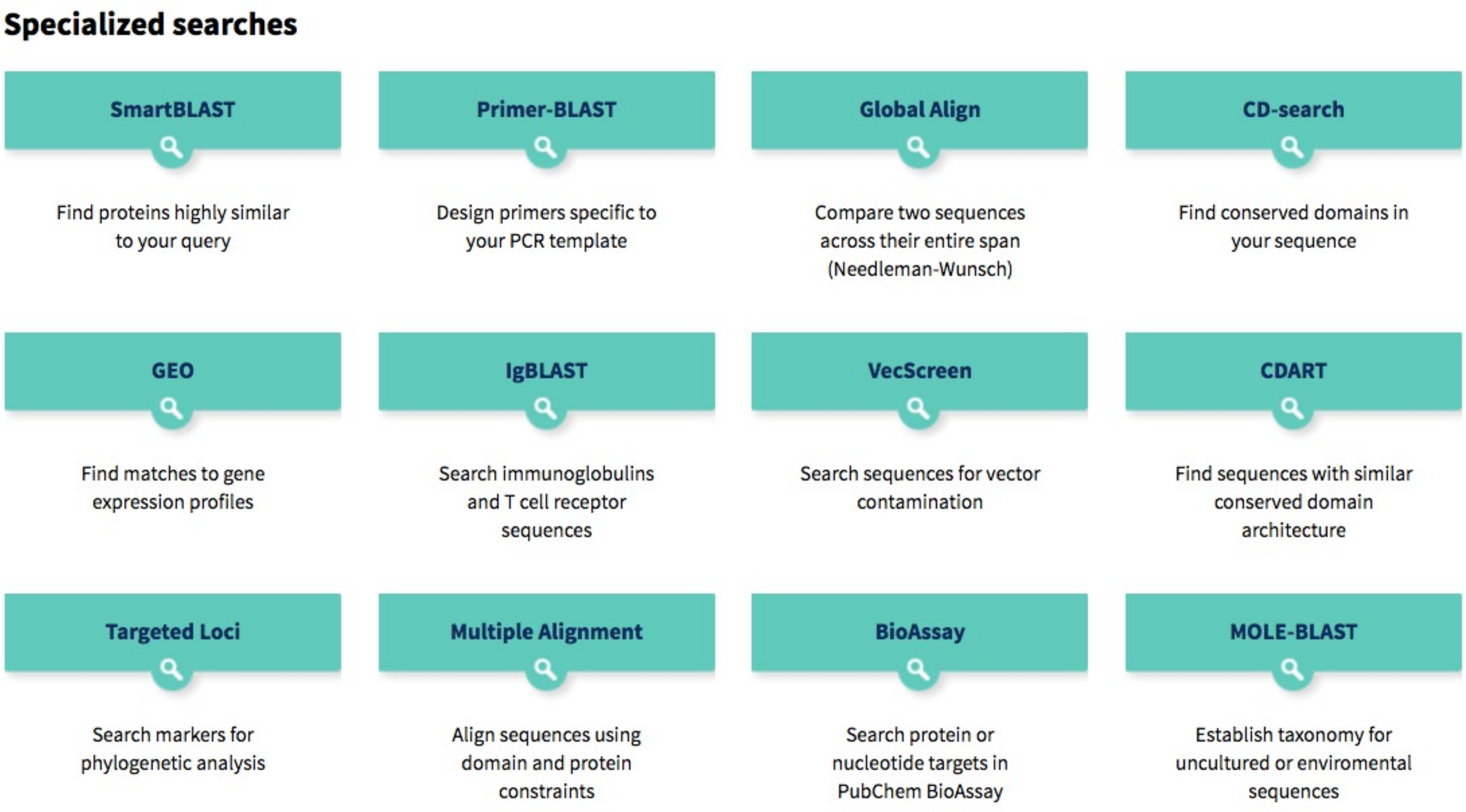
tblastn
protein ▶ translated nucleotide

Protein BLAST
protein ▶ protein

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

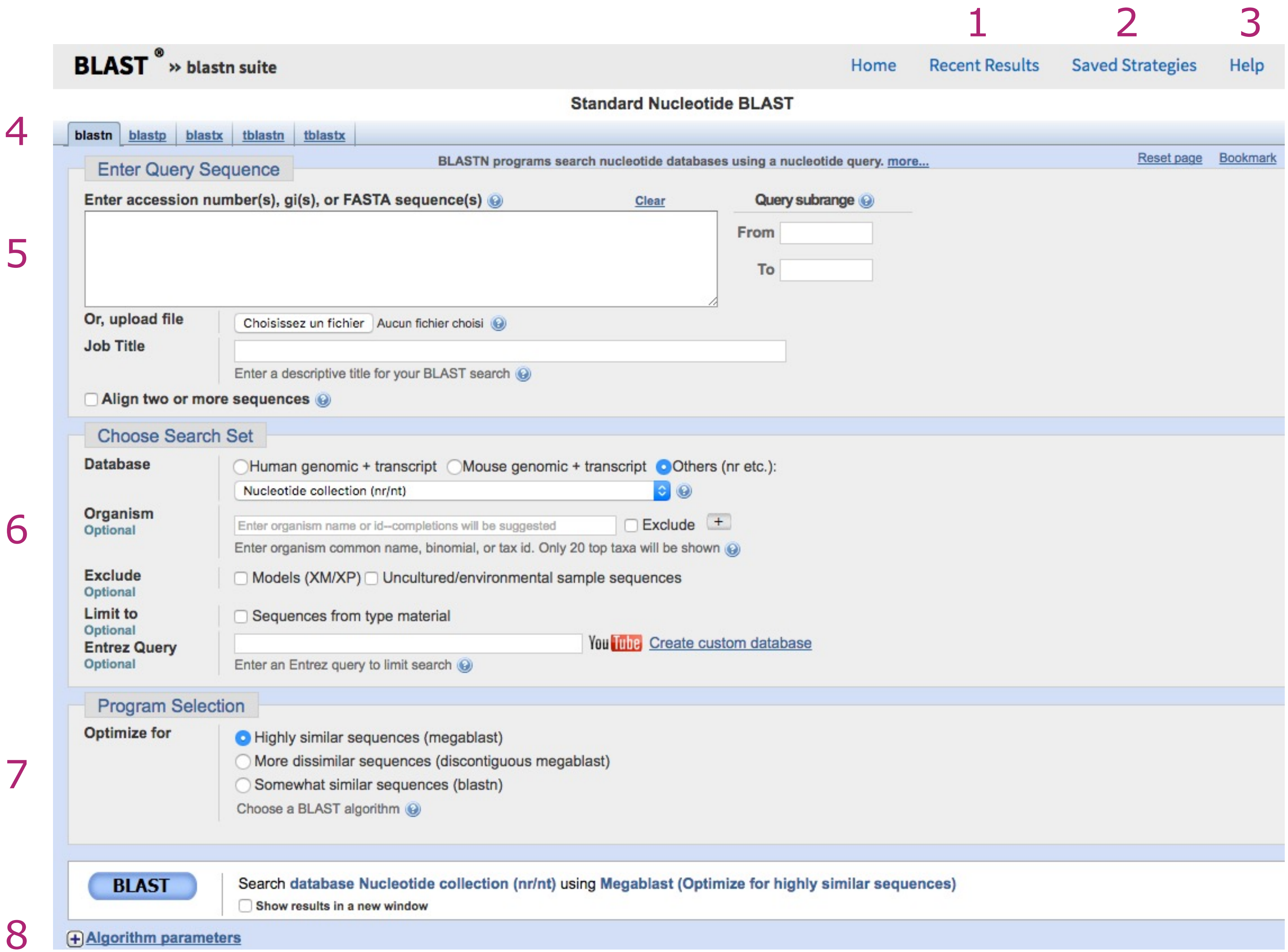
BioAssay

Search protein or nucleotide targets in PubChem BioAssay

MOLE-BLAST

Establish taxonomy for uncultured or enviromental sequences

Graphical user interface



BLAST® » blastn suite

Home Recent Results Saved Strategies Help

Standard Nucleotide BLAST

blastn blastp blastx tblastn tblastx

BLASTN programs search nucleotide databases using a nucleotide query. more...

