

# Utilisation de la base de données et des outils fournis par le NCBI (Bases de données, BLAST)

Bénédicte Condamine et Marie Petitjean

## Résumé

Le but est de comprendre et de savoir utiliser les bases de données hébergées par le NCBI ainsi que les outils rattachés tels BLAST et Global Align.

Comprendre le fonctionnement des bases de données et savoir y rechercher des informations. Utilisation des outils avec les paramètres simples et avancés.

## Table des matières

<b>1</b>	<b>Bases de données NCBI - <i>National Center for Biotechnology Information</i></b>	<b>2</b>
1.1	Recherche d'information sur les gènes/protéines . . . . .	3
1.2	Informations sur les génomes . . . . .	3
1.3	Taxonomie . . . . .	3
1.4	Bibliographie . . . . .	3
1.5	Submission . . . . .	3
<b>2</b>	<b>Recherche de similarité : BLAST - <i>Basic Local Alignment Search Tool</i></b>	<b>4</b>
2.1	Principe de fonctionnement . . . . .	5
2.2	Les paramètres avancés : . . . . .	7
2.3	Recherche d'informations : est-ce que le gène X est présent dans le génome Z? . . . . .	8
2.4	Comparaison avec un Alignement Global . . . . .	8
2.5	Recherche de motifs . . . . .	9
2.6	Ligne de Commande . . . . .	9
2.6.1	Création de la base de données . . . . .	9
2.6.2	Lancement du BLAST . . . . .	9

# 1 Bases de données NCBI - *National Center for Biotechnology Information*

Site du NCBI : [www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov)

## Literature

### PubMed

PubMed® comprises more than 33 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full text content from PubMed Central and publisher web sites.



**Example searches** Search for titles, citations, identifiers and more

[Revealing protein-protein interactions by transcriptome sequencing](#)

[Molecular Biology of the Cell](#)

[PMC7207159](#)

[Cancer Statistics 2021](#)

[Kumon, Cell 2021](#)

### Literature databases

#### Bookshelf

Books and reports

#### MeSH

Ontology used for PubMed indexing

#### NLM Catalog

Books, journals and more in the NLM Collections

#### PubMed

Scientific and medical abstracts/citations

#### PubMed Central

Full-text journal articles

## Data

### Genes

Gene sequences and annotations used as references for the study of orthologs structure, expression, and evolution

#### Gene

Collected information about gene loci

#### GEO DataSets

Functional genomics studies

#### GEO Profiles

Gene expression and molecular abundance profiles

#### HomoloGene

Homologous genes sets for selected organisms

#### PopSet

Sequence sets from phylogenetic and population studies

### Proteins

Protein sequences, 3-D structures, and tools for the study of functional protein domains and active sites

#### Conserved Domains

Conserved protein domains

#### Identical Protein Groups

Protein sequences grouped by identity

#### Protein

Protein sequences

#### Protein Family Models

Models representing homologous proteins with a common function

#### Structure

Experimentally-determined biomolecular structures

### BLAST

A tool to find regions of similarity between biological sequences

#### blastn

Search nucleotide sequence databases

#### blastp

Search protein sequence databases

#### blastx

Search protein databases using a translated nucleotide query

#### tblastn

Search translated nucleotide databases using a protein query

#### Primer-BLAST

Find primers specific to your PCR template

### Genomes

Genome sequence assemblies, large-scale functional genomics data, and source biological samples

#### Assembly

Genome assembly information

#### BioCollections

Museum, herbaria, and other biorepository collections

#### BioProject

Biological projects providing data to NCBI

#### BioSample

Descriptions of biological source materials

#### Genome

Genome sequencing projects by organism

#### Nucleotide

DNA and RNA sequences

#### SRA

High-throughput sequence reads

#### Taxonomy

Taxonomic classification and nomenclature

### Clinical

Heritable DNA variations, associations with human pathologies, and clinical diagnostics and treatments

#### ClinicalTrials.gov

Privately and publicly funded clinical studies conducted around the world

#### ClinVar

Human variations of clinical significance

#### dbGaP

Genotype/phenotype interaction studies

#### dbSNP

Short genetic variations

#### dbVar

Genome structural variation studies

#### GTR

Genetic testing registry

#### MedGen

Medical genetics literature and links

#### OMIM

Online mendelian inheritance in man

### PubChem

Repository of chemical information, molecular pathways, and tools for bioactivity screening

#### BioAssays

Bioactivity screening studies

#### Compounds

Chemical information with structures, information and links

#### Pathways

Molecular pathways with links to genes, proteins and chemicals

#### Substances

Deposited substance and chemical information

(?) Quelles bases connaissez vous ? Que contiennent-elles ?

