BLAST comparison and searching in banks

Useful links

- Entrez
- Prosite
- Interpro
- Blast au NCBI

Presentation of BLAST at NCBI

The BLAST welcome page will guide you following the experiments you want to pursue. You can :

- Search for your sequence in a specific genome (*BLAST RefSeq Assembled Genomes*) : offers you the possibility to interrogate data from full genomes, or genomes under ongoing sequencing. The interrogated data are not redundant : each genome region only appears once in the bank. When data are available, it is also possible to interrogate mRNAs, proteins ... Very useful to interrogate data from a single organism, as there is no redundancy.
- Search for your sequence within a bank of your choice, with a tool of your choice (*Basic BLAST*)
- Make specialized searches (*Specialized BLAST*)

For this practical, unless otherwise specified, we will choose the links from the Basic BLAST section.

Several versions of the software are suggested, depending on the nature of the request sequence, and of the interrogated bank:

Nucleotide

- Compare a nucleic sequence to a nucleic bank (BlastN): useful to study a sequence that does not encode a protein, or to localize an mRNA on a genome and *the other way around*.
- MegaBlast: a software optimized to align sequences that modestly differ, and 10 times faster than BlastN. This software has been conceived two compare two sets of sequences to each other.

An interface dedicated to short sequences with few errors: the parameters are adapted to that kind of data.

Protein

- Compare a protein sequence to a protein bank (BlastP): search for a protein's homologs.
- PHI-Blast (Pattern Hit Initiated BLAST) : this software takes as entry a protein request sequence, and a motif defined by a regular expression (see the syntax). It searches for

proteins that contain the motif, and that are similar to the request protein's sequence neighboring the motif. This software is useful to search for a motif (protein domain, active site, ...) in a protein sequences databank. Its advantage, compared to a simple motif search, resides in the fact that it relies on the request sequence (neighboring study) as to eliminate sequences that contain the motif randomly.

- PSI-Blast (Postion Specific Iterated Blast) : a profile is constructed from the multiple alignment of sequences that got the best scores with the request sequence. This profile is compared to the interrogated bank, and is refined progessively with each iteration. Therefore, the sensitivity of the software is increased. PSI-Blast is useful to detect far away protein family members, and to study the function of unknown proteins.

blastx

Compare a nucleic sequence translated in the 6 reading frames to a protein bank: useful to know if a nucleic sequence codes for a protein, and eventually to localize the positions of the coding part.

tblastn

Compare a protein sequence to a translated nucleic bank, translated in the 6 frames: useful to identify the gene and/or the mRNA which encodes a protein.

tblastx

Compare a nucleic sequence translated in the 6 reading frames to a nucleic bank translated in the 6 reading frames (tBlastX): useful to compare a nucleic sequence of which we know nothing about to an unannotated genome, or when BlastN doesn't give out results. Use with moderation, takes a very long time!