Reset page Bookmark

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

Query subrange
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)

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
YouTube Create custom database
Enter an Entrez query to limit search

Program Selection

Optimize for
● 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

+ Algorithm parameters

Graphical user interface

- **1 Recent Results:** The results of your searches over the last 36 hours. If you are registered on MyNCBI, you can access your results from any machine. If not, only searches from the active browser session are kept

- **2 Saved Strategies :** allows you to save the parameters of a Blast search in order to restart a search with the same parameters later (connection to MyNCBI required)

- **3 Help :** documentations, links and tutorials

- **4 :** type of Blast

Graphical user interface

- **5 Enter Query Sequence**

Copy/paste or upload your query sequence(s). You can also define a search range in your sequences. You can give a title to your search.

The "Align two or more sequences" function allows you to compare sequences between them without using a database

- **6 Choose Search Set**

Select your database. You can limit your search to specific organisms or exclude organisms. You can exclude sequences produced from genome annotation projects or from non-cultured/bred organisms. You can limit your search to model specimens and strains

- **7 Program Selection**

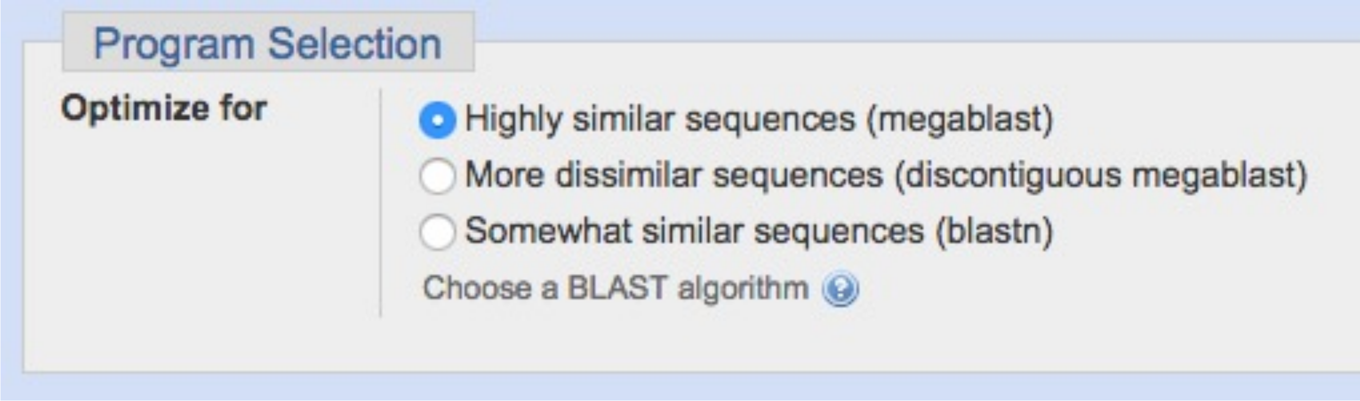
Allows you to optimise your search for different scenarios (e.g. intra or inter species searches)

Graphical user interface

- **8 Algorithm parameters**

This is the place to modify the parameters of the BLAST agoritm that has been selected (see dedicated section)

Nucleotide Blast



Program Selection

Optimize for

● Highly similar sequences (megablast)

○ More dissimilar sequences (discontiguous megablast)

○ Somewhat similar sequences (blastn)

Choose a BLAST algorithm

- **Megablast**:

- a Faster Blast when searching for high similarity

- Implementation: use larger word sizes (28 *vs* 11)

- To be reserved when searching for very similar sequences or when we want to know if our sequence is in the database

Nucleotide Blast

- **Discontigous megablast**:

- Use a spaced seed rather than an exact word (contiguous seed)

- Useful for inter-species comparisons

- Example of contiguous seed: 1 1 1 1 1: an exact word (without mismatch) of 5 nucleotides

- Example of spaced seed: 1 0 1 1 0 1 1: a word of 7 nucleotides, positions 2 and 5 may represent mismatches

Spaced seeds *vs* contiguous seeds

- We consider a sequence *q* of length *l*=26

- A seed (word) of size 6

- We can therefore define a maximum of 26-6+1=21 seeds

- The sequence *ti* is identical to *q*: therefore all seeds can be aligned with *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
```

Spaced seeds *vs* contiguous seeds

```
ATCTGATCGATCGATCGATCGATCGA
||||||||||||||||||||||||||
ATCTGATCGATCGATCGATCGATCGA
111111
 111111
  111111
   111111
    111111
     111111
      111111
       111111
        111111
         111111
          111111
           111111
            111111
             111111
              111111
               111111
                111111
                 111111
                  111111
                   111111
                    111111
```

```
ATCTGATCGATCGATCGATCGATCGA
||||||||||||||||||||||||||
ATCTGATCGATCGATCGATCGATCGA
11101011
 11101011
  11101011
   11101011
    11101011
     11101011
      11101011
       11101011
        11101011
         11101011
          11101011
           11101011
            11101011
             11101011
              11101011
               11101011
                11101011
                 11101011
                  11101011
```

- Spaced seeds behave in the same way as contiguous seeds in this example

Spaced seeds *vs* contiguous seeds

- Now let's introduce a mismatch between *q* and *ti*...

