

République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche
Scientifique

Université Mohamed khider de Biskra
Faculté Des Sciences Exacte Et Science De Nature Et De La
Vie Département Des Sciences De La Nature Et De La Vie



Cours de Bio-informatique

**Destiné aux étudiants de 1^{ière} année Master
Sciences biologiques**

2023-2024

Réalisé par M. REBAI Redouane

Table des matières

Liste des figures

Listes des tables

Préface

Application de l'outil informatique sur les techniques génotypiques

A. METHOD BASÉES SUR LA NON AMPLIFICATION DE L'ACIDE NUCLÉIQUE

1. L'analyse du plasmide 1
2. Le ribotypage 6
3. L'électrophorèse sur gel en champ pulsé (PFGE)..... 7
4. L'analyse des enzymes de restriction..... 12

B. METHOD BASÉES SUR L'AMPLIFICATION DE L'ACIDE NUCLÉIQUE

1. La réaction en chaîne par polymérase *in silico* 16
2. L'ADN polymorphe amplifié au hasard 19
3. La technique de MLST..... 24

C. OUTILS DE LA BIO-INFORMATIQUE

1. Banques de données biologiques..... 30
 - 1.1. Les bases de données et les banques de données biologiques..... 31
 - 1.1.1. Les bases de données biologiques..... 31
 - 1.2. Les banques de données 31
 - 1.2.1. Les banques de données nucléiques 31
 - 1.2.2. Les banques de données protéiques 34
 - 1.2.3. Les bases de données spécialisées..... 35
 - 1.2.4. La structure d'une entrée des banques de données..... 36
1. Alignement des séquences biologiques (séquences génomiques et protéiques)..... 42
 - 2.1. Les méthodes d'alignements des séquences nucléiques et protéiques..... 43
 - 2.2.1. Dotplot 43
 - 2.1.2. L'alignement par programmation dynamique (Needleman et Wunsch)..... 45
 - 2.1.3. L'alignement des séquences protéiques 47

2.1.4. Le programme BLAST.....	52
--------------------------------	----

Bibliographie

Liste des figures

Figure	Titre	Page
1	Interface du PlasmidFinder	1
2	Schéma récapitulatif sur la technique de Southern blot	6
3	Interface du programme OligoArchitect	7
4	Principe de l'électrophorèse en champ pulsé	8
5	Interface de plateforme BioNumerics	9
6	Panneau spécifique à la technique PFEG	10
7	Normalisation d'image du gel	10
8	Création de dendrogramme à partir de comparaison des résultats de gel (PFEG).	12
9	Interface de l'outil in silico.ehu.eus	17
10	Interface du Primer3	18
11	Interface du logiciel GelJ.	21
12	L'ajout des informations aux voies de l'image	22
13	Normalisation des voies de comparaison	23
14	Résultats sous forme de Dendrogramme et d'un Image	24
15	Interface du Plateforme autoMLST	25
16	Etapes de MLST selon le mode de placement	25
17	Etapes de MLST selon le mode de novo	27
18	Dendrogramme construit par autoMLST	29
19	L'interface de la banque de données GenBank	32
20	L'interface de banque de données EMBL	33
21	L'interface de la banque de données DDBJ	33
22	Interface de la banque de données protéiques UniProt	35
23	Structure d'une entrée de la banque Nucleotide (GenBank).	37
24	Format FASTA d'une séquence nucléique	38
25	Représentation graphique des points communs entre les séquences selon la méthode de Dotplot	44
26	La matrice PAM 250	49
27	La matrice BLOSUM 62	50
28	Interface du programme BLAST	53

Liste des tables

Tableau	Titre	Page
1	Principaux outils Web pour l'analyse de digestion par restriction	12
2	Récapitulation des résultats de BLAST	54

Préface

La bio-informatique est un domaine multidisciplinaire qui intègre l'informatique, les mathématiques et la biologie, elle est devenue un outil primordial pour analyser, intégrer des données d'origines très diverses et modéliser les systèmes vivants afin de comprendre et prédire leurs comportements (analyse des séquences nucléiques et protéiques, modélisation de l'évolution d'une population animale dans un environnement donné, modélisation moléculaire, reconstruction d'arbres phylogénétiques...).

Grâce à l'avancée spectaculaire de la biologie moléculaire, à l'analyse informatique et au développement de nouvelles méthodes et outils informatiques, ainsi que l'implication de l'intelligence artificielle, la bio-informatique a permis d'intégrer différents domaines, tels que la médecine prédictive en identifiant les risques des maladies génétiques et en élaborant des méthodes de prévention personnalisées. En agriculture, elle a contribué à optimiser les rendements des cultures, à développer des variétés résistantes aux maladies des plantes et à réduire l'utilisation des produits chimiques dangereux. Ainsi que l'étude des microorganismes du sol.

Au bout du compte, on constate que la bio-informatique révolutionne les méthodes d'étude de la vie sous de nombreux aspects.

Le but de ce polycopié est d'offrir un cours étayé en bio-informatique, permettant aux étudiants d'en comprendre les notions de base et constitue pour les étudiants déjà à l'aise avec les rudiments de cette discipline, un recueil d'un ensemble de méthodes permettant de mettre en pratique divers outils informatiques.

Objectifs généraux assignés

1. Familiarisation et apprentissage des notions de base de la bio-informatique telles que les approches *in silico* appliquées aux techniques de génotypage.
2. Acquisition et renforcement des connaissances en matière d'utilisation des banques de données génomiques et protéiques.
3. Maîtrise d'outils élémentaires de statistiques, pour permettre leur application dans l'analyse et la comparaison des séquences biologiques
4. Développement de compétences concernant les différentes techniques et les méthodes expérimentales utilisées pour la caractérisation et l'analyse de l'ADN et des protéines.

Partie I : Méthodes basées sur la non amplification de l'acide nucléique

1. L'analyse du plasmide

Les plasmides bactériens sont des molécules d'ADN extra-chromosomiques circulaires fermées capables de se répliquer de façon autonome. Les plasmides peuvent contenir des gènes pour une variété de traits phénotypiques, tels que la résistance aux antibiotiques, la virulence ou les activités métaboliques, bien que certains plasmides comprennent des gènes ne conférant aucun phénotype détectable. Certains plasmides avoir la capacité de transférer des copies d'eux-mêmes à d'autres souches ou espèces bactériennes.

Avec l'évolution spectaculaire du séquençage du génome entier (WGS) et des données de séquences de plasmides entiers générées par les plates-formes de séquençage à haut débit, il est nécessaire de pouvoir identifier les gènes de résistance aux médicaments et les plasmides à l'aide de données de séquence brutes.

Des outils en ligne sont disponibles tels que PlasmidFinder pour l'identification de plasmides dans les séquences du génome entier en recherchant les séquences de réplicons plasmidiques des espèces *Enterobacteriaceae* et à Gram-positif qui sont organisées dans une base de données. Cet outil, basé sur l'alignement, identifie les plasmides de ces taxons avec une grande précision en indiquant un organisme source sur la base du réplicon le mieux adapté.

Pour utiliser l'outil en ligne PlasmidFinder, l'utilisateur doit saisir l'URL suivant : <https://cge.food.dtu.dk/services/PlasmidFinder/>.

L'interface fournie est conviviale et permet aux utilisateurs d'identifier les plasmides dans des isolats de bactéries séquencés totaux ou partiels.

Center for Genomic Epidemiology

Home Services Publications Contact

PlasmidFinder 2.1

Service Instructions Output Article abstract Citations

Software version: 2.0.1 (2020-07-01)
Database version: (2023-01-18)
[Test sequence](#)

Select database
Gram Positive
Enterobacteriales

Select threshold for minimum % identity
95 %

Select minimum % coverage
60 %

Select type of your reads

The database is curated by:
Henrik Hasman and Alessandra Carattoli
(click to contact)

Figure 1 : Interface du PlasmidFinder.

Plusieurs étapes à suivre pour aboutir à un résultat de recherche des plasmides à partir des bases de données :

Pour commencer à utiliser cet outil, l'utilisateur doit sélectionner la base de données dans laquelle les plasmides seront recherchés. Par exemple, les entérobactéries.

Ensuite, il serait indispensable de fixer un seuil pour le % minimal d'identité, afin de sélectionner le pourcentage minimum de nucléotides identiques entre le gène de résistance le mieux correspondant dans la base de données et la séquence correspondante dans le génome. Dans deuxième temps, l'utilisateur va sélectionner aussi le taux de couverture, qui correspond aux plasmides avec un pourcentage de couverture égal ou supérieur au seuil sélectionné (fig.1). Cette étape est suivie par le choix du type de lecture à l'aide du menu déroulant pour charger le fichier sous format 'FASTA' de séquence annotée dans lequel l'utilisateur souhaite identifier les plasmides.

Après la soumission de la séquence à identifier via le serveur PlamidFinder, ce dernier affichera une sortie similaire à l'exemple ci-dessous :

PlamidFinder-2.0 Server - Results

Organism(s): *Enterobacteriaceae*

Enterobacteriaceae, <i>Aerobacter baumannii</i>						
Plasmid	Identity	Query / Template length	Contig	Position in contig	Note	Accession number
IncFIB(AP001918)	96.84	538 / 682	NODE_151_length_1547_cov_574.472534	1..538		AP001918
IncFII(pRSB107)	97.7	261 / 261	NODE_103_length_1790_cov_579.962585	539..799		AJ851089
IncI1-(Gamma)	97.89	142 / 142	NODE_266_length_500_cov_522.737976	61..202		AP005147

[extended output](#)

Input Files: *resfindertest.fa*

[Results as text](#) [Results tsv](#) [Hits in genome seqs](#) [Plasmid sequences](#)

Les données contenues dans le tableau ci-dessus présentent des informations explicatives des résultats, à savoir :

- Organisme(s) : L'espèce sélectionnée.
- Plasmide : Plasmide contre lequel la séquence d'entrée a été alignée.
- % d'identité : pourcentage d'identité dans l'alignement entre le plasmide le mieux correspondant dans la base de données et la séquence correspondante dans le génome d'entrée (également appelée paire de segments à score élevé (HSP)).