First steps

This exercise deals with the sequence analysis of conversion enzymes of angiotensin I to angiotensin II, also called ACE. Below, the nucleotide sequence of the leech's ACE mRNA:

```
>Leech, ACE
aatttaaaaatgaatttaataaatttttcatacttaaatttgctttttggtgccggtttatttagcgttttagaaagcgc
tacaatattaaataccgaatcggatgctaaaaaatggctgacaacgtataacgatgaagccggaaaatatatttacgatg
caactgaagcagaatggaattacaacaccaacctgactgatcacaatttaggaatttctattaaaaaatcaaatgatttg
gctacttttacggaacaaaaggcaatcgaggccaataaaaaatttgtatggaaaaattttactgatccacttttgaaaag
agaattttcaaaaataactgacattggtactgctagcctttcagatgaagactttcaaaagatgtcaggtttgaactctg
atctaacaaaaatttacagcactgcaaaagtttgtaacaagcctaacgacccatctggaaaatgctatcctttagatcct
gatttgtccgacataatctccaagtcaaacgatctcgaggaattgacctgggcatggaaaggttggagggatgcgtctgg
caaacatatgcccgataaatatgatgaatttgttcaactgctcaacaaagctgctaagattcatggatatgaagacaacg
ggattattggaggtcctggtacgagtcccccacgttcagaaaggattgtgaagatttgtggcaggagatcaaaccattc
tacgaacaactgcatgcatacgtcagaaggaagctgcagaagaagtatccccaaattgcattccccaaggaggggcccat
ccctgctcatctgctcggcaacatgtgggcccaatcgtgggagaacatagagtacttgttatgggcccaatcgtgggaga
acatagagtacttgttaaggcccgctcctgaccttcctagcatggacatcactgaggaactcgtcaaacagaactacacg
gcattgaaactcttccaactgtcggacacatttttcaaatccttgggtctcatccagatgcctcagccgttttgggaaaa
gtcgatgatcgagaaaccagctgatcgggatgtgttcagaatcaaacaatgcgtttgccatgcgtcagcctgggacttct
acaatcgcaaggatacggttgtggacatgcactggttcatgacgactcaccatgagatgggacacatcgaatactacctc
cactacaaggaccaacccatcagtttcagatctggcgctaatccaggatttcatgaggccattgccgatattgcatcact
gtcagtggccacacctgaatatatgcaatccgtcagcctgttgcctaatttcactgacgatccaaatggcgatttaaact
tcttaatgaaccaagccttaacgaaggtggccttcctaccattcggttacctgatcgaccagtggagatgggacgtgttc
tcgggagataccctcgaccaaaatacaactccaagtggtggcacaacaggtgtaagtaccagggcatatatcctccagt
gaaaaggtcagagcaagattttgatgccggttccaagttccatgtacccaacaacactccatacatcaggtactttgttg
ctcacgtcatccaattccaattccatgaagccctgtgcaaggctgccaacaacagcagacctctacatagatgtaacatc
gccaattccaaggaagctggagagaaactggctgaattgatgaaatctggatcttcaattccgtggcctaaagttctaga
aaatcttactggatcggaaaaaatgtcagcgaaatctctcatggcctattacaaaccgttgatcgattggcctgaaaaaa
gaaaaccaagggcagaaaattggatgggaggaaaaatgtcctcctggatcatttgaaccatgaaattatttatttgattt
tatgtcatttcataatttttctaccacttttttaataaacttaggtgcctattgaatatgttcttgcaatttgaaaaaaa
aaaaaaaaaaaaaaaaaaaaaaaaaaaaa
```

DNA sequence request against nucleic bank

Discovery

Follow the link *nucleotide blast* in the section *Basic BLAST*. Copy-Paste the sequence below in the corresponding box, choose the databank (database) *Other : Nucleotide collection (nr/nt)*, select the software *Somewhat similar sequences (blastn)*, limit the search to entries from Man. To do so, in the *Organism* field, enter `Homo sapiens`, and launch the request.
[results]

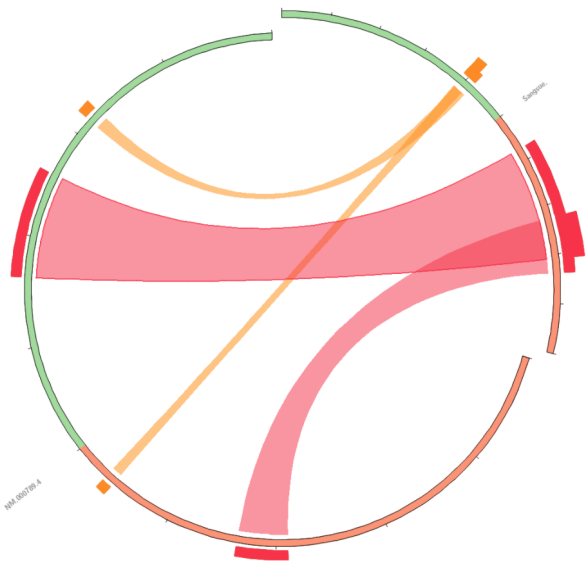
> **Question 1**
>
> How many human sequences within the bank look like ours (check the number of "hits")?
> Do the obtained alignments seem relevant from a biological point of view?

> **Question 2**
>
> What do the terms "Total Score" and "Query Coverage" in the results table mean ?