```
ATCTGATCGATCGATCGATCGATCGA
|||||.||||||||||||||||||||
ATCTGCTCGATCGATCGATCGATCGA
111111
 111111
  111111
   111111
    111111
     111111
      111111
       111111
        111111
         111111
          111111
           111111
            111111
             111111
              111111
               111111
                111111
                 111111
                  111111
                   111111
                    111111
```

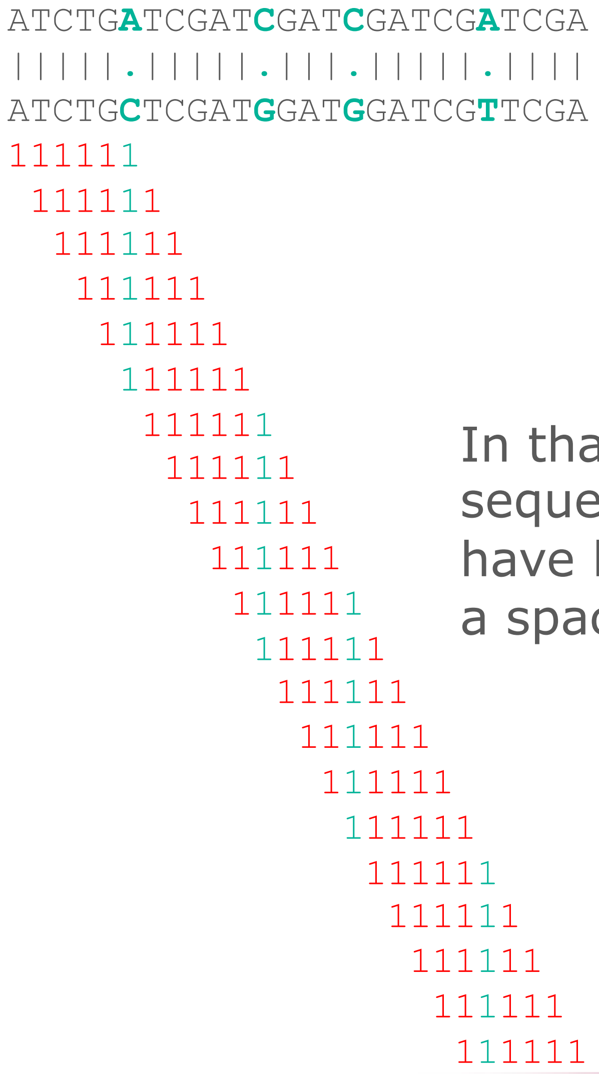
```
ATCTGATCGATCGATCGATCGATCGA
|||||.||||||||||||||||||||
ATCTGCTCGATGGATGGATCGTTCGA
11101011
 11101011
  11101011
   11101011
    11101011
     11101011
      11101011
       11101011
        11101011
         11101011
          11101011
           11101011
            11101011
             11101011
              11101011
               11101011
                11101011
                 11101011
                  11101011
```

Some seeds (in red) are lost

Spaced seeds can be permissive

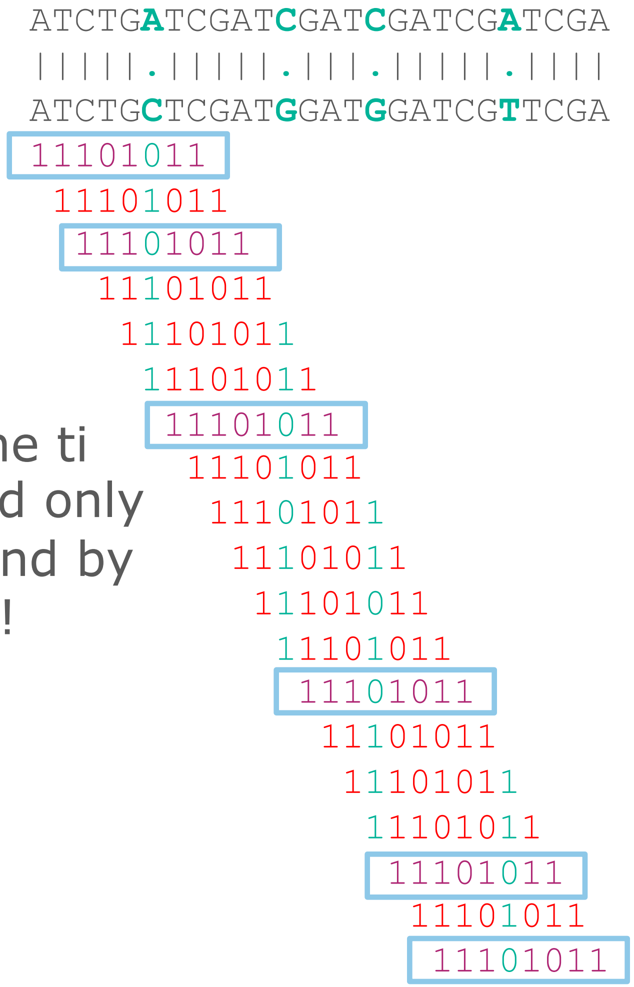
Spaced seeds *vs* contiguous seeds

- Then, Let's introduce more mismatches between q and ti . …



```
ATCTGATCGATCGATCGATCGATCGA
|||||.|||||.|||.|||||.||||
ATCTGCTCGATGGATGGATCGTTCGA
111111
 111111
  111111
   111111
    111111
     111111
      111111
       111111
        111111
         111111
          111111
           111111
            111111
             111111
              111111
               111111
                111111
                 111111
                  111111
                   111111
                    111111
```

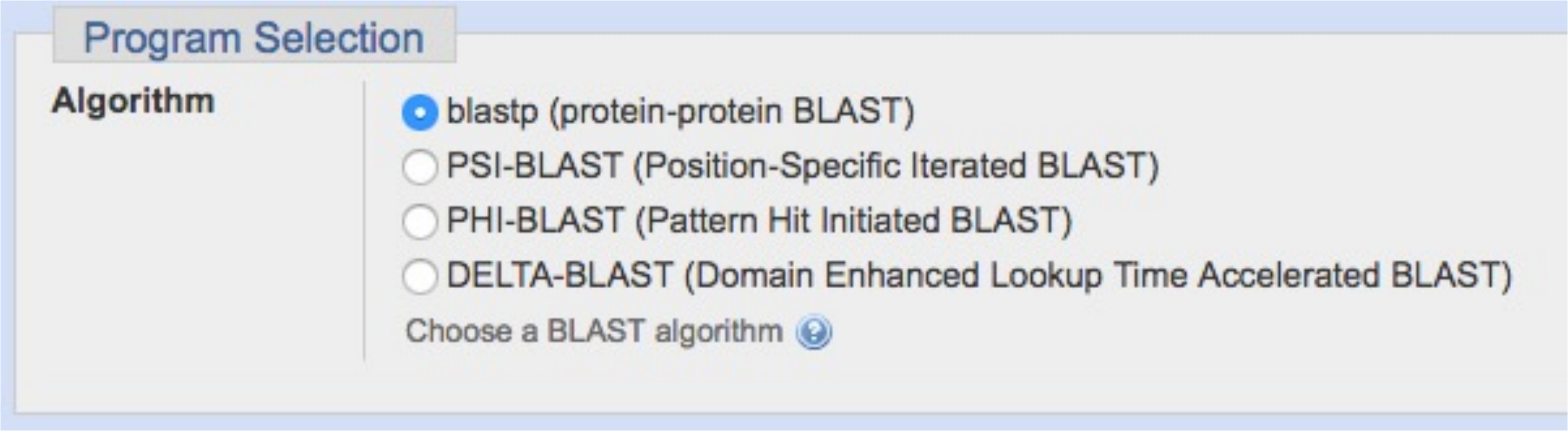
In that case the ti sequence could only have been found by a spaced seed!