Un alignement parfait est de 100%, mais doit également couvrir toute la longueur du plasmide dans la base de données (comparer les exemples 1 et 3).

- Longueur de requête/HSP : la longueur de requête est la longueur du plasmide le mieux correspondant dans la base de données, tandis que la longueur HSP est la longueur de l'alignement entre le plasmide le mieux correspondant et la séquence correspondante dans le génome (également appelée paire de segments à score élevé (HSP)).
- Contig : Nom du contig dans lequel se trouve le plasmide.
- Position dans le contig : Position de départ du gène trouvé dans le contig.
- Numéro d'accession du plasmide : Numéro d'accession pour le plasmide de référence au sein de la banque de données, Genbank.

D'autres programmes disponibles sur le Web, s'intéressent à l'analyse des marqueurs de diverses souches bactériennes lorsqu'un système de typage appelé le profilage plasmidique. Dans ces procédés, les molécules d'acide désoxyribonucléique plasmidique partiellement purifiées sont séparées selon la taille moléculaire par électrophorèse sur gel d'agarose. Dans une seconde procédure, l'ADN plasmidique qui a été clivé par des endonucléases de restriction (enzymes de restriction) peut être séparé par électrophorèse sur gel d'agarose et le motif de fragments résultant peut être utilisé pour vérifier l'identité des isolats bactériens.

Les séquences cibles où coupent ces enzymes sont dites palindromiques, certaines coupent les deux brins de l'ADN exactement au même endroit en produisant ainsi des bouts francs, les autres coupent chacun des brins de l'ADN à des endroits différents en créant ainsi des bouts cohésifs où l'ADN est sous forme monocaténaire. Les bouts collants peuvent s'hybrider entre elles, permettant de mettre bout à bout deux séquences, créant ainsi un ADN recombinant. Par exemple, l'enzyme *EcoRI* reconnaît la séquence :

5' - GAATTC - 3'

3' - CTTAAG - 5'

Cette enzyme coupe toujours entre les résidus 5'G et A, produisant ainsi des bouts cohésifs :

5' - G AATTC - 3'

3' - CTTAA G - 5'

Une autre enzyme *Hea III*, coupe au point de symétrie pour produire des extrémités franches:

5' - GAA TTC- 3'

3' - CTT AAG- 5'

En effet, ils existent plusieurs sites pour étudier les caractéristiques d'un plasmide et prévoir la taille du/des fragments attendus suite à une restriction enzymatique *in silico*. A titre d'exemple, nous voudrions simuler une restriction enzymatique d'un plasmide en utilisant un logiciel dénommé « **NEBcutter** » qui est un outil simple et rapide pour analyser une molécule d'ADN, qu'elle soit plasmidique ou phagique (il est également possible d'importer sa propre séquence ou une séquence provenant d'une banque de donnée).

L'interface de cet outil apparaît comme suit :

Il suffit d'insérer la séquence de l'ADN plasmidique directement ou introduire le numéro d'accension qui est obtenu après avoir consulté la banque de données nucléique « Nucleotide ». Par exemple, nous allons analyser la séquence du plasmide dénommée, "plasmid pEQ2" (Escherichia coli souche 63743). Son numéro d'accension (ID) fourni par la banque génomique est : « KF362122 ».

Méthodes basées sur la non amplification de l'acide nucléique

Local sequence file: Choisir un fichier Aucun fichier choisi

GenBank number: **KF362122** **1** [Browse GenBank]

or paste in your DNA sequence: (plain or FASTA format)

2

Standard sequences: # Plasmid vectors # Viral + phage

Submit **4**

The sequence is: ☐ Linear ☒ Circular **3**

Enzymes to use: ☒ NEB enzymes ☐ All commercially available specificities ☐ All specificities ☐ All + defined oligonucleotide sequences ☐ Only defined oligonucleotide sequences [define oligos]

Minimum ORF length to display: 100 a.a.

Name of sequence: (optional)

Earlier projects:

Note: Your earlier projects will be deleted 2 days after they were last accessed. You need to have cookies enabled in your browser for this feature to work.

☐ Disable NEBcutter cookies

Delete projects

Démarche :

- 1- Insérer le numéro d'accèsion de la séquence plasmidique.
- 2- Aussi, il y a la possibilité d'insérer la séquence directement.
- 3- Spécifier le type de séquence d'ADN (dans notre cas, il s'agit d'un plasmide, c-à-d un ADN circulaire.
- 4- Il suffit que cliquer sur le bouton « soumettre » pour lancer l'analyse.

Résultat :

Le résultat de la digestion du plasmide apparait dans la page suivante :

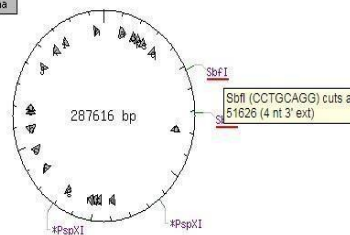
Display: - NEB double cutter restriction enzymes
- Main non-overlapping, min. 100 aa ORFs

GC=47%, AT=53%

ORFs:

a:	1432 aa	k:	794 aa
b:	1329 aa	l:	717 aa
c:	1111 aa	n:	717 aa
d:	1058 aa	n:	706 aa
e:	1011 aa	c:	694 aa
f:	988 aa	p:	679 aa
g:	933 aa	q:	670 aa
h:	893 aa	r:	634 aa
i:	818 aa	s:	618 aa
j:	809 aa	t:	611 aa

Cleavage code	Enzyme name code
blunt end cut	Available from NEB
5' extension	Has other supplier
3' extension	Not commercially available
cuts 1 strand	*: cleavage affected by CpG meth.
	: cleavage affected by other meth.
	(enz.name): ambiguous site



Main options
New DNA
Custom digest
View sequence
ORF summary
Translate GB file
Save project

Availability
All commercial
All

Display
1 cutters
3 cutters
Linear

List
0 cutters
2 cutters
All sites
Save all sites

Minimum ORF length to display: 100 aa OK

2. Le Ribotypage

Le ribotypage est l'une des techniques pour étudier la diversité microbienne. Cette technique implique la digestion de restriction de l'ADN génomique suivie d'un transfert sur la membrane et l'hybridation avec sonde contre les gènes ARNr 16S et 23S spécifiques à l'espèce étudiée. L'ADN digéré est séparé par électrophorèse sur un gel d'agarose et transféré sur une membrane de nylon ou de nitrocellulose. Les fragments d'ADN sur la membrane sont ensuite sondés avec des fragments d'ADN marqués complémentaires des séquences géniques d'ARNr.

L'une des techniques utilisées pour le ribotypage est celle de l'hybridation par Southern blot (fig.2)

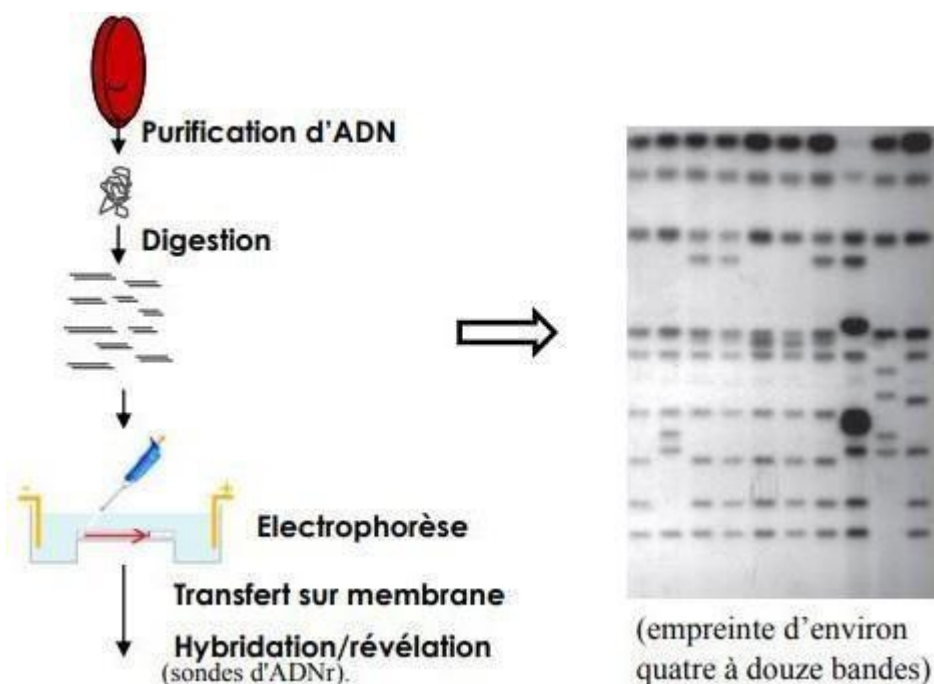


Figure 2 : Schéma récapitulatif sur la technique de Southern blot L'introduction des outils informatiques pour analyser les résultats de techniques du ribotypage devient impératif pour l'identification et la classification des bactéries (les dendrogrammes). En tant que terme, le ribotypage fait référence à l'utilisation de sondes d'acide nucléique reconnaissant les gènes ribosomiaux. En effet, ils existent plusieurs outils pour la conception des sondes oligonucléotidiques qu'ils vont s'hybrider avec l'ADN codant pour l'ARNr.