The results' graphical representation indicates the regions of the request sequence (scale) that align with the sequences from the bank (colored rectangles). Only the rectangles linked by a black line are from the same bank entry.

> **Question 3**
>
> How many common regions between the sequence with accession number `NM_000789` and the leech's are represented on the graph? (Slide the cursor on the colored lines so that the entry number is displayed just above the *Color key for alignment scores*)
> How many regions are truly common to the 2 sequences? (Check the alignments by clicking either on the associated traits on the graph, or on the alignment score in the results table under the graph).

The below graph represents a reciprocal projection of the query (leech sequence) and of the subject (`NM_000789`).



> **Question 4**

What can we learn from this graph, that we couldn't learn from the Blast graph?

The E-value

Question 5

What is the score obtained for the best alignment of the human sequence `NM_000789` ?
What is the corresponding e-value ?
How does the e-value vary with the score (compare it for different alignments of this sequence)?

Save this resultst page in *complete web page* format (or keep it in an open tab): we will need it later on for the rest of the practical.

Now, launch BLAST, limiting the data to the `refseq_rna` databank (keeping the entries limited to the ones from `Homo sapiens`). For this, choose the appropraite databank in the interrogation form.
[results]

Question 6

Do we find the ACE sequence `NM_000789` ?
Has the score of the best alignment changed ?
How has the e-value varied ?
Why is there a change in e-value without a change in score ?

Choice of software

Question 7

How to check (as efficiently as possible) if the sequence we have is present in the entire nr bank?
Is it indeed present ?

[result]

Low complexity filter

Low complexity regions are part of sequences composed of few different letters. By default, in Blast, the option "Low complexity" is checked. Low complexity regions in the request sequence are replaced by `N` in the case of DNA, and by `X` in the case of proteins. Therefore, those regions are not aligned to sequences from the bank. We will study the interest of this option.

You can observe the low complexity region present in the leech sequence by creating a dotplot of the sequence against itself.

Use BLAST again, with the leech sequence against all the sequences from the bank (keeping limiting to Human and to `refseq_rna`), but by unchecking the "Low complexity regions" option in "Algorithm parameters".

The obtained results are then different from the previous ones. Yet, the request sequence and the bank are the same.
[result]

Question 8

How many sequences from the bank now look like ours?
Which part of our sequence is more detected than previously? On what type of sequences does it match? In your mind, why?
Which entries, among the ones from the previous results graph, no longer appear when we deactivate low complexity filters (Use `CTRL + F` to make a search) ? In your mind, where are they now ?

DNA sequence request agains protein databank.

In this case, the request sequence is first blindly translated into the six reading frames by BlastX. The six obtained peptides are then searched for, and aligned with the proteins from the bank (including the stop codons of the request peptides, which are replaced by stars).

BlastX

Launch a BlastX with the leech sequence, restraining the search to mammals (*Mammalia*).
[result]

Question 9

How many sequences from the bank look like ours?
What is the E-value of the 2 first sequences from the list ?
The one of the last 2 ?
Compare the found values to those obtained with BlastN.

Question 10

Are more sequences from the ACE family found (don't count!) ?

Question 11

What organism does the first found sequence come from ?

In the results, find the ACE protein corresponding to the human sequence.

Question 12

How many hits are there on this protein sequence?
Compare this alignment to the one previously obtained with BlastN.
What do you observe (E-value, alignment coverage, result quality)?

Sensitivity to parameters

Modification of word size

Make a request with the help of **BlastN** ("nucleotide blast") against the *nr* bank with the MAKORIN1 gene from *Seriola quinqueradiata* (keep the window open for later).
Now, change, in "Algorithm parameters", the size of the searched exact words, as to be more specific.

[résultat pour w=7,résultat pour w=11, résultat pour w=15]

Question 13

What size do you choose ?
Compare the results with those obtained with the default word size. What do you observe?

Modification of the "Expect threshold" value

Question 14

What is the default value of this parameter? What does it mean?

Now, change the *Expect threshold* value as to be more specific.

[result]

Question 15

What value do you choose ?
Compare the results to those obtained with the default value. What do you observe?