```
ATCTGATCGATCGATCGATCGATCGA
|||||.|||||.|||.|||||.||||
ATCTGCTCGATGGATGGATCGTTCGA
11101011
 11101011
  11101011
   11101011
    11101011
     11101011
      11101011
       11101011
        11101011
         11101011
          11101011
           11101011
            11101011
             11101011
              11101011
               11101011
                11101011
                 11101011
                  11101011
```

Université de Lille

Protein Blast



Program Selection

Algorithm

● blastp (protein-protein BLAST)

○ PSI-BLAST (Position-Specific Iterated BLAST)

○ PHI-BLAST (Pattern Hit Initiated BLAST)

○ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm

- **PSI-BLAST :**

- Initial search with blastp

- Construction of a multiple alignment and then a profile from the best hits → position-specific score matrix (PSSM)

- New search with the profile

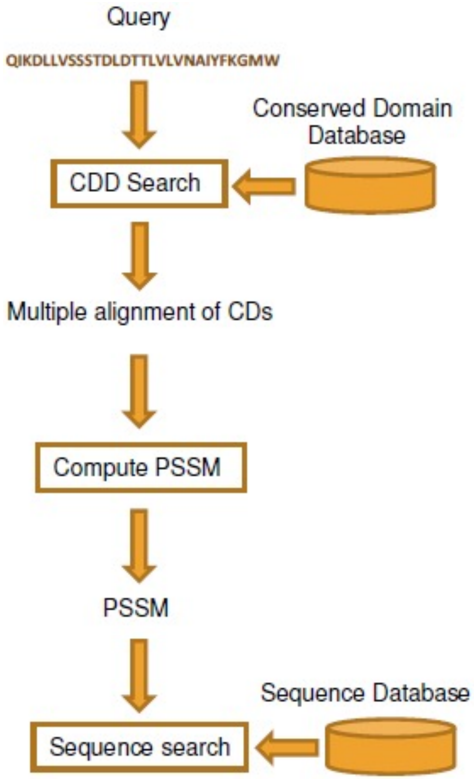
Protein Blast

- **PHI-BLAST :**

- Input: a protein sequence and a motif (regular expression)
- Restriction of the library to sequences for which the motif is found

- **DELTA-BLAST**