Le logiciel **OligoArchitect** permet de concevoir différents type sondes oligonucléotidiques, soit pour les puces à ADN ou pour les technique d'hybridation moléculaire.

OligoArchitect™ est un outil sert pour la conception d'amorces et de sondes optimisé par la norme industrielle "Beacon Designer". Ce programme est disponible en ligne, entièrement

gratuit et convivial, il intègre les paramètres des algorithmes les plus récents, permettent de prendre en charge des modèles ayant jusqu'à 10 000 paires de bases, tout en permettant de faire varier de nombreux paramètres tels que le pourcentage en G/C, la température d'hybridation et la longueur d'oligonucléotides ...

Cet outil en ligne est accessible via l'URL suivant :

<http://www.oligoarchitect.com/LoginServlet>

OligoArchitect™ Online

[Technical Help](#)

SYBR® Green I Dual-Labeled Probe Molecular Beacon LightCycler® Probe Scorpions® Probe

Assay

Assay Name :

Description : [Optional]

Sequence

☒ Open sequence from Entrez (Accession/GI number):

☐ Enter nucleotide sequence *Only A, T, G and C bases

SNP Position : Base : Mutant Base :

Figure 3 : Interface du programme OligoArchitect

3. Electrophorèse sur gel en champ pulsé (PFGE)

L'électrophorèse sur gel en champ pulsé (PFGE) est une technique de laboratoire utilisée pour l'obtention d'empreinte moléculaire spécifique d'une souche bactérienne au sein d'une même espèce en fonction du polymorphisme de taille des fragments de restriction du génome.

Le pouvoir résolutif de l'électrophorèse est excellent pour des fragments d'ADN dont la taille est inférieure à 50 000 paires de bases (50 kilobases ou kb), mais, au-delà de cette taille, les fragments d'ADN ne sont plus séparés. L'électrophorèse en champ électrique pulsé permet de séparer des fragments de très grande taille, jusqu'à 10 Mb. Cette technique est basée sur la

digestion de l'ADN par des enzymes de restriction ayant peu de sites de coupure, d'où la génération des fragments d'une taille trop grande (10 à 800 kb). Ainsi les fragments d'ADN générés d'ADN. Ensuite, ces derniers vont subir une migration sur gel en champs pulsé (un champ électrique à orientation variable et qui change régulièrement de direction selon un angle défini) et une analyse bio-informatique des empreintes pour la calibration et la comparaison des empreintes à la base de données.

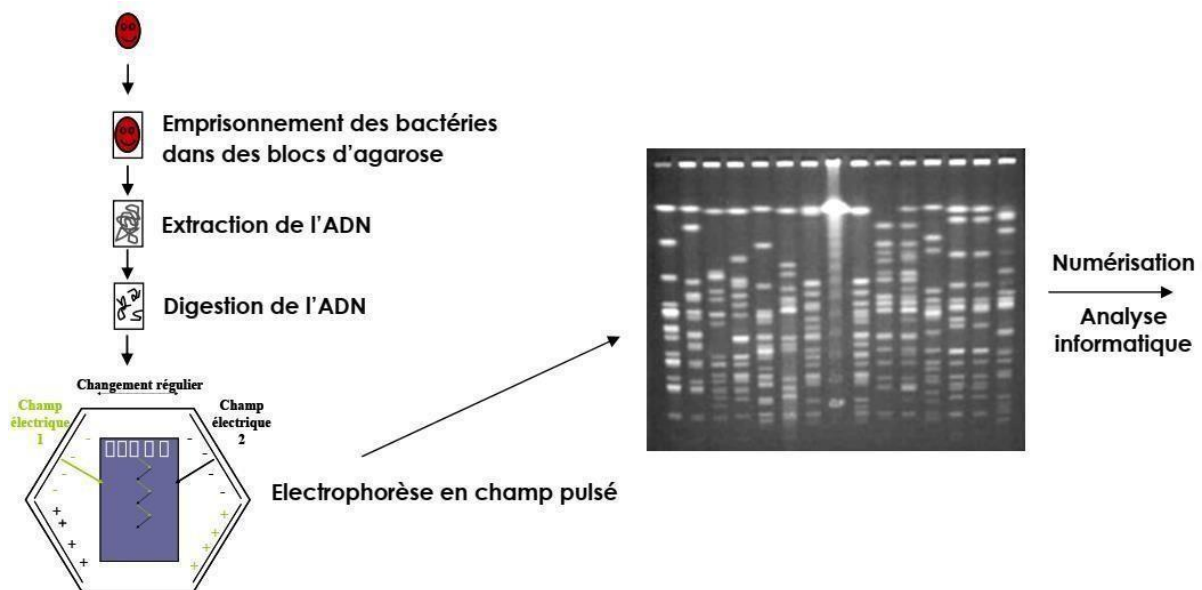


Figure 4 : Principe de l'électrophorèse en champ pulsé

-Analyse par l'outil informatique : parmi les outils le plus utilisé dans le monde entier, surtout en épidémiologie pour analyser le profil électrophorétique issu de la technique de typage moléculaire PFGE, nous citons la plateforme « **Bionumériques** » qui est dotée des outils pour l'analyser des données épidémiologiques et stocker des images de gel dans une seule base de données.

Bionumerics est un logiciel payant, nous pouvons le télécharger en consultant la plateforme de la société bioMérieux. (URL: <https://www.applied-maths.com/applications>).



Figure 5 : Interface de plateforme BioNumerics

Etapes de l'analyse de résultats PFEG : on peut les résumées dans 2 opérations très importantes, la première correspond à l'importation et la normalisation des images de gel après avoir subi une numérisation. La deuxième opération consiste à effectuer des comparaisons entre les différentes souches (selon le profil électrophorétique) pour établir des dendrogrammes.

Démarche :

➤ **Importer et normaliser l'image de gel**

Cette opération se déroule en 4 étapes successives ayant pour but de traiter l'image du gel.

La 1^{ière} étape consiste à crée une base de données pour stocker les données spécifiques aux empreintes moléculaires analyse par la technique PFEG.

Ensuite, il faut désigner le type des résultats et expérimentation à analyser (séquençage, analyse spectrale, empreintes moléculaires...) pour pouvoir importer l'image du gel.

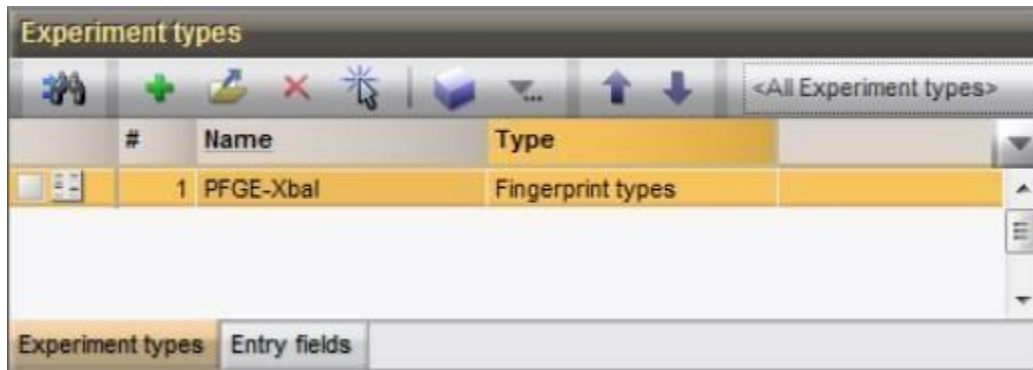


Figure 6 : Panneau spécifique à la technique PFEG (Xbal : l'enzyme de restriction utilisée).

Après avoir importé l'image du gel (format TIFF), cette dernière subira les traitements suivants :

- Localisation des bandes : recadrer l'image pour supprimer l'espace vide et à définir les bandes
- Définition des courbes densitométriques.
- Etablissement d'un système de référence par l'utilisation des poids moléculaires standard pour nommer les positions de référence (fig.6).

