Utilisation de la base de données et des outils fournis par le NCBI

(Bases de données, BLAST)

Bénédicte Condamine et Marie Petitjean

Résumé

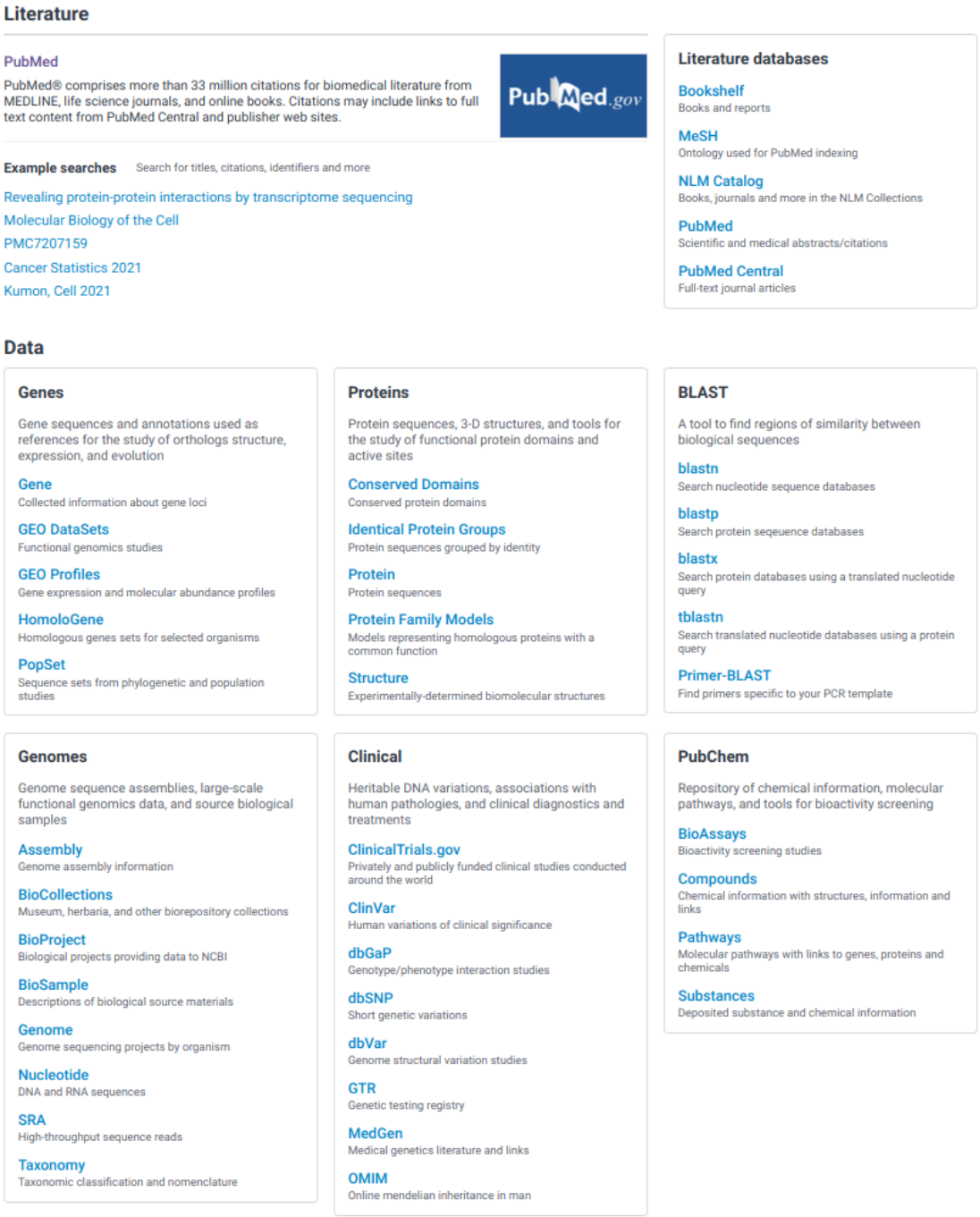
Le but est de comprendre et de savoir utiliser les bases de données hébergées par le NCBI ainsi que les outils rattachés tels BLAST et Global Align.