- Use of PSSMs built from a NCBI CDD (conserved domain database)
- Faster than PSI-BLAST, also more sensitive

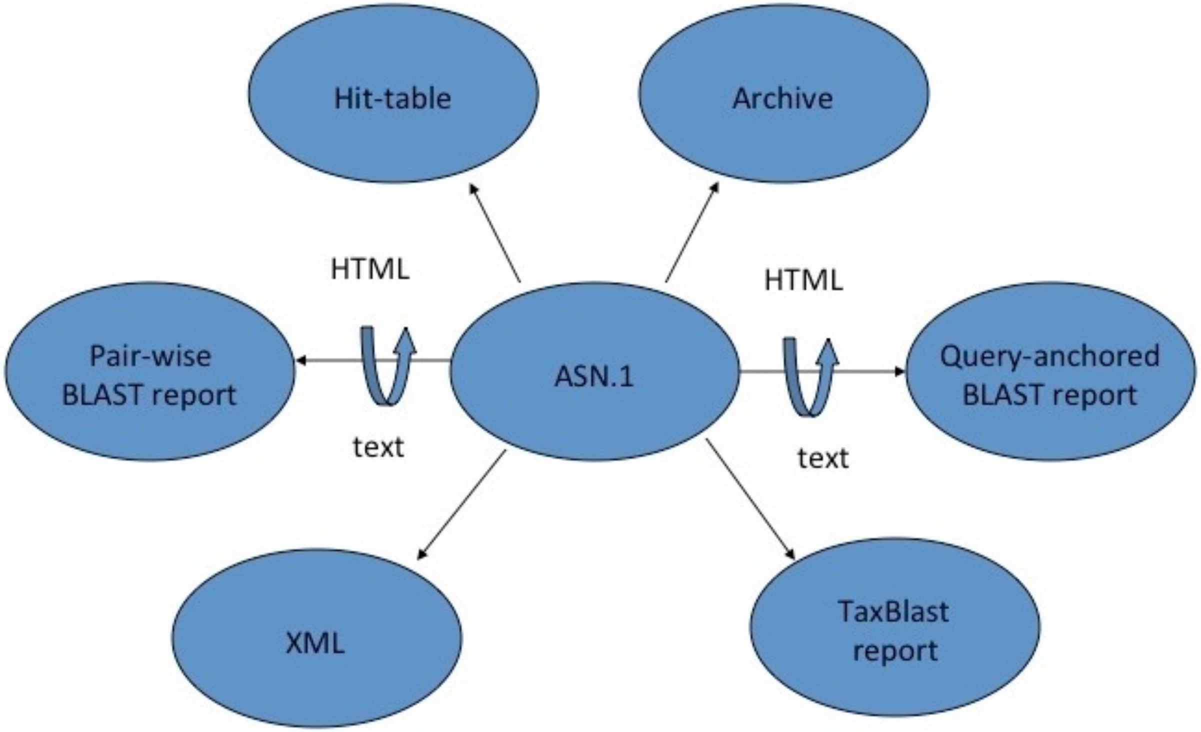

Query
QIKDLLVSSSTDLDTTLVLVNAIYFKGMW
Conserved Domain Database
CDD Search
Multiple alignment of CDs
Compute PSSM
PSSM
Sequence Database
Sequence search

Boratyn et al 2012

Blast results

Results

- Structured results: a flexible output

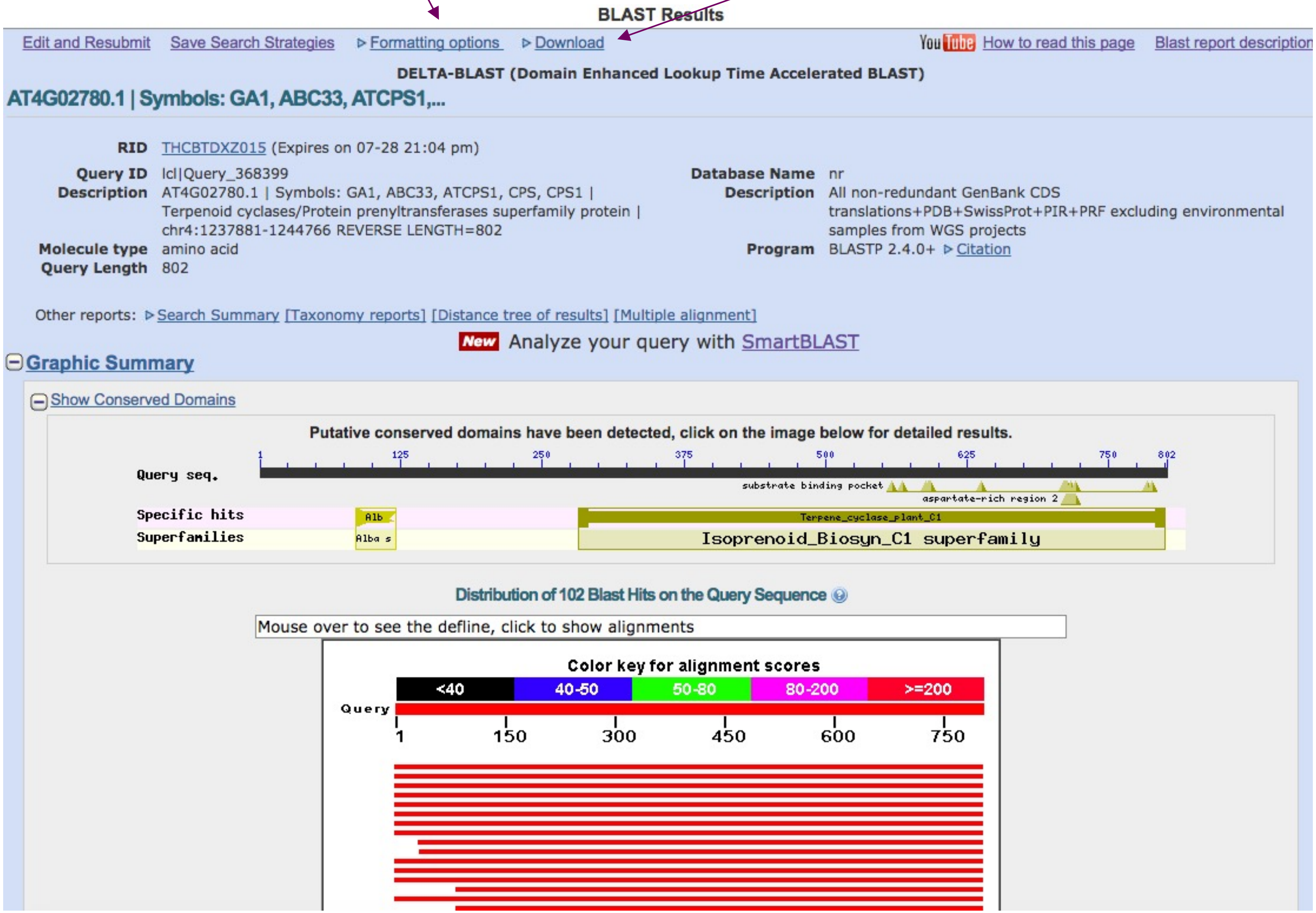


Madden, 2013

Results

Results formatting

Downloads



BLAST Results

Edit and Resubmit   Save Search Strategies   ▷ Formatting options   ▷ Download   YouTube How to read this page   Blast report description

DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

AT4G02780.1 | Symbols: GA1, ABC33, ATCPS1,...

RID THCBTDXZ015 (Expires on 07-28 21:04 pm)

Query ID lcl|Query_368399
Description AT4G02780.1 | Symbols: GA1, ABC33, ATCPS1, CPS, CPS1 | Terpenoid cyclases/Protein prenyltransferases superfamily protein | chr4:1237881-1244766 REVERSE LENGTH=802
Molecule type amino acid
Query Length 802

Database Name nr
Description All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
Program BLASTP 2.4.0+ ▷ Citation