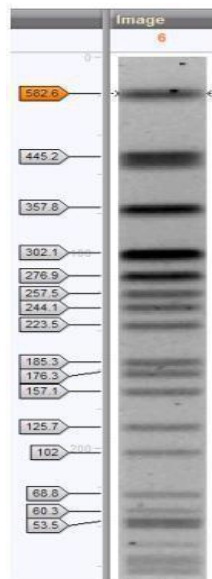
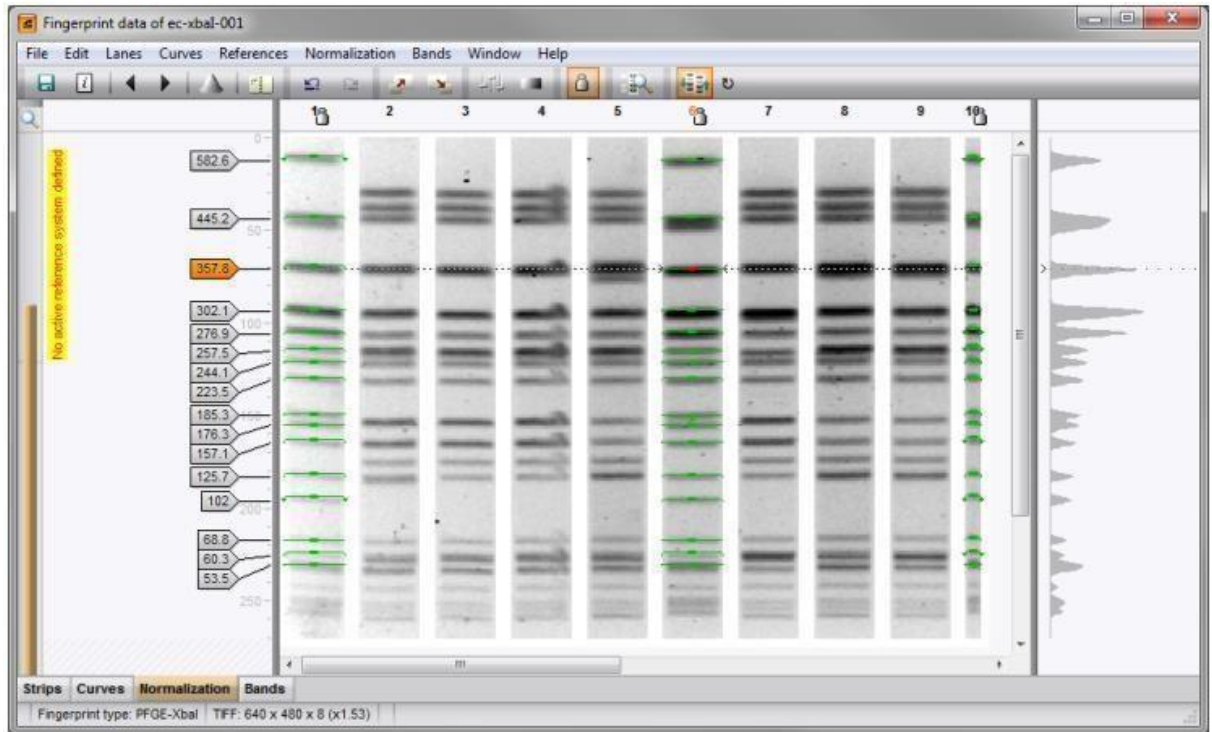
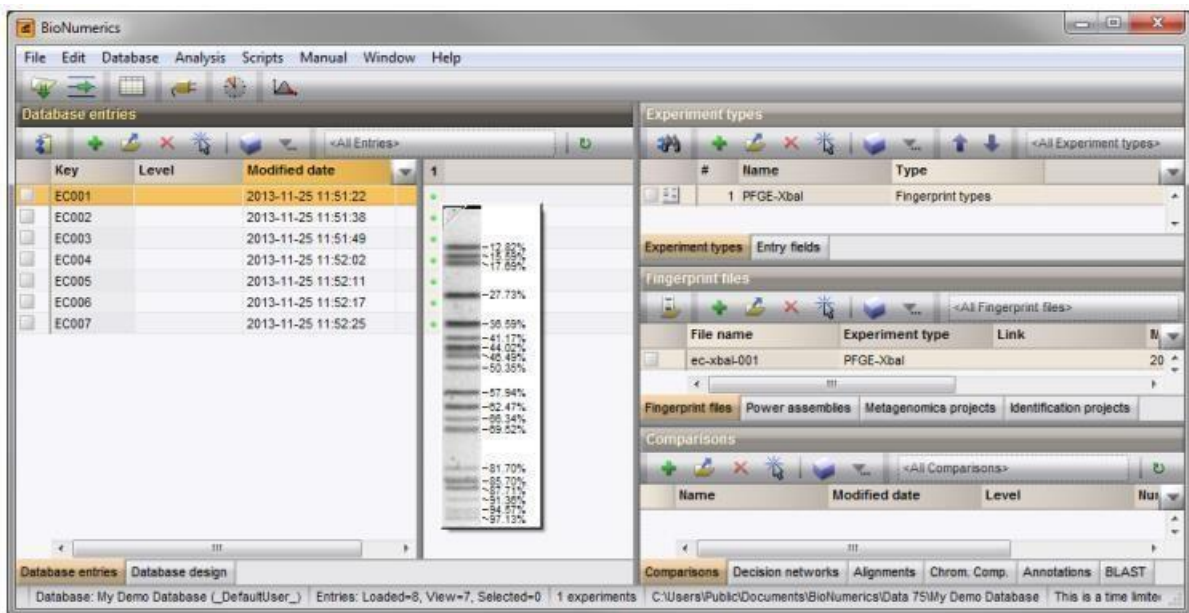


Figure 7 : Normalisation d'image du gel

Méthodes basées sur la non amplification de l'acide nucléique



- Liaison de données d'empreintes moléculaire aux entrées de la base de données que nous avons créée au départ.



➤ Création des dendrogrammes

La création des dendrogrammes nécessite une comparaison des résultats issus de la technique PFEG. Ceci est réalisé par une analyse de cluster qui repose sur l'utilisation d'une matrice de similarité résultante est convertie en un dendrogramme avec un algorithme de clustering.

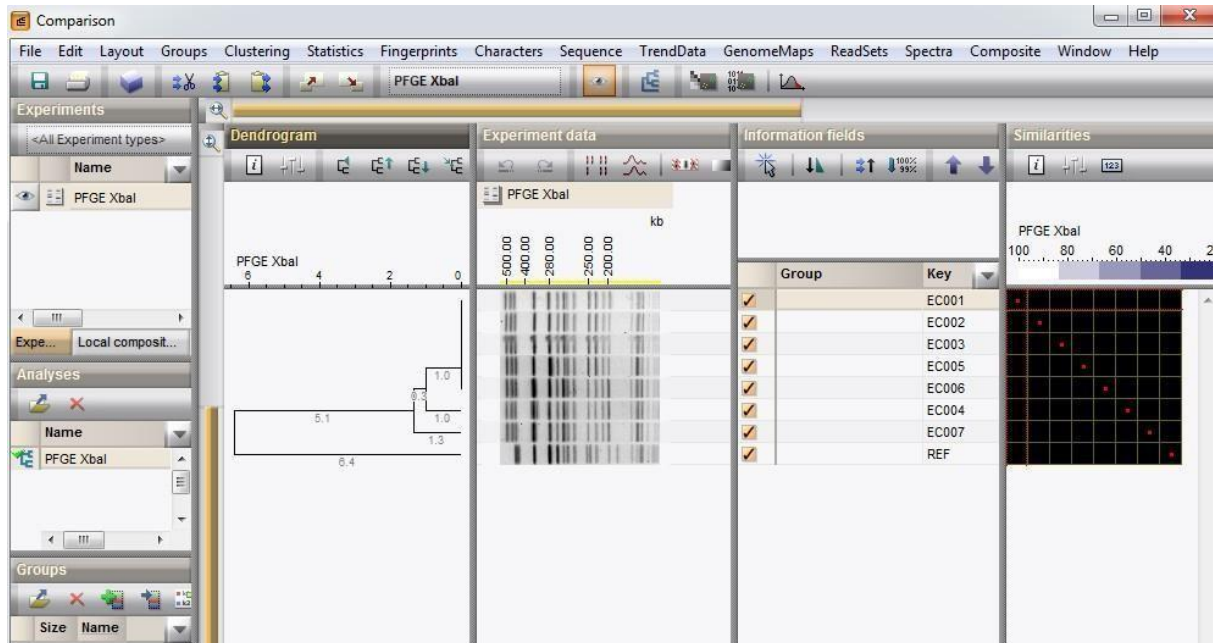


Figure 8 : Création de dendrogramme à partir de comparaison des résultats de gel (PFEG).

4. Analyse des enzymes de restriction

La digestion par restriction est réalisée par incubation de la molécule d'ADN cible avec des enzymes de restriction. Ces dernières reconnaissent et se lient à des séquences d'ADN spécifiques et clivent des nucléotides spécifiques soit à l'intérieur de la séquence de reconnaissance ou à l'extérieur. La digestion par restriction peut produire d'extrémités franches (extrémités d'une molécule d'ADN se terminant par une paire de bases) ou d'extrémités collantes (extrémités d'une molécule d'ADN se terminant par un surplomb de nucléotides).

À l'heure actuelle, il existe une panoplie d'outils qui sont capables d'induire une digestion par restriction *in silico* d'un grand nombre de séquences à la fois, de manière interactive et, en sortie, de produire des séquences de fragments de restriction. Dans ce contexte, différents serveurs ont été développés présentant une interface utilisateur graphique pour l'analyse de digestion par restriction *in silico* de séquences de gènes ou de génomes (Tableau 1).

Tableau 1 : Principaux outils Web pour l'analyse de digestion par restriction

L'outil Web	URL
Webcutter (Max Heiman, U.S.A.).	https://sites.unimi.it/camelot/tools/cut2.html
WatCut (Michael Palmer, University of Waterloo, Canada)	https://diyhl.us/~bryan/irc/protocol-online/protocol-cache/template.php

Restriction Analyzer (Vladimír Cermák, molbiotools.com)	https://molbiotools.com/restrictionanalyzer.php
Restriction Digest.	https://www.geneinfinity.org/sms/sms_redigest.html
In silico restriction digest of complete genomes (University of the Basque Country, Spain)	http://insilico.ehu.es/digest/

Nous prenons le dernier outil web dans le tableau ci-dessus pour donner un exemple sur la digestion de séquence génomique par des endonucléases.