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

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

Site du NCBI : www.ncbi.nlm.nih.gov



(?) 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

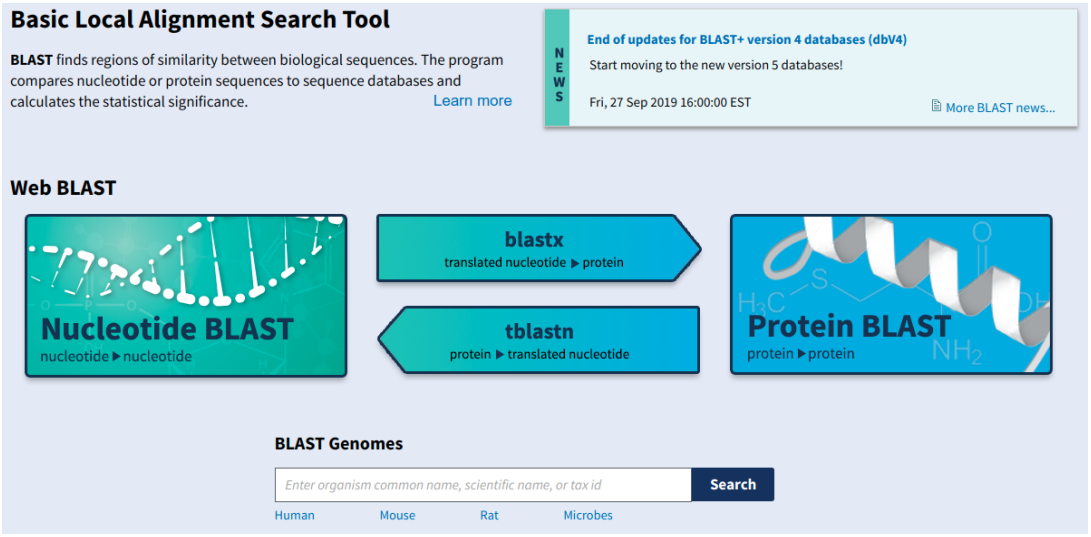
1.5 Submission

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)

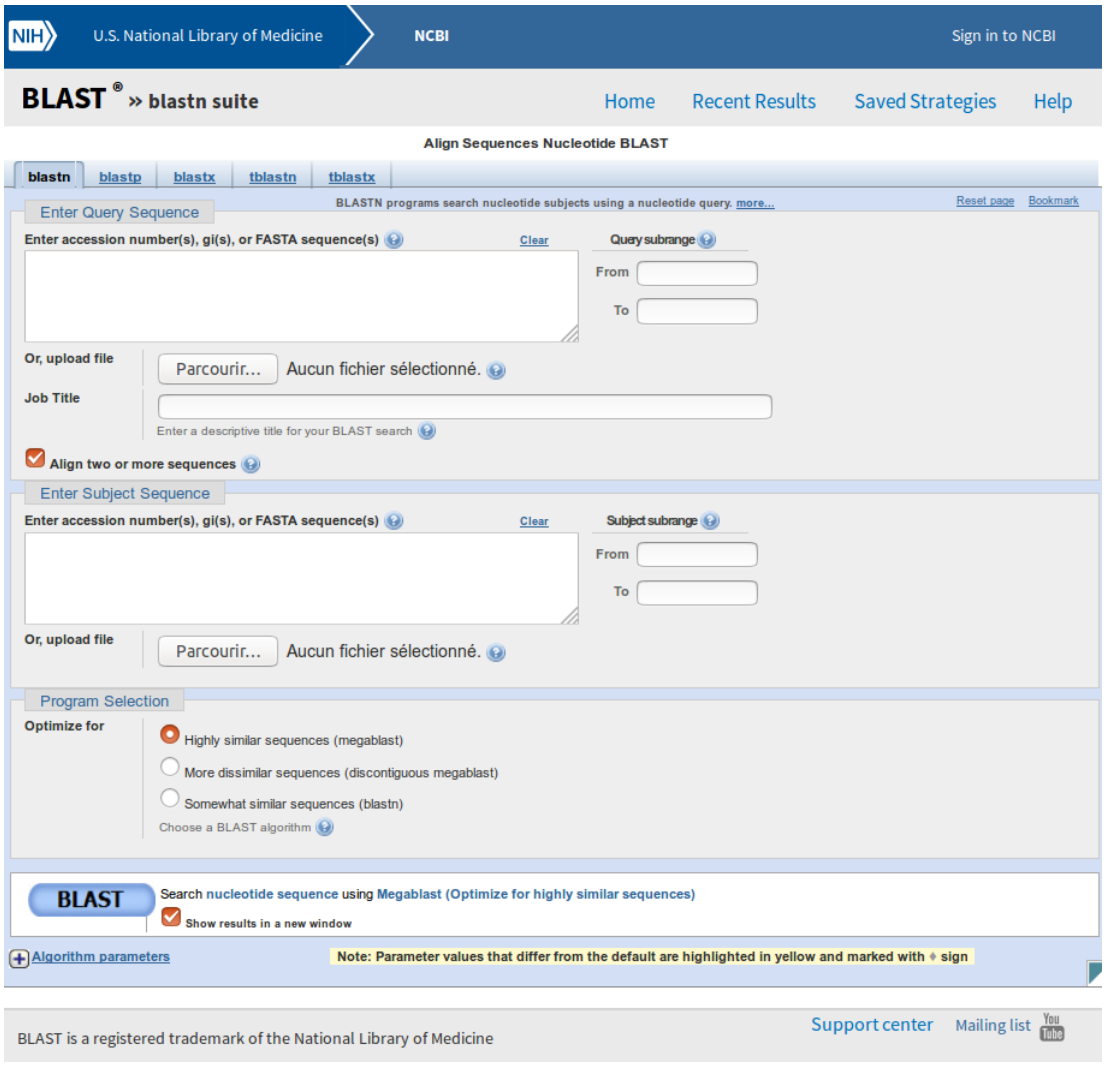
(?) Chercher *Pseudomonas aeruginosa* DHS01 dans les bases de données suivantes *BioProject*, *BioSample*, *Assembly* et *Nucleotide*. Quelles informations peut-on tirer de chaque base de données ?