Other reports: ▷ Search Summary [Taxonomy reports] [Distance tree of results] [Multiple alignment]

New Analyze your query with SmartBLAST

⊖ Graphic Summary

⊖ Show Conserved Domains

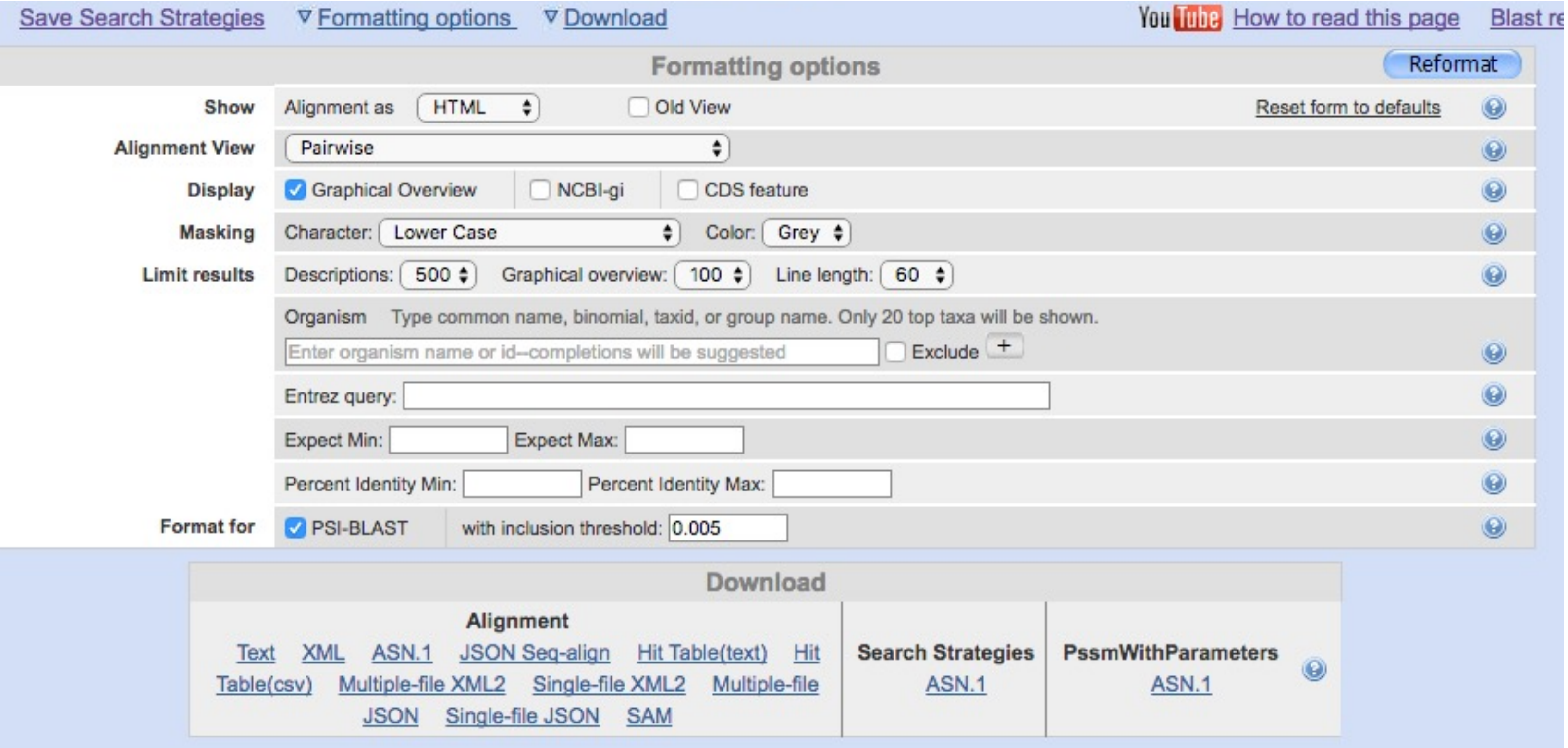
Putative conserved domains have been detected, click on the image below for detailed results.

Distribution of 102 Blast Hits on the Query Sequence

Mouse over to see the defline, click to show alignments

Results

- Different format and export possibilities...



Save Search Strategies ▿ Formatting options ▿ Download YouTube How to read this page Blast re

Formatting options Reformat

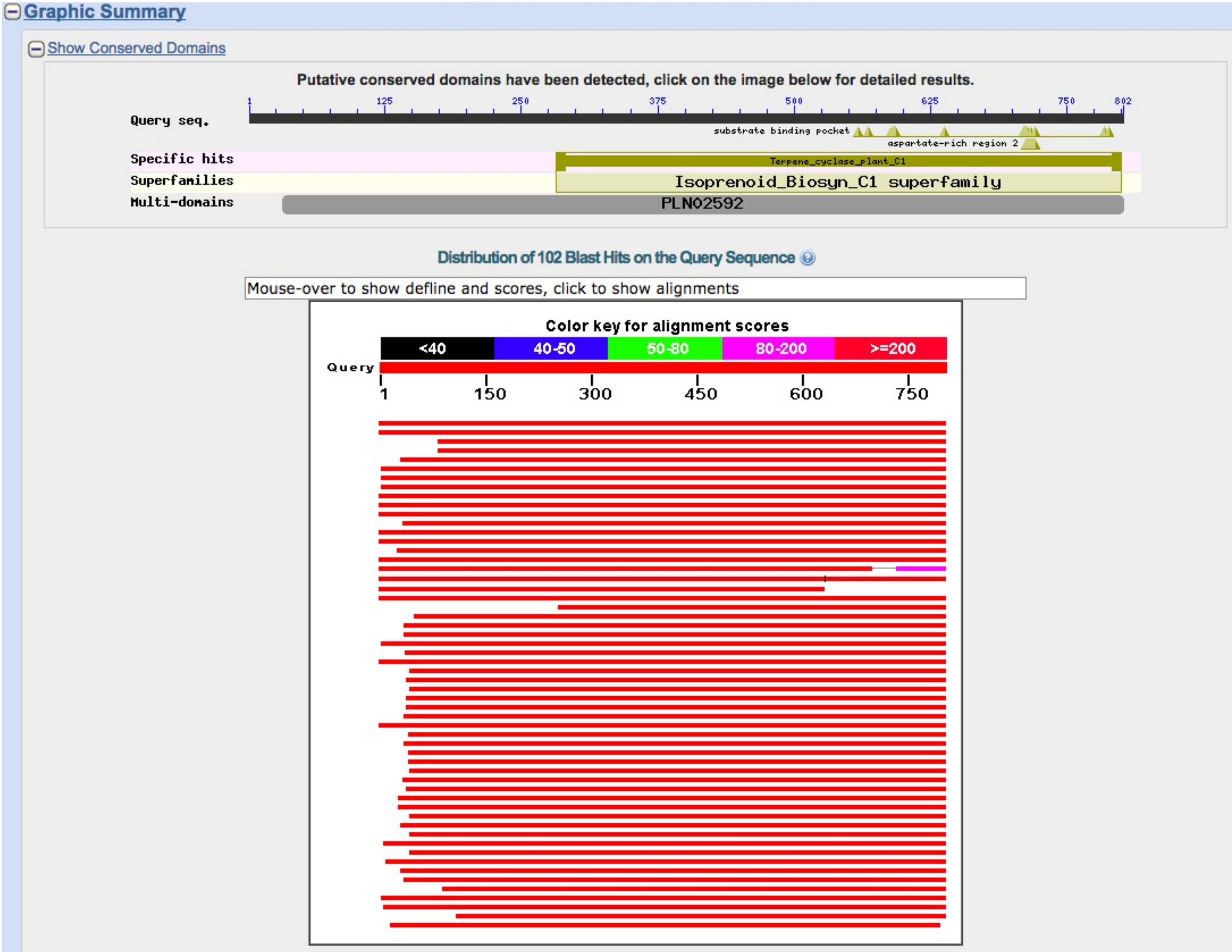
Show	Alignment as HTML □ Old View Reset form to defaults
Alignment View	Pairwise
Display	☑ Graphical Overview □ NCBI-gi □ CDS feature
Masking	Character: Lower Case Color: Grey
Limit results	Descriptions: 500 Graphical overview: 100 Line length: 60
	Organism Type common name, binomial, taxid, or group name. Only 20 top taxa will be shown. Enter organism name or id--completions will be suggested □ Exclude +
	Entrez query:
	Expect Min: Expect Max:
	Percent Identity Min: Percent Identity Max:
Format for	☑ PSI-BLAST with inclusion threshold: 0.005

Download

Alignment	Search Strategies	PssmWithParameters
Text XML ASN.1 JSON Seq-align Hit Table(text) Hit Table(csv) Multiple-file XML2 Single-file XML2 Multiple-file JSON Single-file JSON SAM	ASN.1	ASN.1

Results

- Graphic summary



Graphic Summary

Show Conserved Domains

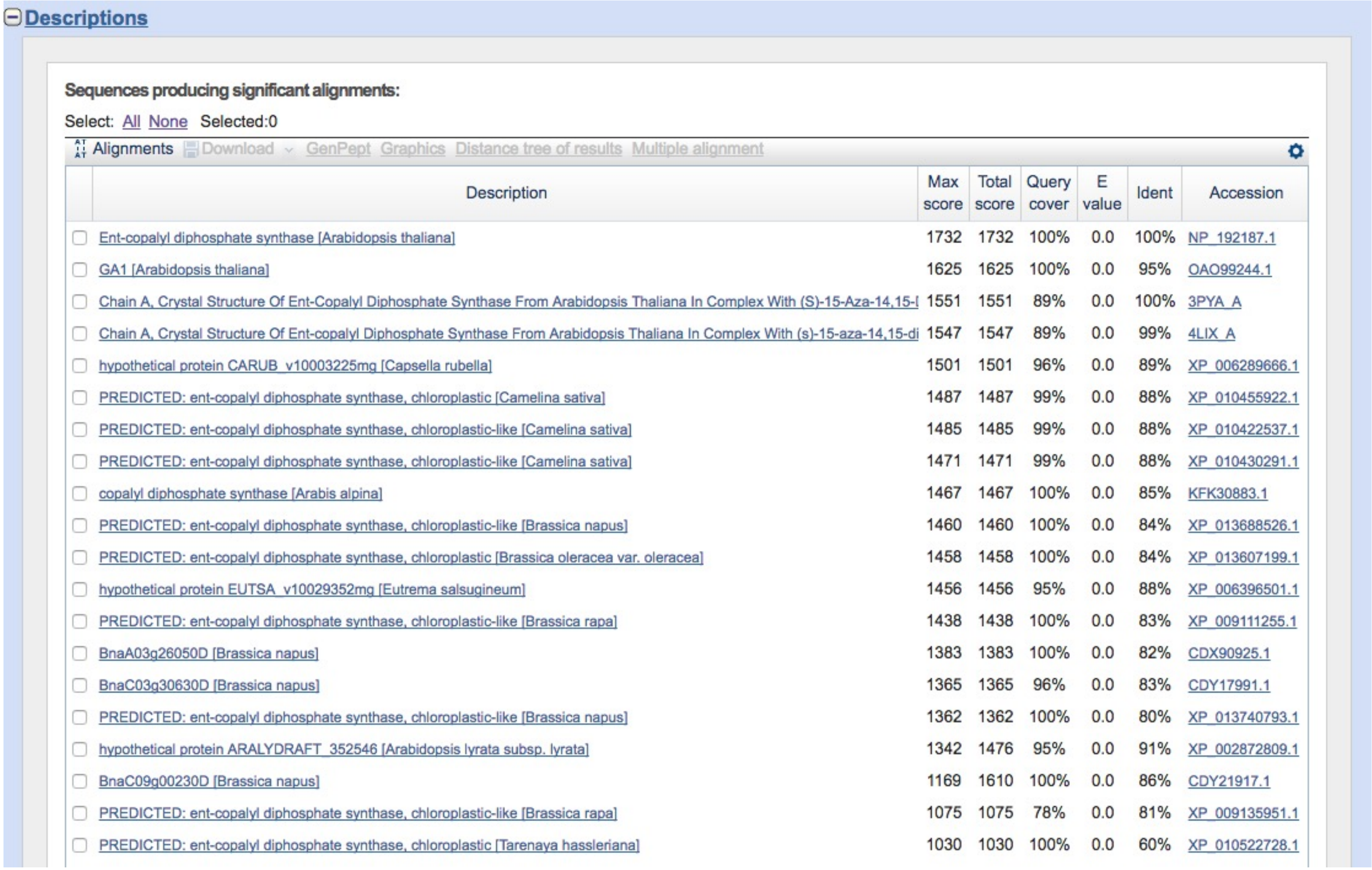
Putative conserved domains have been detected, click on the image below for detailed results.

Distribution of 102 Blast Hits on the Query Sequence

Mouse-over to show defline and scores, click to show alignments

Results

- Descriptions



⊖ Descriptions

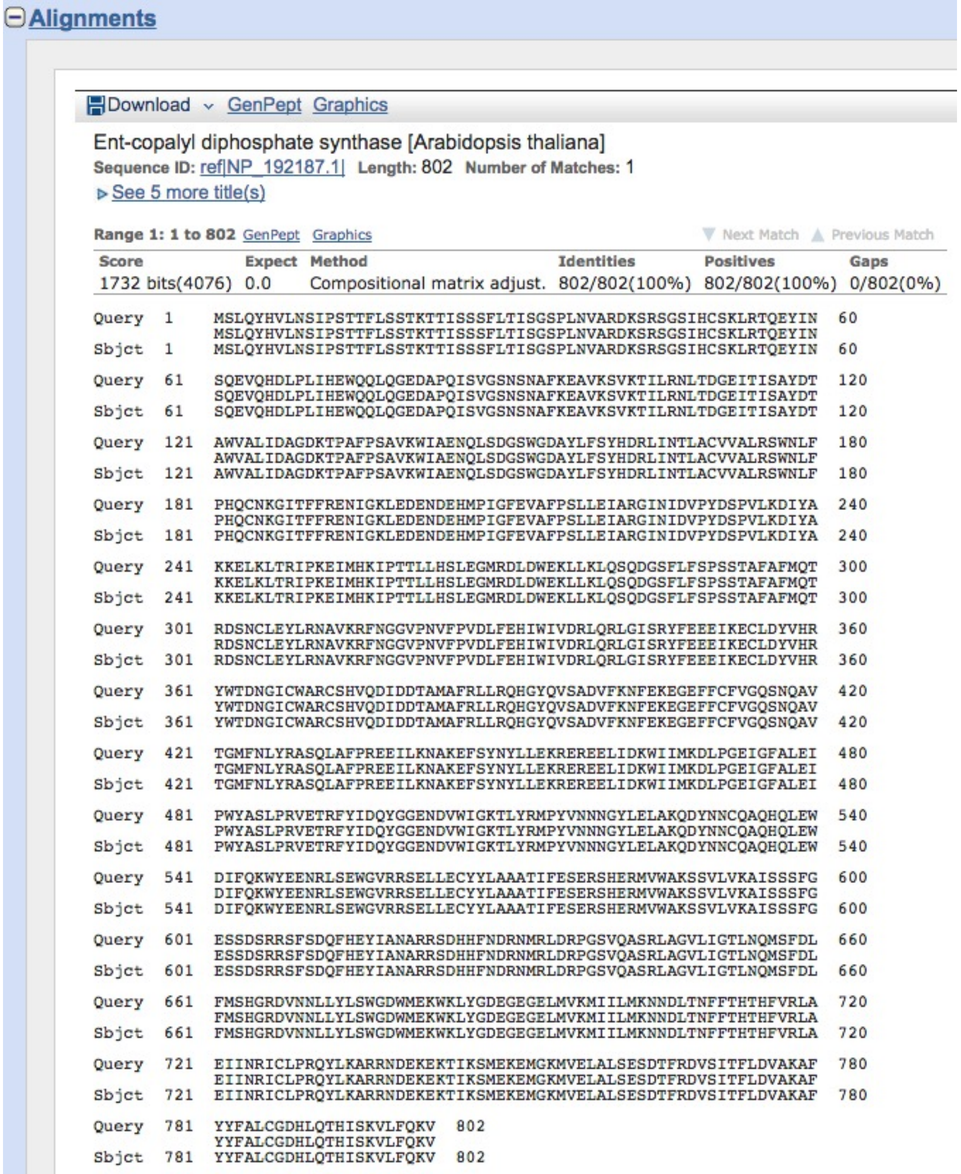
Sequences producing significant alignments:

Select: All None Selected:0

	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-I	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

Results

- Alignments



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

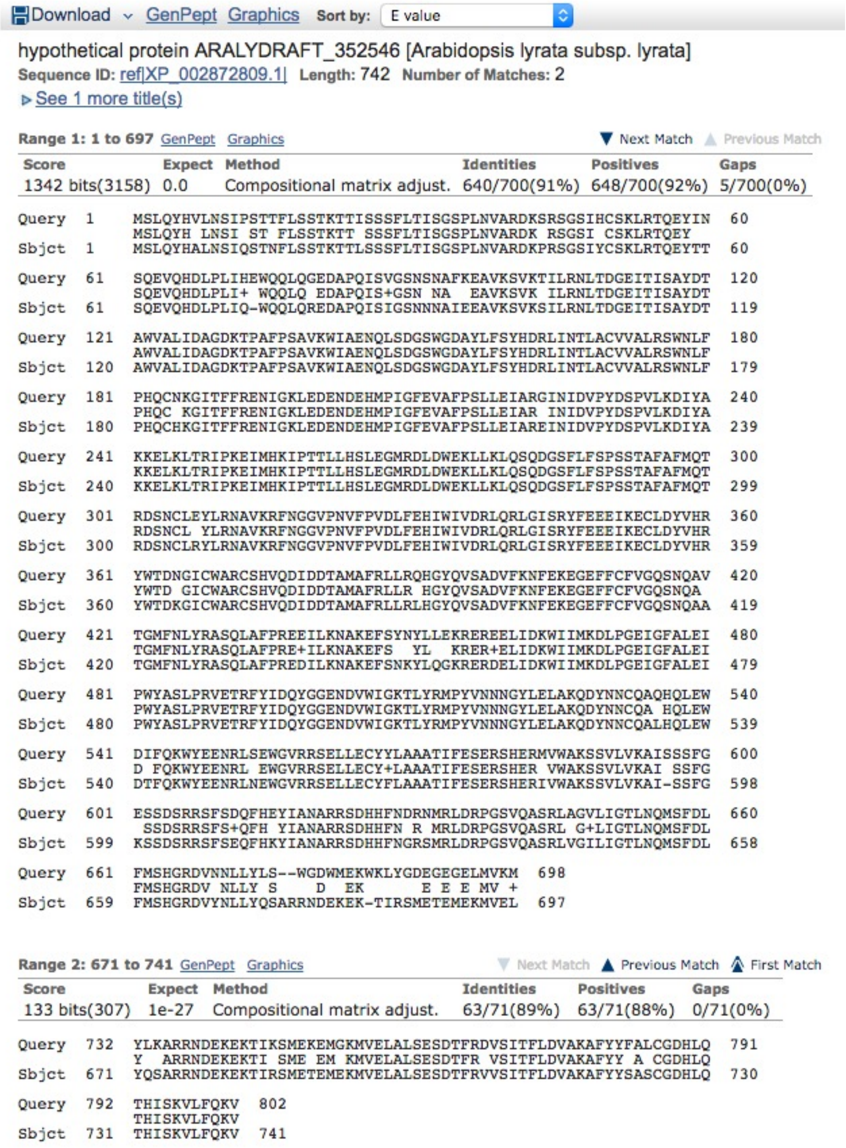
Score	Expect	Method	Identities	Positives	Gaps
1732 bits(4076)	0.0	Compositional matrix adjust.	802/802(100%)	802/802(100%)	0/802(0%)