Après avoir saisi l'URL : <http://insilico.ehu.es/digest/>, il y aura l'apparition de l'interface suivante :

***In silico* simulation of molecular biology experiments**

About, Citing this site Last update: 2023/11/29 (2758 prokaryotic genomes)



Experiments against prokaryotic genomes

- [PCR amplification](#)
- [Restriction digest and PFGE](#)
- [PCR-RFLP](#)
- [T-RFLP](#)
- [Double Digestion fingerprinting](#)
- [AFLP-PCR](#)
- [SAMPL](#)
- [SRF](#)
- [DDSL](#)
- [resAP-PCR](#)
- [DNA fingerprinting](#)
- [cDNA-AFLP](#)

[Microsatellite Repeats](#)

[Find ORF by name](#)

[Sort sequence locator](#)

PCR against complete prokaryotic genomes

[pcc.ehu.eus](#)

Experiments against user's sequences

[Main](#)

Online exercises

[Design of PCR and PCR-RFLP experiments](#)

[Counting Chamber](#)

Restriction digest of DNA

[Translate DNA to protein](#)

[Palindromic sequences finder](#)

[Coloured sequences for presentations](#)

[Discriminatory Power Calculator](#)

[Molecular Weight Calculator](#)

[Basic Tm calculation](#)

[RCF / rpm conversion](#)

[Dice + UPGMA analysis of PFGE patterns](#)

[DNA/Protein Alignment \(Smith-Waterman\)](#)

[Experiments against eukaryotic genomes](#)

[Multiple Sequence Alignment \(ClustalW\)](#)

[GScompare.ehu.eus](#)

Recommended sites:

[GScompare.ehu.eus](#)

[Biophp.org](#)

Pour procéder à la digestion *in silico*, l'utilisateur doit choisir sur 'Restriction of digest of DNA', puis insérer la séquence à analyser sous format 'FASTA'

Une fois la séquence a été saisie.

Restriction enzyme digest of DNA

with commercially available restriction enzymes

Beta version: correct behaviour in being checked

Paste the sequence in the textbox as **plain text**, [fasta](#) or [upload sequence](#).

```
>X06534.1 E. coli plasmid Ent P307 oriT region
AGATCTCATTTATAAACATCAGGCAGATGGCTAACATCCATTTTTTCATTTTTCCACCTCTGGTGACTTT
ATCCGTAAATAATTTAACCCACTCCACAAAAGGCTCAACAGTTGGTGGTTCTCACCACCAAAAGCACC
ACACCCTACGCAAAAACAAGTTTTTGCTGATTTGATTTTAATATCATGTGCTTATATTCATGAATTTATA
TTATTTAAATCAGATTTATTTAAACAGCGGTGCAGGCGCGGCTATGGCACCGTGTCTGCACCGCTTTGT
AGGGGTGGTGGTCTGACTATTTTTATAAAAAACATTGTTTTATATTAGGGGGGGGCTGCTAGCGGCGCGGTGT
GTTTTTTTATAGGATACCGCCAGGGGCGCTGCTAGCGGTGCGTTCCTGTTTTTCATTGTGAATTTAGTGT
```

[Tidy Up](#) [Reverse](#) [Complement](#)

Display of Results: ☒ Table ☐ Tabulated text

☒ Show sequence(s) on top

Settings: Default [Change](#)

Select specific endonucleases: [Select](#)

Le clic sur "Get list of restriction enzymes" permet l'obtention d'un tableau récapitulatif, comportant les différentes enzymes de restriction qui pourraient couper la séquence cible, avec le nombre et la position de la coupure.

Remarque : pour avoir le site et la position exacte de coupure, l'utilisateur doit cliquer sur le chiffre situé dans la colonne "Position" dans l'interface présentée dans la capture ci-dessous.

Méthodes basées sur la non amplification de l'acide nucléique

>X06534.1 E. coli plasmid Ent P307 oriT region

Length of code: 536

G+C=38%

```
AGATCTCATT TATAACATC AGGCAGATGG CTAACATCCA TTTTTCATT TTTCCACCTC TGGTGACTTT ATCCGTAAAT AATTTAACCC ACTCCACAAA 100
AAGGCTCAAC AGGTTGGTGG TTCTCACCAC CAAAAGCACC ACACCTACG CAAAAACAAG TTTTGTCTGA TTGATTTTA ATATCATGTG CTTATATTCA 200
TGAATTTATA TTATTTAAAT CAGATTTATT TAAACAGCG GTGCAGGCGC GGCTATGGCA CCGTGTCTGC ACCGCTTTGT AGGGGTGGTG CTGACTATTT 300
TTATAAAAAA CATTGTTTTA TATTAGGGGG GGCTGCTAGC GGCCTGGTGT GTTTTTTTAT AGGATACCGC CAGGGGCGCT GCTAGCGGTG CGTTCTCTGT 400
TTCATTGTGA ATTTTAGTGT TTCGAAATTA ACTTTGTTTT ATGTTTAAAA AAGGTAATCT CTAATGGCTA AGGTGAACCT GTATATCAGC AATGATGCTT 500
ATGAAAAAAT AAATGCGATT ATTGAAAAAC GTCGAC
```

Restriction enzyme	Cuts	Positions
AccBII , BanI , BshNI , BspT107I G [^] GYRC_C	1	257
AccI , FblI , XmiI GT [^] MK_AC	1	532
AccII , Bsh1236I , BspFNI , BstFNI , BstUI , MvnI CG [^] CG	2	249 344
AciI , BspACI , SsiI C [^] CG_C or G [^] CG_G	7	238 249 272 339 344 367 385
AcsI , ApoI , XapI R [^] AATT_Y	2	202 409

Partie II : Méthodes basées sur l'amplification de l'acide nucléique

La réaction en chaîne par polymérase (PCR)

La réaction en chaîne par polymérase (PCR) est une technique *in vitro* courante utilisée pour réaliser de nombreuses copies d'une région particulière de l'ADN. Cette région d'ADN peut être tout ce qui intéresse l'expérimentateur. Il peut s'agir d'un gène ou d'un marqueur génétique qu'on souhaite comprendre la fonction.

En règle générale, l'objectif de la PCR est de produire suffisamment de région d'ADN cible pour qu'elle puisse être analysée ou utilisée d'une autre manière. Par exemple, l'ADN amplifié par PCR peut être envoyé pour séquençage, visualisé par électrophorèse sur gel ou cloné dans un plasmide pour des expériences ultérieures.

La PCR est utilisée dans de nombreux domaines de la biologie et de la médecine, notamment la recherche en biologie moléculaire, le diagnostic médical et même certaines branches de l'écologie.

Actuellement, l'utilisation des outils bio-informatiques devient indispensable pour de nombreuses techniques de biologie moléculaire et de génie génétique.

1. La réaction en chaîne par polymérase *in silico*

L'objectif de la PCR *in silico* est de fournir une méthode simple pour obtenir des résultats théoriques prometteurs de PCR à partir de l'ADN (génomique bactérienne, des plasmides facultatifs lorsqu'ils sont disponibles).

Il existe plusieurs outils en Web qui permettent de réaliser une réaction de PCR *in silico*. Par exemple nous citons le site insilico.ehu.eus. Ce dernier a été développé par une équipe de chercheur de l'université de du Pays Basque.

***In silico* simulation of molecular biology experiments**[About, Citing this site](#)

Last update: 2023/11/29 (2758 prokaryotic genomes)



insilico.ehu.eus

Experiments against prokaryotic genomes PCR amplification Restriction digest and PFGE PCR-RFLP T-RFLP Double Digestion fingerprinting AFLP-PCR SAMPL SRF DDSL resAP-PCR DNA fingerprinting cDNA-AFLP	Microsatellite Repeats Find ORF by name Sort sequence locator PCR against complete prokaryotic genomes pcr.ehu.eus Experiments against user's sequences Main Online exercises Design of PCR and PCR-RFLP experiments Counting Chamber	Restriction digest of DNA Translate DNA to protein Palindromic sequences finder Coloured sequences for presentations Discriminatory Power Calculator Molecular Weight Calculator Basic Tm calculation RCF / rpm conversion Dice + UPGMA analysis of PFGE patterns DNA/Protein Alignment (Smith-Waterman) Experiments against eukaryotic genomes Multiple Sequence Alignment (ClustalW) GScompare.ehu.eus Recommended sites: GScompare.ehu.eus Biophp.org
---	---	--

Figure 9 : Interface de l'outil insilico.ehu.eus

Ce site Web dispose de nombreux applications et logiciels qui aident à effectuer des analyses exhaustives sur le génome procaryote à savoir, PCR-RLFP, ALFP-PCR, recherche des ORFs, des alignement multiples...

La réaction *in silico* de PCR nécessite certains éléments comme le couple d'amorces et surtout l'ADN matrice :

- **La matrice d'ADN** : correspond à l'ADN bactérien séquencé sélectionné dans le formulaire et le plasmide peut être inclus dans l'expérience s'il est disponible.

Sélectionnez le genre sur la page principale du PCR et l'espèce ou la souche sur le formulaire correspondant.

Dans l'expérience, l'ADN chromosomique bactérien sera considéré comme circulaire à moins qu'il ne soit linéaire (pe *Agrobacterium tumefaciens* , chromosome linéaire). Les plasmides seront toujours considérés comme circulaires.

- **Les amorces** : au cours des essais *in silico*, deux amorces peuvent être utilisées, doivent être introduites dans le sens 5'-3', et l'amplification théorique considérera toutes les combinaisons possibles qui permettent l'amplification (amorce 1 vers amorce 1, amorce 2 vers amorce 2, amorce 1 vers amorce 2 et amorce 2 vers amorce 1).