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



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



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 matrice BLOSUM ou PAM sont utilisés 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 sanctionner par un score négatif plus faible que son extension.

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

	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

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

(?) 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 :

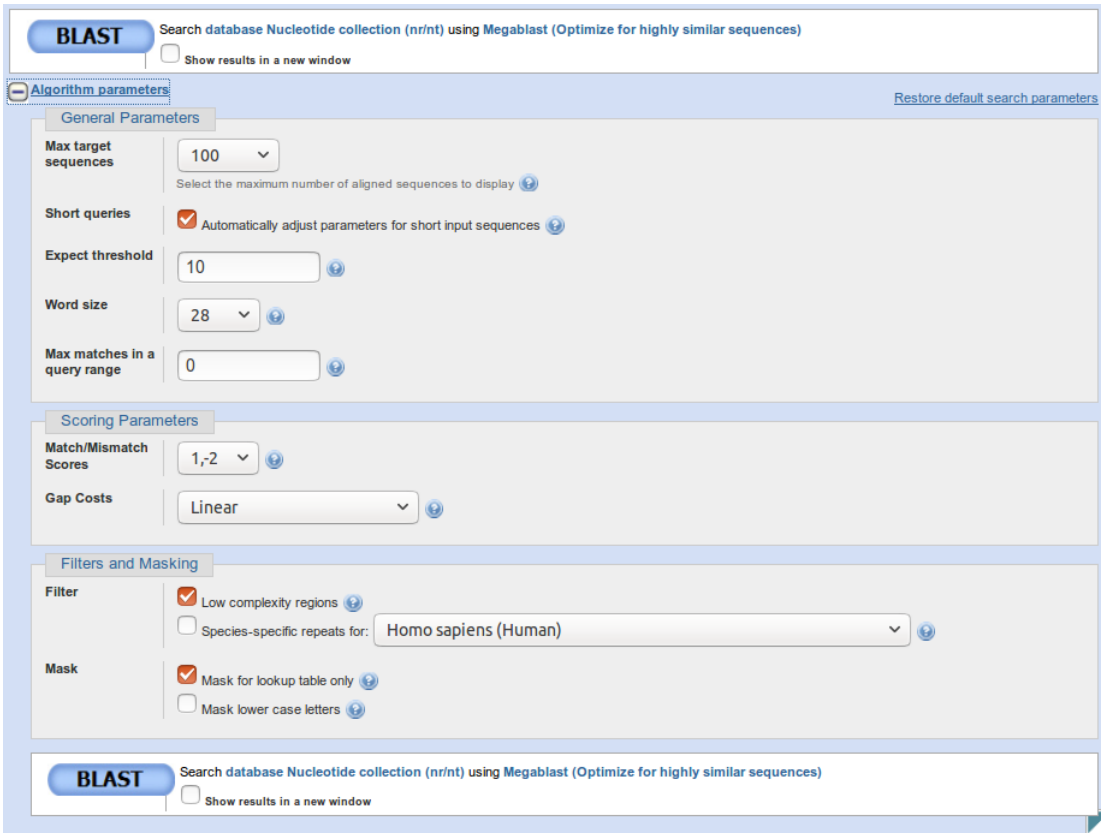