Modification of penalties and of the score matrix in BlastN

Searching how to edit those parameters.

Question 16

What are the default parameters (opening and extension gap penalty, match and mismatch scores)?

Make a BlastN request against the *nr* bank with the MAKORIN1 from *Seriola quinqueradiata* with the least penalizing gap penalties (keep the window open for later).

[result]

Question 17

What differences do you observe with the request with default parameters?
Search for hits on the MAKORIN gene, limiting the search to pig (*Sus scrofa*). Explain the results obtained with the two sets of parameters.

MegaBLAST

Launch a MegaBLAST request with the MAKORIN1 gene from *Seriola quinqueradiata*.

[result]

Question 18

What are the differences with BlastN ?
Observe the results obtained on *Zebrafish* sequences, why do we not see the same results with MegaBLAST ?

PSI-Blast

Here is a protein from *E. coli* :

```
>trpc
MMQTVLAKIVADKAIWVEARKQQQPLASFQNEVQPSTRHFYDALQGARTAFILECKKASP
SKGVIRDDFDPARIAAIYKHYASAISVLTDEKYFQGSFNFLPIVSQIAPQPILCKDFIID
PYQIYLARYYQADACLLMLSVLDDDQYRQLAAVAHSLEMGVLTEVSNEEEQERAIALGAK
VVGINNRDLRDLSIDLNRTRELAPKLGHNVTVISESGINTYAQVRELSHFANGFLIGSAL
MAHDDLHAAVRRVLLGENKVCGLTRGQDAKAAYDAGAIYGGLIFVATSPRCVNVEQAQEV
MAAAPLQYVGVFRNHDIADVVDKAKVLSLAAVQLHGNEEQLYIDTLREALPAHVAIWKAL
SVGETLPAREFQHVDKYVLDNGQGGSGQRFDWSLLNGQSLGNVLLAGGLGADNCVEAAQT
GCAGLDFNSAVESQPGIKDARLLASVFQTLRAY
```

Launch PSI-blast on it, in the NR bank.
[result iteration 1, result iteration 2]

> **Question 19**
>
> During the first iteration, what are the found protein functions? Are there several? Save the entire web page in your working directory.

Launch the second iteration (button available on the results page obtained during the first iteration, under the hits diagram).

> **Question 20**
>
> Are the results different? According to you, why? What are the sequences that appeared? Those that disappeared?

cDNA Analysis

The sequence we're going to study comes from a fish (of which you'll search the description on fishbase). It is the DNA copy of an mRNA.

```
>Gasterosteus aculeatus cDNA clone CLJ188-G12 5', mRNA sequence
AATTGGACATGACAGTTCGGTCCGGAATCCCGGGATGGAGATGCCATCCGTTGGATCCGGATCTTCAGAA
GATCATGGCGGAGTCCAGGGATTATGACGAACTGCTGTTTGCGTGGAAAGGATGGAGAGATTCTGCCGGC
AAAGTGCTTCGCCAGGATTACAAGAGATATGTTGAACTGGCCAACATGGCCGCCAAACTCAACGGTCACT
CCGACAACGGGGCTTCCTGGCGCTCCCTGTATGAAACCCCCAGCTTCGAGGAGGACCTGGAGGCTCTGTG
GAAGGAGCTGGAGCCGCTCTATCAGAATGTGCACGCCTATGTGCGCAGGGCCCTGTACAAAAAGTATGGC
TCCCAGCACATCAACCTGAAGGGAGCCATCCCGGCTCATTTGCTGGGCAACATGTGGGCCCAGACGTGGT
CGGGCATAATGGATTTGGTCATGCCCTACCCGCATGCCACGCAGGTGGACGCCACGCCCGCCATGGTTTC
ACAGGGCTGGAACGCCACCAGAATGTTCCAGGAATCCGACAATTTTTTCACCTCTCTGGGTCTTTTGCCA
ATGCCCCAAGAGTTCTGGGACAAATCCATGCTAGAGAAGCCGTCTGGTGGACGCCAGGTGGTGTGCCACG
CCTCCGCATGGGACTTCTATAACCGAAAAGACTTCAGGATCAAACAGTGCACCGTGGTGACTATGGACGA
TTCCGCGACGGAGCCAACCCCGGCTTCCACGAGGCCATTGGCGACGTGTTGGCCCTGTCAGTGTCTACGC
CCTGATCACGGCGCACCATGAGATGGGCCACATTCAGTACTTCCTGCAGTACAAAGACCAGCCCGTGTCC
CCAAACACCTGCAGAGCATCGGCCTGCTGGACAAAGTGGAGAGCAACCATGAGAGCGATATCAACTTCCT
GATGAGCATGGCGCTCGACAAGATCGCCTTCTTACCCTTTCGCTACCTGATGGATCAGTGNAGATGGAAG
GGTGTCGATGGNCGTATCCCATCGACTGAGTANCATAAAGAATGGNTGGAACTCAGAATGAAGTACCAGG
GCCTCTGTCCCACTGTAACCCGCACAGAGGAAGACTTCGACCANGTGCAAAAGTCACATCCCTGCTACGT
GCCATACGTGAAGAACTTTGTCACCTCATCATCAGGTCCAGGTTCCAAAGCTCTCTGGGATGCCCCAAAA
CGAAGGGGCCTGGAAACTGGAAATTTTAAATTCGGAAAACCCGGACCTCTTGCGGACGATGAAACCCGTT
TCTAAAACCTGGCCCGGGGAAA
```