Dans tous les cas, on considérera que les amorces ont été conçues correctement, de sorte qu'elles ne formeront pas de dimères, d'épingles à cheveux ou tout autre résultat aberrant.

Les longueurs d'amplicons peuvent atteindre 10 000 pb et peuvent être personnalisées. Nous avons supposé 3 000 pb comme valeur par défaut. La longueur sélectionnée comprend les amorces.

Les amorces utilisées dans cette réaction de PCR ont été conçues par un autre outil disponible et accessible via le net, Primer3 (<https://primer3.ut.ee/>).

Primer3web version 4.1.0 - Pick primers from a DNA sequence.

Select the Task for primer selection: **generic**

Template masking before primer design (available species)

Select species: Example: Mus musculus

Nucleotides to mask in 5' direction: 1

Primer failure rate cutoff: < 0.1

Nucleotides to mask in 3' direction: 0

Paste source sequence below (5' to 3', string of ACGTNacln -- other letters treated as N -- numbers and blanks ignored). FASTA format ok. Please N-out undesirable sequence (vector, ALUs, LINEs, etc.) or use a [Mispriming Library \(repeat library\)](#).

(NONE)

☒ Pick left primer, or use left primer below ☐ Pick hybridization probe (internal oligo), or use oligo below ☒ Pick right primer, or use right primer below (5' to 3' on opposite strand)

Pick Primers | Download Settings | Reset Form

Sequence Id: A string to identify your output.

Targets: E.g. 50,2 requires primers to surround the 2 bases at positions 50 and 51. Or mark the source sequence with [and]: e.g. ...ATCT[CCCC]TCAT.. means that primers must flank the central CCCC.

Overlap Junction List: E.g. 27 requires one primer to overlap the junction between positions 27 and 28. Or mark the source sequence with - : e.g. ...ATCTAC-TGTCAT.. means that primers must overlap the junction between the C and T.

Excluded Regions: E.g. 401,7 68,3 forbids selection of primers in the 7 bases starting at 401 and the 3 bases at 68. Or mark the source sequence with < and >: e.g. ...ATCT<CCCC>TCAT.. forbids primers in the central CCCC.

Figure 10 : Interface du Primer3

En effet, le choix d'amorces conditionne le bon déroulement de réaction de PCR, la raison pour laquelle, le couple d'amorces sélectionné pour la réaction doit remplir certains critères.

La définition et la synthèse d'amorces nécessitent généralement de connaître la séquence du fragment que l'on souhaite amplifier.

A partir des séquences d'amorces disponibles, il est nécessaire de déterminer la séquence nucléotidique à partir de laquelle la polymérase peut synthétiser un nouveau brin d'ADN.

Les règles de base pour choisir un pair d'amorces sont :

- Les 2 amorces doivent être complémentaires à un brin d'ADN existant.
- Dans la direction « correcte » pour permettre la synthèse d'ADN de 5' à 3'.
- Les températures d'appariement des deux oligonucléotides doivent être similaires ou identiques.
- Contiennent d'environ 50% de GC.

La longueur des amorces utilisées pour la PCR est généralement de 17 à 25 pb, selon l'application définie.

Les amorces sont obtenues par synthèse chimique à la demande, c'est-à-dire dans un ordre déterminé par l'expérimentateur.

2. L'ADN polymorphe amplifié au hasard (Random Amplified Polymorphic DNA)

La technique de RAPD est utilisée le plus souvent pour étudier la relation phylogénétique entre les espèces animales et végétales.

Il s'agit de l'une des technologies les plus polyvalentes depuis son développement en 1990.

Il est très pratique et rapide, nécessite très peu d'ADN qui n'a pas besoin d'être très pur, ne nécessite aucune connaissance des informations de séquence antérieures et de nombreux organismes peuvent être analysés rapidement et facilement distingué en même temps.

L'ADN polymorphe amplifié aléatoirement (RAPD) est une technique basée sur la PCR qui utilise des amorces arbitraires pour amplifier l'ADN en se liant à des sites non spécifiques de l'ADN. Lorsque ces fragments amplifiés sont transférés sur un gel d'agarose, des différences dans les modèles de bandes sont observées.

RAPD est une technique au cours de laquelle est un fragment d'ADN amplifié par PCR à l'aide d'amorces synthétiques courtes (généralement 10 pb) avec des séquences aléatoires.

Ces oligonucléotides fonctionnent à la fois comme amorces directes et inverses et peuvent généralement amplifier simultanément des fragments de 1 à 10 sites génomiques.

Les fragments amplifiés, généralement de taille 0,5 à 5 kb, sont séparés par électrophorèse dans des gels d'agarose et des polymorphismes, par exemple la présence ou l'absence de bandes d'une taille particulière, sont détectés après coloration au bromure d'Ethidium.

Bien que l'on pense que ces polymorphismes sont principalement dus à des variations dans les sites d'hybridation des amorces, ils peuvent également résulter de différences dans la longueur des séquences amplifiées entre les sites d'hybridation.

Le principal avantage du RAPD est que le test est rapide et simple.

Puisqu'il s'agit d'une PCR, seule une petite quantité d'ADN matrice est requise.

Des amorces aléatoires sont disponibles dans le commerce, donc aucune donnée de séquence n'est requise pour la conception des amorces.

Les marqueurs moléculaires basés sur le polymorphisme de l'ADN ont montré une grande aptitude à décrire la variabilité génétique ainsi que sa répartition au sein des populations grâce à l'utilisation des outils avancés de bio-informatique qui peuvent jouer un rôle important dans la découverte et l'analyse des marqueurs moléculaires. Parmi les outils informatiques utilisés pour identifier différents marqueurs génétiques, nous abordons ici ceux qui servent à analyser les données des images de gel et comparent les modèles d'ADN résultant d'expériences avec des marqueurs génétiques moléculaires et génère également des arbres phylogénétiques

GelJ un des logiciels les plus utilisés par les chercheurs et les étudiants travaillant en biologie moléculaire et en génétique car il est gratuit, open source, facile à utiliser, possède une interface utilisateur graphique conviviale.

GelJ est une application pour l'analyse d'empreintes génétiques sur gel 1D. Il peut être exécuté sous Windows, GNU/Linux et Mac OS X. La seule exigence est l'installation de Java 7 ou supérieur. Il n'y a aucune exigence matérielle particulière pour exécuter GelJ.

Avant de commencer à expliquer les fonctionnalités de GelJ, nous devons préciser trois notions de base pour procéder à l'analyse de données par ce logiciel : Etude, Expérimentation, Comparaison.

- a. L'étude : contient toutes les informations nécessaires pour estimer les relations entre les échantillons issus d'expériences biologiques. Une étude se compose d'entrées correspondant aux organismes individuels étudiés. Chaque étude est identifiée par un identifiant unique et des informations complémentaires attribuées par l'utilisateur.
- b. L'expérience : représente des informations numériques liées aux expériences biologiques réalisées pour estimer les relations entre les échantillons.

L'expérience vient toujours de l'étude. L'ajout d'une expérience à votre étude implique plusieurs étapes (voir la section 'Analyse d'image').

- c. Une comparaison permet à l'utilisateur de comparer les pistes d'expérimentations. Une comparaison permet à l'utilisateur de calculer et afficher des dendrogrammes et générer des matrices de similarité. Une comparaison vit toujours d'une étude, et les voies de cette comparaison appartiennent à l'expérience d'une telle étude.

L'unité d'information principale de GelJ s'appelle une étude. Une étude contiendra des expériences et des comparaisons. La fenêtre principale de GelJ permet à l'utilisateur de gérer ses études.

Au lancement initial du logiciel, la majorité des fonctionnalités demeurent inactives lorsqu'une fenêtre GelJ est ouverte et ne s'activent qu'à leur utilisation. La seule option accessible à ce moment-là est celle de créer une nouvelle étude.