## 1.1 Recherche d'information sur les gènes/protéines

**Nucleotide** : Séquences ADN et ARN  
**Protein** : Séquences protéiques (+ régions fonctionnelles)  
**Gene** : Informations sur les gènes (contexte génomique, transcription, bibliographie)  
=> Bases en lien les unes avec les autres

- (?) Recherche du gène *lasR* chez *Pseudomonas aeruginosa* PAO1
- (?) Comparer les résultats en fonction des trois bases citées précédemment. Qu'observe t-on ?
- (?) Regarder la recherche avancée (opérateurs AND, OR et NOT ; balises ORGANISM, SLEN)

## 1.2 Informations sur les génomes

**Genome** : Informations sur l'organisme (séquence de référence, taxonomie, bibliographie)  
**Assembly** : Informations sur le séquençage d'un génome (souche, chromosome(s))  
=> Lien avec la base de données nucléotides qui contient les séquences des chromosomes

- (?) Recherche du génome de DHS01
- (?) Comment rechercher de l'information dans le génome ? gène *gyrB* ?
- (?) Récupération de données : genbank, fasta

## 1.3 Taxonomie

**Taxonomy** : Retrouver la taxonomie d'un organisme

- (?) Est-ce que les organismes *Helicobacter pylori* et *Pseudomonas aeruginosa* sont très éloignés ?
- (?) Idem pour *Escherichia coli* et *Klebsiella pneumoniae*

## 1.4 Bibliographie

**Pubmed** : Articles scientifiques

- (?) Trouver les articles parlant de *Pseudomonas aeruginosa* dans *BMC microbiology* depuis 2015

## 1.5 Submission

**sra** : archives partagées du haut-débit (SRA du NCBI, ENA de l'EBI et DDRA de la DDBJ). Lors du téléchargement, les données nécessitent l'utilisation de SRA Toolkit pour convertir les données dans le format d'intérêt) (<https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software>)  
**BioSample** : description de matériel biologique utilisé pour les expériences  
**BioProject** : collection de données biologiques liées entre elles (en lien avec un projet ou une organisation)

- (?) Chercher *Pseudomonas aeruginosa* DHS01 dans les bases de données suivantes *BioProject*, *BioSample*, *Assembly* et *Nucleotide*. Quelles informations peut-on tirer de chaque base de données ?

## 2 Recherche de similarité : BLAST - *Basic Local Alignment Search Tool*

**BLAST** (acronyme de **b**asic **l**ocal **a**lignment **s**earch **t**ool) est une méthode de recherche heuristique utilisée en bio-informatique. Il permet de trouver les régions similaires entre deux ou plusieurs séquences de nucléotides ou d'acides aminés, et de réaliser un alignement de ces régions homologues.

On différencie un alignement local d'un alignement global. L'alignement global compare des séquences homologues (apparentées) sur toute leur longueur.

Étant donné une séquence introduite par l'utilisateur, BLAST permet de retrouver rapidement dans des bases de données, les séquences répertoriées ayant des zones de similitude avec la séquence d'entrée. Cette méthode est utilisée pour trouver des relations fonctionnelles ou évolutives entre les séquences et peut aider à identifier les membres d'une même famille de gènes. (Source : Wikipedia)

<https://blast.ncbi.nlm.nih.gov/Blast.cgi>

The screenshot shows the BLAST web interface. At the top, it says "Basic Local Alignment Search Tool". Below this, a description states: "BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance." There is a "Learn more" link. To the right, a "NEWS" banner reads: "End of updates for BLAST+ version 4 databases (dbV4). Start moving to the new version 5 databases! Fri, 27 Sep 2019 16:00:00 EST. More BLAST news...".

The "Web BLAST" section features three main options:

- Nucleotide BLAST**: nucleotide → nucleotide
- blastx**: translated nucleotide → protein
- tblastn**: protein → translated nucleotide
- Protein BLAST**: protein → protein

At the bottom, there is a "BLAST Genomes" section with a search input field: "Enter organism common name, scientific name, or tax id" and a "Search" button. Below the input field are links for "Human", "Mouse", "Rat", and "Microbes".