FIGURE 1 – Options avancées - blastn

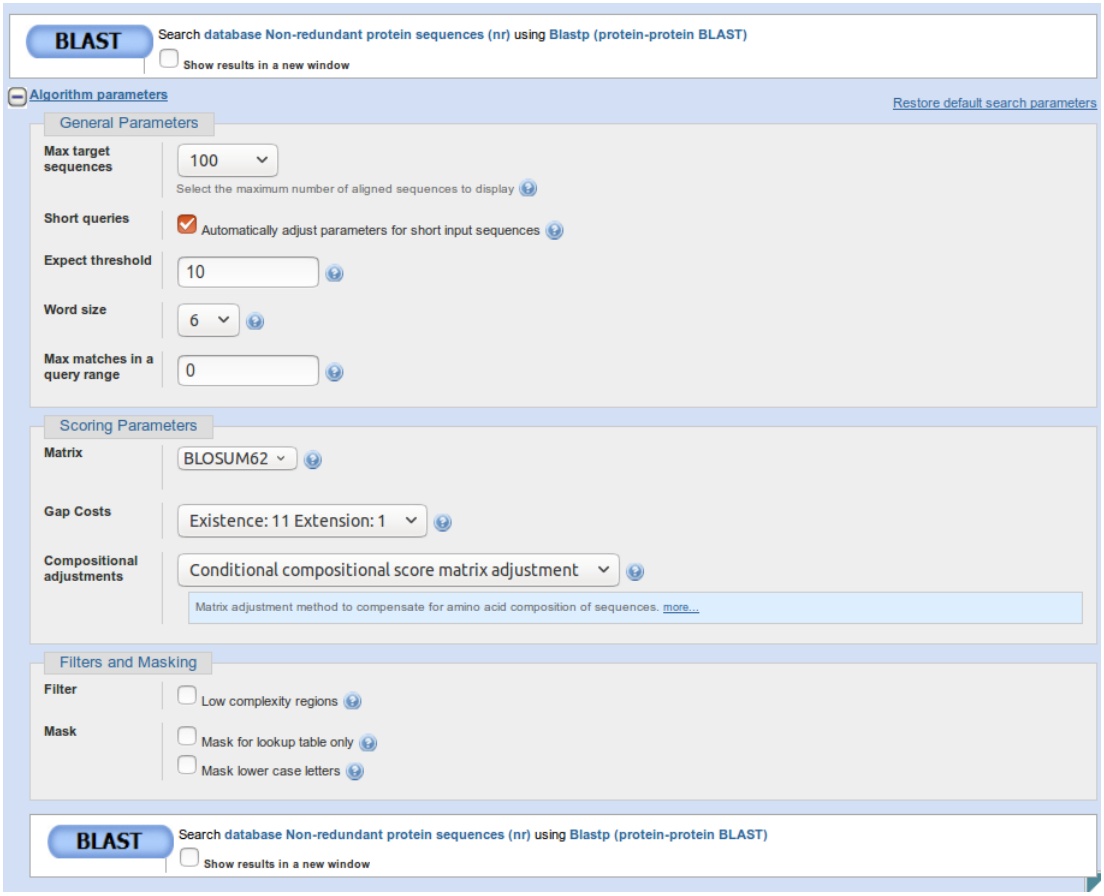


FIGURE 2 – Options avancées - blastp

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.	evalue	expect value
12.	bitscore	bit score

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

Identifiant	Signification
qseqid	Query Seq-id
qgi	Query GI
qacc	Query accesion
qaccver	Query accesion.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
evalue	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
btop	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)
sskingdoms	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