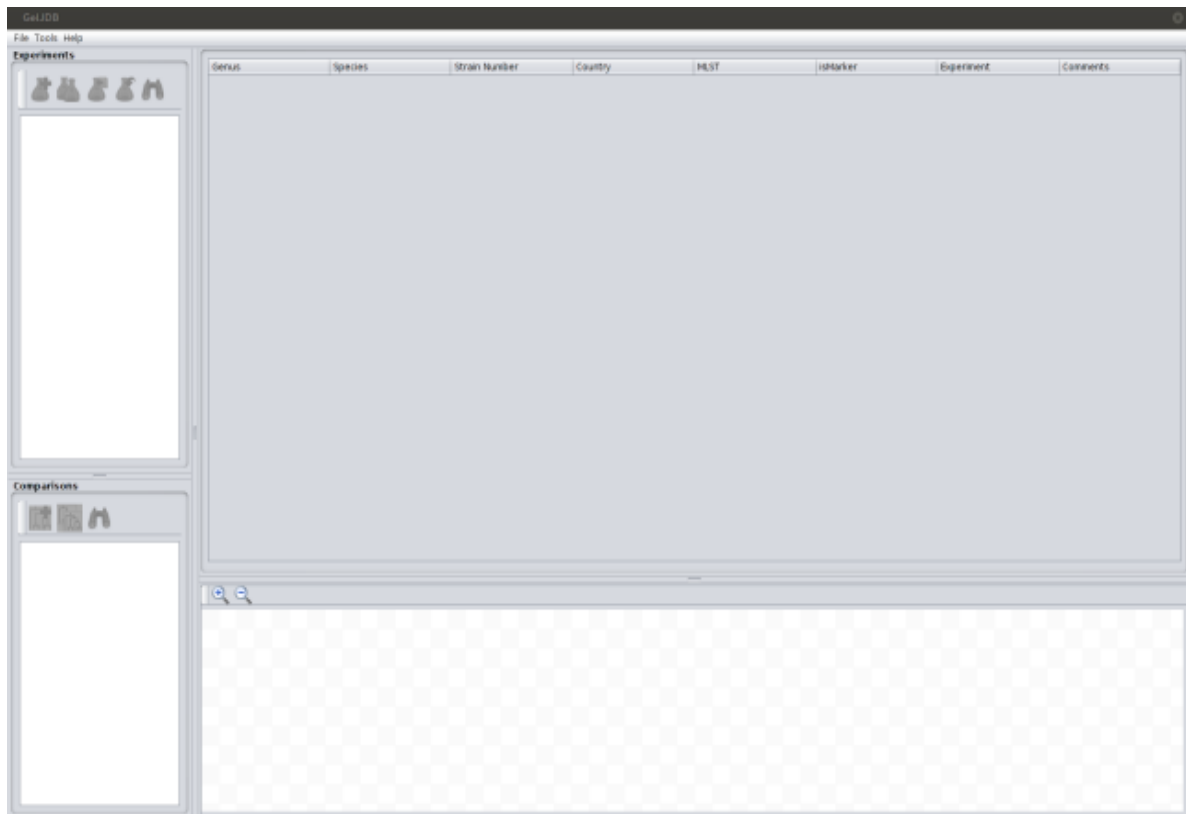


Figure 11 : Interface du logiciel GelJ.

La première étape dans l'analyse d'images de sélectionner l'image à analyser.

Une fois l'image ouverte, il y a 4 étapes principales pour analyser l'image : le prétraitement, la détection de trajectoire, la détection de bande et la normalisation.

- **Prétraitement** : Dans cette étape, l'utilisateur peut prétraiter l'image avec différentes options.

- **Détection de trajectoire :**

Dans un premier temps, l'utilisateur doit indiquer si le fond de l'image est sombre ou clair. Ces informations seront utilisées pour détecter automatiquement les voies de l'image. Ensuite, L'utilisateur peut ajuster la position, l'épaisseur et la forme des voies en sélectionnant simplement une voie puis en interagissant avec elle dans l'image.

- **Détection de bandes :**

Cette étape permet à l'utilisateur de sélectionner une bande d'image. Depuis cette fenêtre, l'utilisateur peut fixer un seuil pour chaque voie de l'image. Cela fournit un meilleur ajustement plus fin pour localiser automatiquement les bandes que le seuil global.

- **Normalisation :** Dans cette étape, l'utilisateur peut normaliser l'image à l'aide des trajectoires de référence.

Une fois qu'une expérience est créée ou importée, elle est ajoutée au panneau « Expériences » du fenêtre principale de GelJ.

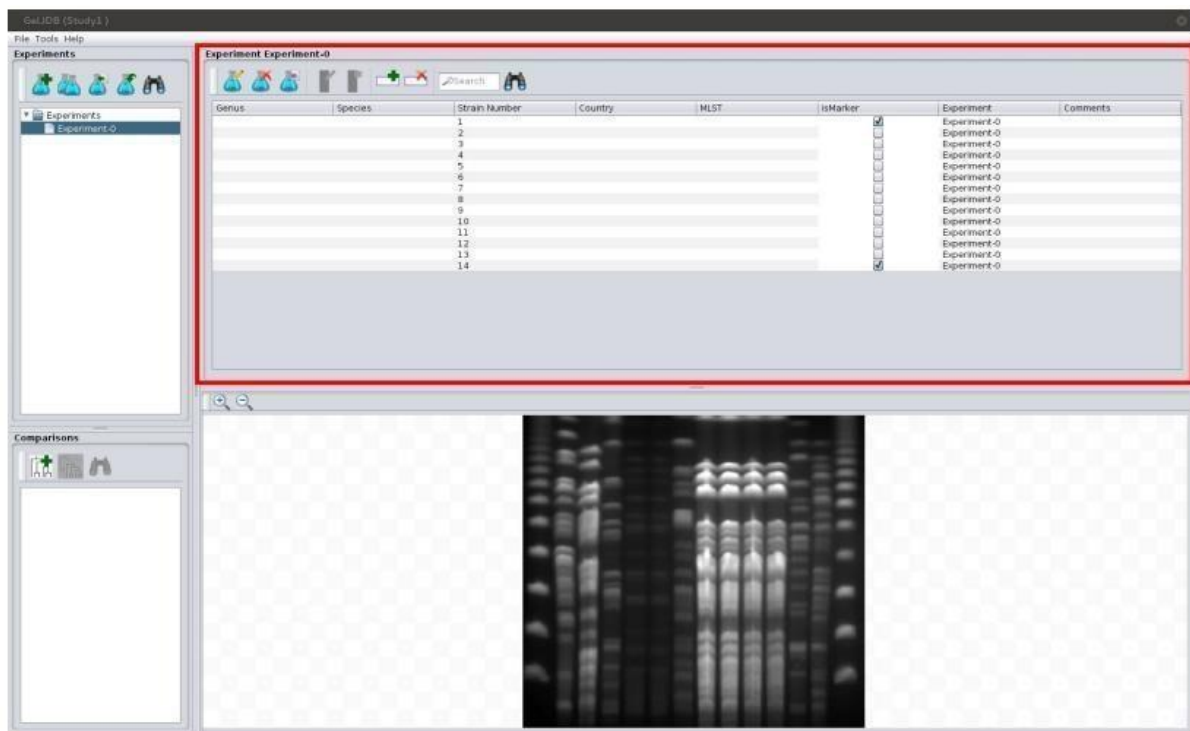


Figure 12 : L'ajout des informations aux voies de l'image.

Dans la fenêtre principale, les utilisateurs peuvent ajouter des informations sur chaque voie de l'expérience par la sélection de l'un des chemins (les images associées à l'expérience seront affichées dans le panneau inférieur droit, et dans ces images, la voie sélectionné sera mis en

surbrillance pour aider l'utilisateur à identifier le chemin), puis la saisie des informations telles que dans une feuille de calcul.

La recherche de similarités entre les différentes bandes lorsque ce bouton est enfoncé, la fenêtre suivante apparaît :



Cette option permet à l'utilisateur de rechercher des voies similaires à la voie sélectionnée.

- **Comparaison** : une fois qu'une expérience est ajoutée à l'étude, le panneau de comparaison est activé et la fonctionnalité permettant de gérer les comparaisons est activée, ce qui permet à l'utilisateur de créer une nouvelle comparaison générant comme résultat un dendrogramme. Lorsque cette option est sélectionnée, l'assistant « New comparaison » apparaît et guide l'utilisateur dans la création de la comparaison.

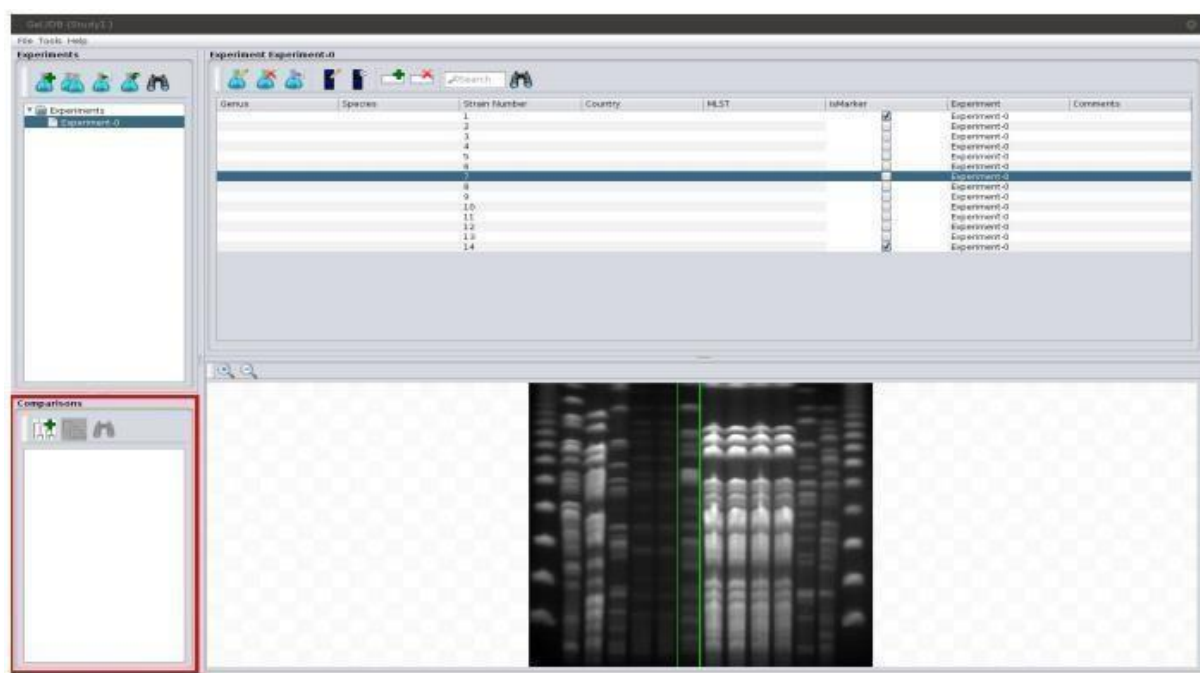


Figure 13 : Normalisation des voies de comparaison

Au cours de cette étape, l'utilisateur peut sélectionner les expériences contenant les voies qui seront incluses dans la comparaison. L'utilisateur peut choisir si les marqueurs des expériences sont inclus et sélectionner les voies qui seront incluses dans la comparaison.