Les différents types de BLAST :

**blastn** : recherche une séquence nucléotidique dans une base de nucléotides

**blastp** : recherche une séquence protéique dans une base de protéines

**blastx** : recherche une séquence nucléotidique dans une base de données protéique, la séquence est traduite (3 cadres de lectures et les 2 sens) afin d'effectuer les recherches

**tblastn** : recherche d'une séquence protéique dans une base de données nucléotidique, la base de données de nucléotides est traduite pour la recherche

**tblastx** : recherche d'une séquence de nucléotides traduite en protéine dans une base nucléotide traduite en protéine

NIH U.S. National Library of Medicine NCBI Sign in to NCBI

**BLAST®** » blastn suite Home Recent Results Saved Strategies Help

Align Sequences Nucleotide BLAST

blastn blastp blastx tblastn tblastx

Enter Query Sequence BLASTN programs search nucleotide subjects using a nucleotide query. [more...](#) [Reset page](#) [Bookmark](#)

Enter accession number(s), gi(s), or FASTA sequence(s) [Clear](#) Query subrange [↓](#)

From

To

Or, upload file  Aucun fichier sélectionné. [↓](#)

Job Title

Enter a descriptive title for your BLAST search [↓](#)

Align two or more sequences [↓](#)

Enter Subject Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [Clear](#) Subject subrange [↓](#)

From

To

Or, upload file  Aucun fichier sélectionné. [↓](#)

Program Selection

Optimize for

Highly similar sequences (megablast)

More dissimilar sequences (discontiguous megablast)

Somewhat similar sequences (blastn)

Choose a BLAST algorithm [↓](#)

**BLAST** Search nucleotide sequence using Megablast (Optimize for highly similar sequences)

Show results in a new window

[Algorithm parameters](#) **Note: Parameter values that differ from the default are highlighted in yellow and marked with a sign**

BLAST is a registered trademark of the National Library of Medicine [Support center](#) [Mailing list](#) [YouTube](#)

## 2.1 Principe de fonctionnement

1. : Décomposition de la séquence à tester en mots de longueur  $k$  (graines) pour former un dictionnaire
2. : Balayage de la base avec le dictionnaire, si trouvé, il essaye de voir si les séquences en amont et en aval sont similaires
3. : Calcul du score et de l'E-value

### Calcul du score en utilisant une matrice de substitution

Une matrice de substitution permet d'associer un score à chaque paire de résidus que l'on trouve dans un alignement. On utilise des matrices spécifiques selon le type de BLAST. Les matrices BLOSUM ou PAM sont utilisées pour les protéines, elles prennent en compte les caractères des acides aminés (hydrophobes, aromatique, polaire, basique et acide). Les scores positifs indiquent des substitutions fréquentes. Les valeurs négatives indiquent des mutations rares. Pour les séquences nucléotidiques, on utilise généralement des pénalités identiques pour toutes les substitutions. Pour un alignement donné, le score est la somme des scores de chaque paire de résidus. Dans le cas de gap, l'ouverture est sanctionnée par un score négatif plus faible que son extension.

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

Exemple de matrice de substitution - BLOSUM 62

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	9																			
S	-1	4																		
T	-1	1	5																	
P	-3	-1	-1	7																
A	0	1	0	-1	4															
G	-3	0	-2	-2	0	6														
N	-3	1	0	-2	-2	0	6													
D	-3	0	-1	-1	-2	-1	1	6												
E	-4	0	-1	-1	-1	-2	0	2	5											
Q	-3	0	-1	-1	-1	-2	0	0	2	5										
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8									
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5								
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5							
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5						
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4					
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4				
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4			
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6		
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7	
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11

- (?) Prenons les séquences suivantes : GPAFDYSTVHA et GCPRFETHVA, quels sont les scores des alignements suivants (ouverture de gap = -11 et extension = -1, valeurs par défaut de blastp) ?
- (?) Idem avec une ouverture de gap et une extension de -1

(?) Qu'observe t'on ?

```
G - P A F D Y S T - V H A
G C P R F E - - T H V - A
```

et

```
G - P A F D Y S T V H A
G C P R F E - - T H V A
```

## 2.2 Les paramètres avancés :

The image shows a screenshot of the BLAST search interface. At the top, there is a search bar with the text "BLAST" and "Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar sequences)". Below this, there is a checkbox for "Show results in a new window". The main interface is divided into several sections:

- Algorithm parameters:** A tab that is currently selected, with a "Restore default search parameters" link on the right.
- General Parameters:**
  - Max target sequences:** A dropdown menu set to "100". Below it is the text "Select the maximum number of aligned sequences to display".
  - Short queries:** A checkbox labeled "Automatically adjust parameters for short input sequences" which is checked.
  - Expect threshold:** A text input field set to "10".
  - Word size:** A dropdown menu set to "28".
  - Max matches in a query range:** A text input field set to "0".
- Scoring Parameters:**
  - Match/Mismatch Scores:** A dropdown menu set to "1,-2".
  - Gap Costs:** A dropdown menu set to "Linear".
- Filters and Masking:**
  - Filter:** A checkbox labeled "Low complexity regions" is checked. Below it is a checkbox labeled "Species-specific repeats for:" followed by a dropdown menu set to "Homo sapiens (Human)".
  - Mask:** A checkbox labeled "Mask for lookup table only" is checked. Below it is a checkbox labeled "Mask lower case letters".

At the bottom of the interface, there is another search bar with the text "BLAST" and "Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar sequences)", and a checkbox for "Show results in a new window".

FIGURE 1 – Options avancées - blastn

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

**Algorithm parameters** [Restore default search parameters](#)

**General Parameters**

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

Short queries:  Automatically adjust parameters for short input sequences

Expect threshold: 10

Word size: 6

Max matches in a query range: 0

**Scoring Parameters**

Matrix: BLOSUM62

Gap Costs: Existence: 11 Extension: 1

Compositional adjustments: Conditional compositional score matrix adjustment  
Matrix adjustment method to compensate for amino acid composition of sequences. [more...](#)

**Filters and Masking**

Filter:  Low complexity regions

Mask:  Mask for lookup table only  
 Mask lower case letters

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

FIGURE 2 – Options avancées - blastp

### 2.3 Recherche d'informations : est-ce que le gène X est présent dans le génome Z ?

- (?) On recherche lasR dans le génome de DHS01 avec un blastn
- (?) Pareil pour mexR, avec les différents types de BLAST. Quels résultats pertinents ?

### 2.4 Comparaison avec un Alignement Global

Le NCBI permet aussi de réaliser un alignement global avec Global Align. Ce logiciel permet de comparer deux séquences sur toutes leurs longueurs grâce à l'algorithme de Needleman-Wunsch. <https://blast.ncbi.nlm.nih.gov/Blast.cgi>

- (?) Faire un alignement entre le gène mexR de *Pseudomonas aeruginosa* PAO1 et mexR venant de *Pseudomonas citronellolis* strain P3B5
- (?) Quelles différences peut-on observer par rapport aux résultats de BLAST ?

## 2.5 Recherche de motifs

- (?) Reprendre les séquences précédentes et rechercher des motifs, y a t'il des domaines en commun entre les deux ?
- (?) Quels sont les domaines protéiques d'intérêt dans la séquence des fichiers unknown.fasta et unknown2.fasta ? (/home/progs/docker/Club\_bioinfo/Blast/data/fasta)

## 2.6 Ligne de Commande

### 2.6.1 Création de la base de données

On utilise la commande "**makeblastdb**" avec plusieurs options :

- in** : le fichier d'entrée à partir du quel sera fait la base de données
- dbtype** : type de base de données : nucl ou prot
- out** : le nom que portera la base de données (optionnel, par défaut le nom du fichier d'entrée)

```
$ makeblastdb -in file.fa -dbtype nucl -out nameDB
```

- (?) Créer une base de données avec les gènes de l'opéron Arginine

### 2.6.2 Lancement du BLAST

Les différents types de BLAST se lancent grâce à leurs noms : **blastn**, **blastp**, **blastx**, **tblastn** et **tblastx**. Pour avoir accès aux différentes options utiliser l'option **-help**.

```
$ blastp -help
```

Les différentes options et paramètres du blastn (liste non exhaustive) :