With Discontiguous MegaBLAST

Compare this cDNA (from *stickleback*) to the **nucleic** bank "nr/nt" by using *More dissimilar sequences (discontiguous megablast)*.

[result]

> **Question 21**

> Do the found sequences really look like the request sequence ? Have we managed to localize the mRNA on the genome of *Gasterosteus aculeatus* ? Can we get an idea of the function of the protein encoded by the studied mRNA ?

The genome of *Gasterosteus aculeatus* is undergoing sequencing: the data are not in the "nr/nt" bank, but in the "wgs" bank ("whole genome shotgun").

Interrogate (still with *More dissimilar sequences, discontiguous megablast*) the sequences from the "wgs" bank for the *Gasterosteus aculeatus* organism.

[result for the entire WGS, result for the WGS limited to Gasterosteus]

> **Question 22**
>
> Can we localize the mRNA ? Can we position it on the genome ? Why is that not possible? What is the accession number and the size of the found contig ? Why is the alignment between the mRNA and the contig split up? Note on the graph the black vertical bars between the red hits, indicating that these hits come from the same bank sequence.

With Ensembl

We will make a search on the Ensembl website, of the cDNA against the *Gasterosteus aculeatus* genome, as to get more information on the positions predicted by *Ensembl* for the whole genome.

Click on the *BLAST/BLAT* link from Ensembl, give your sequence, and choose the *Gasterosteus aculeatus* species on which this sequence will be localized. Launch the comparison, then click on "[View results]".

> **Question 23**
>
> On which chromosome is the gene present? In which direction is the gene represented on the contig ?

In *Genomic location*, it is possible to vizualize hits found by BLAT *visually* on the chromosome. To do so, click on one of the entries, on the link right before "[sequence]" : it will lead you to the "Region in detail" interface for this entry.

You should have a fairly complex image described as "Region in detail", having the **hit** (the **hits** if you dezoom) of BLAT/BLAST in the forward sense, as well as in the inverse complementary sense.

Compare the positions of the exons found on our sequence ("BLAT/BLAST hits" in red/brown) with those of the *predicted* exons by Ensembl ("Genes [Ensembl]").

> **Question 24**
>
> Are the main hits of our request rather at the start, or at the end of the gene?

On the protein bank `nr`

By coming back to the NCBI interface, now compare the **cDNA** sequence to the **protein** bank "nr", by choosing the **right** program.

Question 25

Are the results more satisfying (better similarity) than those obtained with the cDNA against the nucleic bank? According to you, why?
What can we learn about the protein's function ? What do you notice by consulting the obtained alignments (for example, consult the first hit on a `sp` "Swiss-Prot" entry) ? Does the studied sequence contain a full coding sequence ?