```
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
```

Results

- Alignments

HSPs



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
```

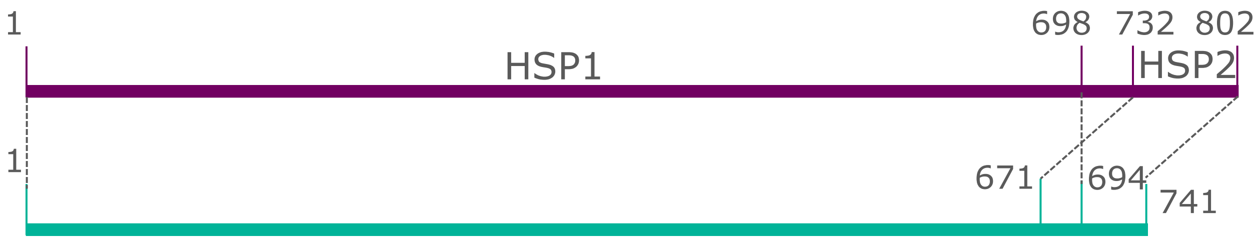
Range 2: 671 to 741 GenPept Graphics ▼ Next Match ▲ Previous Match ▲ First Match

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

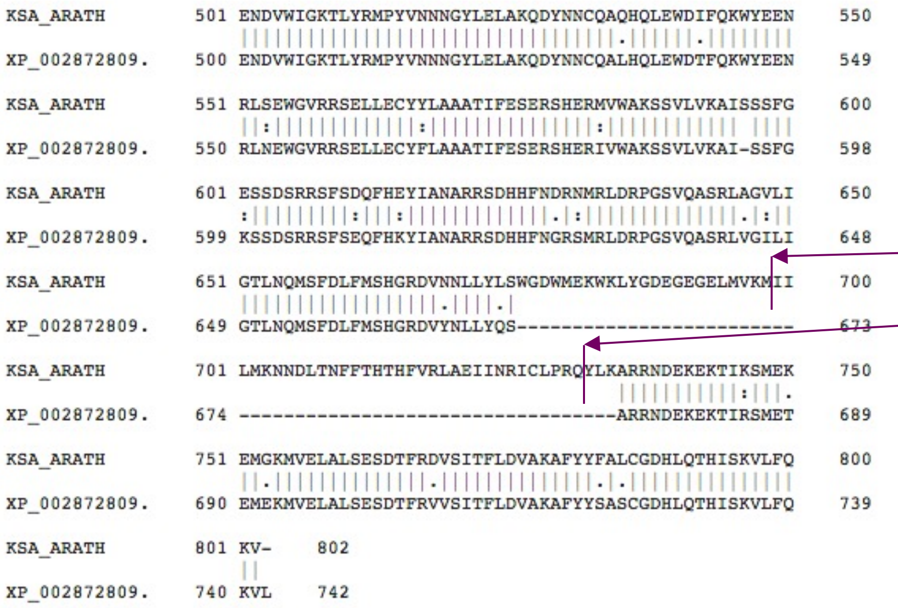
```
Query  732  YLKARRNDEKEKTIKSMEKEMGKMVELALSESDTFRDVSITFLDVAKAFYYFALCGDHLQ  791
            Y  ARRNDEKEKTI SME EM KMVELALSESDTFR VSITFLDVAKAFYY A CGDHLQ
Sbjct  671  YQSARRNDEKEKTIRSMETEMEKMVELALSESDTFRVVSITFLDVAKAFYYSASCGDHLQ  730

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

Blast vs global alignment

- Graphical overview of the Blast alignment



- Global alignment obtained using Needle



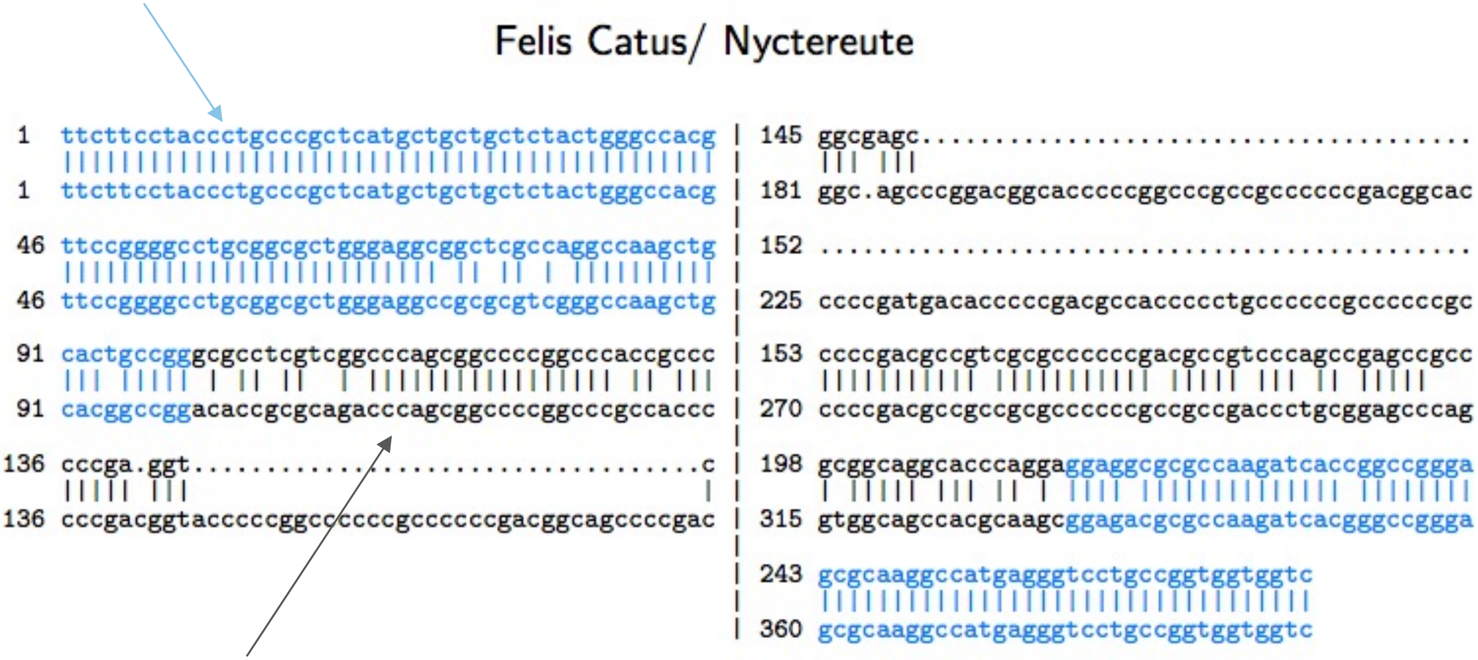
```
KSA_ARATH        501 ENDVWIGKTLYRMPYVNNNGYLELAKQDYNNCQAQHQLEWDIFQKWYEEN    550
                     |||||||||||||||||||||||||||||||||.|||||||.||||||||
XP_002872809.    500 ENDVWIGKTLYRMPYVNNNGYLELAKQDYNNCQALHQLEWDTFQKWYEEN    549

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
```

End HSP1 with Blast

Start HSP2 with Blast

Blast vs global alignment

Blast alignment



Felis Catus/ Nyctereute

```
  1 ttcttcctaccctgcccgctcatgctgctgctctactgggccacg | 145 ggcgagc......................................
    ||||||||||||||||||||||||||||||||||||||||||||| |     ||| |||
  1 ttcttcctaccctgcccgctcatgctgctgctctactgggccacg | 181 ggc.agcccggacggcacccccggcccgccgccccccgacggcac
                                                  |
 46 ttccggggcctgcggcgctgggaggcggctcgccaggccaagctg | 152 .............................................
    |||||||||||||||||||||||||| || || | |||||||||| |
 46 ttccggggcctgcggcgctgggaggccgcgcgtcgggccaagctg | 225 ccccgatgacacccccgacgccacccccctgcccccgccccccgc
                                                  |
 91 cactgccgggcgcctcgtcggcccagcggccccggcccaccgccc | 153 ccccgacgccgtcgcgccccccgacgccgtcccagccgagccgcc
    ||| ||||| | || ||  | ||||||||||||||||| || ||| |     ||||||||||| ||||||||||| ||||| ||| || |||||
 91 cacggccggacaccgcgcagacccagcggccccggcccgccaccc | 270 ccccgacgccgccgcgccccccgccgccgaccctgcggagcccag
                                                  |
136 cccga.ggt...................................c | 198 gcggcaggcacccaggaggaggcgcgccaagatcaccggccggga
    ||||| |||                                   | |     | ||||| ||| || | |||| |||||||||||||| ||||||||
136 cccgacggtacccccggccccccgccccccgacggcagccccgac | 315 gtggcagccacgcaagcggagacgcgccaagatcacgggccggga
                                                  |
                                                  | 243 gcgcaaggccatgagggtcctgccggtggtggtc
                                                  |     ||||||||||||||||||||||||||||||||||
                                                  | 360 gcgcaaggccatgagggtcctgccggtggtggtc
```

Global alignment 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