- task** : algorithme BLAST (blastn, megablast, ...)
- query** : fichier contenant la ou les séquences que l'on veut comparer à la base de données. (obligatoire)
- db** : nom de la base de données d'intérêts
- subject** : fichier contenant la ou les séquences pour faire du blast deux à deux
- out** : nom du fichier de sortie
- evalue** : seuil pour la e-value
- word\_size** : taille des mots
- gapopen** : pénalité d'ouverture de gap
- gapextend** : pénalité d'extension de gap
- penalty** : pénalité de mismatch
- reward** : récompense de match
- num\_descriptions** : nombre de descriptions demandé
- num\_alignments** : nombre d'alignements demandé
- perc\_identity** : seuils d'identité
- num\_threads** : nombre de threads
- remote** : lancement du blast en local avec une base de données à distance sur le NCBI
- outfmt** : permet de choisir un format de sortie grâce à un nombre de 0 à 18

Il existe différents formats :

Options	Significations
0	Pairwise,
1	Query-anchored showing identities,
2	Query-anchored no identities,
3	Flat query-anchored showing identities,
4	Flat query-anchored no identities,
5	BLAST XML,
6	Tabular,
7	Tabular with comment lines,
8	Seqalign (Text ASN.1),
9	Seqalign (Binary ASN.1),
10	Comma-separated values,
11	BLAST archive (ASN.1),
12	Seqalign (JSON),
13	Multiple-file BLAST JSON,
14	Multiple-file BLAST XML2,
15	Single-file BLAST JSON,
16	Single-file BLAST XML2,
17	Sequence Alignment/Map (SAM),
18	Organism Report

Par défaut le format 6, "tabular", présente 12 colonnes sans entêtes :

Column	Name	Significations
1.	qseqid	query (e.g., gene) sequence id
2.	sseqid	subject (e.g., reference genome) sequence id
3.	pident	percentage of identical matches
4.	length	alignment length
5.	mismatch	number of mismatches
6.	gapopen	number of gap openings
7.	qstart	start of alignment in query
8.	qend	end of alignment in query
9.	sstart	start of alignment in subject
10.	send	end of alignment in subject
11.	eval	expect value
12.	bitscore	bit score

L'utilisation de l'option **-outfmt** permet aussi de personnaliser sa sortie :  
 \$ -outfmt "6 qseqid sseqid eval"

Identifiant	Signification
qseqid	Query Seq-id
qgi	Query GI
qacc	Query accession
qaccver	Query accession.version
qlen	Query sequence length
sseqid	Subject Seq-id
sallseqid	All subject Seq-id(s), separated by a ','
sgi	Subject GI
sallgi	All subject GIs
sacc	Subject accession
saccver	Subject accession.version
sallacc	All subject accessions
slen	Subject sequence length
qstart	Start of alignment in query
qend	End of alignment in query
sstart	Start of alignment in subject
send	End of alignment in subject
qseq	Aligned part of query sequence
sseq	Aligned part of subject sequence
evaluate	Expect value
bitscore	Bit score
score	Raw score
length	Alignment length
pident	Percentage of identical matches
nident	Number of identical matches
mismatch	Number of mismatches
positive	Number of positive-scoring matches
gapopen	Number of gap openings
gaps	Total number of gaps
ppos	Percentage of positive-scoring matches
frames	Query and subject frames separated by a '/'
qframe	Query frame
sframe	Subject frame
btot	Blast traceback operations (BTOP)
staxids	Subject Taxonomy ID(s), separated by a ','
sscinames	Subject Scientific Name(s), separated by a ','
scomnames	Subject Common Name(s), separated by a ','
sblastnames	Subject Blast Name(s), separated by a ',' (in alphabetical order)
sckingdoms	Subject Super Kingdom(s), separated by a ',' (in alphabetical order)
stitle	Subject Title
salltitles	All Subject Title(s), separated by a '<>'
sstrand	Subject Strand
qcovs	Query Coverage Per Subject
qcovhsp	Query Coverage Per HSP

(?) Récupérer les assemblages des génomes 1, 2, 3 et 4

- (?) Retrouver si et où sont positionnés les gènes de l'opéron Arginine dans les génomes 1, 2, 3 et 4 en utilisant la DB créée et en générant les sorties aux formats 1, 6 (par défaut et en le personnalisant), 8 et 12
- (?) Retrouver si et où sont positionnés les gènes de l'opéron Arginine dans les génomes 1, 2, 3 et 4 en utilisant le fichier fasta et en générant les sorties aux formats 1, 6 (par défaut et en le personnalisant), 8 et 12
- (?) Comparer les résultats