Enfin, l'utilisateur peut configurer la manière dont le dendrogramme est généré, choisissant ainsi une méthode de construction convenable :

- **Méthode de similarité**

-Méthodes basées sur les bandes : Dice, Jaccard, Ochiai et Jeffrey's X.

-Méthodes basées sur les courbes : coefficient de Pearson, corrélation cosinus, distance euclidienne et distance de Manhattan.

- **Liaison (linkage)** : UPGMA, UPGMC, liaison simple, Ward.

Une fois toutes ces options configurées, un dendrogramme est généré.

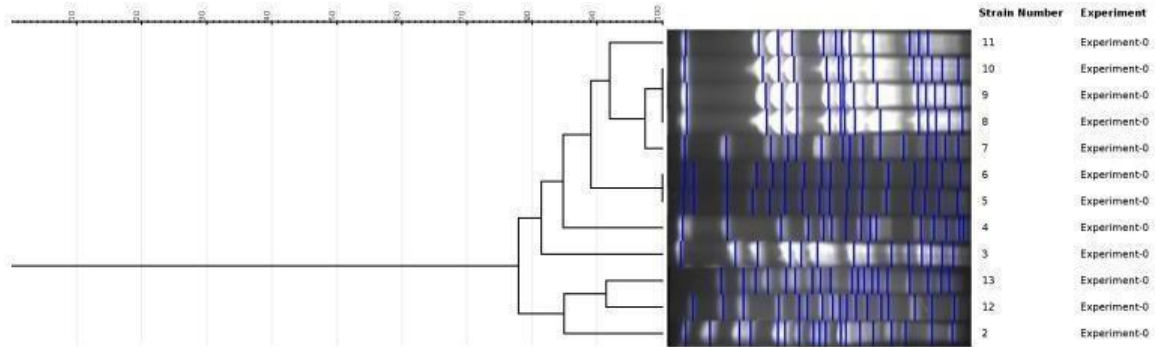


Figure 14 : Résultats sous forme de Dendrogramme et d'un Image.

3. La technique de MLST (Multi-Locus Sequence Typing)

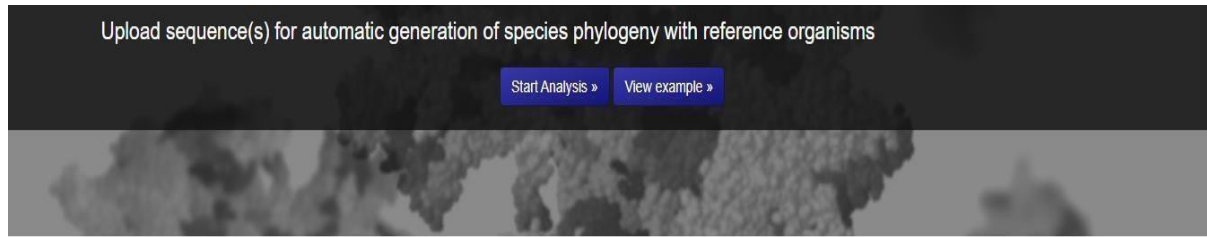
Comprendre le contexte évolutif des isolats bactériens a des applications dans un large éventail d'études. Cependant, établir une phylogénie précise des espèces reste un défi.

L'utilisation de l'ADNr 16S pour l'identification des espèces reste très populaire malgré sa faible résolution au niveau des espèces en raison de la conservation élevée des séquences.

La méthode est basée sur le séquençage d'un ensemble de fragments d'ADN, amplifiés par PCR, provenant d'autant de gènes de ménage qui codent pour des protéines essentielles chez les bactéries. Ces séquences se caractérisent par leur stabilité dans le temps et par leur polymorphisme suffisant pour distinguer les souches les unes des autres.

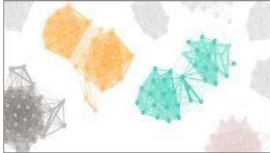
Cependant, l'analyse de séquences multilocus (MLST) par typage est entravée par les étapes manuelles intensives en aval du traitement des fichiers de données de séquence brutes d'où la nécessité d'utiliser des outils informatiques pour l'édition et l'analyse des séquences d'ADN afin de réduire considérablement le temps nécessaire au traitement des données.

La plateforme autoMLST a été développée pour fournir un outil rapide en un clic pour simplifier ce flux de travail sur : <https://automl.st.ziemertlab.com>. Il permet de réaliser des analyses phylogénétiques rigoureuses plus rapides.



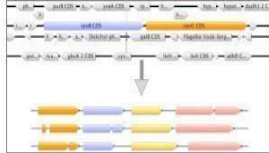
Overview

Rapid ANI estimation



Nearest reference organisms (NCBI RefSeq) are identified and collated

MLST gene selection



Single copy genes with dN/dS values < 1 are identified and prioritized

Tree building



Phylogeny is built using all MLST genes from query and reference organisms using either a concatenated gene or coalescent tree method.

Figure 15 : Interface du Plateforme autoMLST

La première étape pour réaliser une analyse par l'outil autoMLST, l'utilisateur doit choisir le flux du travail selon deux modes :

- A. Le mode placement : les séquences soumises par l'utilisateur sont automatiquement ajoutées à l'arbre de référence sélectionné par l'utilisateur.

Select reference genus:

Auto

Sequences

Accession numbers and uploaded files

Remove selected sequences

Upload Genomes

NCBI Accession Number

Enter a NCBI accession number

+ Upload

Sequence(s) (one genome per file)

Choose Files No file chosen

+ Upload

Options

Enter email address (optional)

(Optional)

Enter job title (optional)

(Optional)

Save links

☐ Save link in cookies (viewable on results page under "last results")

Submit Job »

Figure 16 : Etapes de MLST selon le mode de placement

1. Comme illustré dans la figure 8, l'utilisateur doit Sélectionnez l'arbre de référence auquel les séquences téléchargées seront ajoutées. Par défaut, autoMLST tentera de détecter lui-même le genre le plus proche.
 2. Récupérez la séquence sur le portail NCBI en fonction de son numéro d'accession, puis téléchargez un ou plusieurs fichiers de séquence dans l'un des formats acceptés, c'est-à-dire le format FASTA ou Genbank, et chaque fichier doit contenir le génome d'une seule espèce.
 3. Un maximum de 50 séquences peut être ajouté à l'arbre.
 4. Plusieurs options supplémentaires pour enregistrer les résultats des tâches et être informé de l'état des tâches sont disponibles, telle que la saisie de l'email et l'attribution d'un titre au travail.
- B. Le mode de novo : l'utilisateur peut choisir subtilement des espèces supplémentaires pour construire l'arbre et sélectionner les gènes à utiliser pour le MLST.

Dans ce mode, l'utilisateur doit ignorer la sélection manuelle des organismes pour la construction de l'arbre (étape 2) ou des gènes à aligner (étape 3) et procéder automatiquement avec les 50 meilleurs organismes ou les 100 meilleurs gènes lorsqu'ils sont trouvés, ensuite, la sélection de l'option ModelFinder a permis de trouver le modèle optimal pour la construction d'arbres.

La soumission des séquences pour construire l'arbre se fait via la récupération d'une séquence par son numéro d'accès au niveau de NCBI ou le téléchargement un ou plusieurs fichiers de séquence. Sous un format acceptable (FASTA, Genbank, EMBL)

Par défaut, autoMLST sélectionne toutes les séquences soumises par l'utilisateur et les organismes les plus proches qu'il détecte pour continuer, jusqu'à un total de 50. Les utilisateurs peuvent ajouter des organismes de la liste d'organismes à cette sélection ou les supprimer.

Step 1 Enter sequences » Step 2 » Step 3 » Results

Advanced Mode: After uploading your sequences, choose from which species to construct the tree for them, and choose the genes to use for the MLST.

Options

- ☐ Select default nearest organisms (skip Step 2)
- ☐ Select default MLST genes (skip Step 3)
- ☐ Perform IQ-TREE Ultrafast Bootstrap analysis (1000 replicates)
- ☐ Run ModelFinder
- ☒ Concatenated alignment
- ☐ Coalescent tree

Sequences

Accession numbers and uploaded files

Remove selected sequences

Upload Genomes

NCBI Accession Number

Sequence(s) (one genome per file) No file chosen

Options

Enter job title (optional)

Save links ☐ Save link in cookies (viewable on results page under "last results")

Submit Job »

Figure 17 : Etapes de MLST selon le mode de novo

Un maximum de 50 organismes sera utilisé pour la construction d'arbres ; les espèces supplémentaires dépassant ce nombre ne seront pas utilisées. Les séquences dépassant cette limite sont surlignées en rouge.

Un tableau des organismes les plus proches de leurs séquences soumises, tel que déterminé par une estimation de (ANI) l'identité moyenne des nucléotides. (Le tableau est trié par défaut par ANI le plus élevé).

Méthodes basées sur l'amplification de l'acide nucléique

Mash-based ANI estimation

Show 10 entries Search:

Query organism	Reference assembly ID	Reference name	Mash distance	estimated ANI	P-value	Genus	Order
Hyalangium_minutum_DSM_14724.fa	GCF_000737315	Hyalangium minutum	0.0000	100.0%	0.0000	Hyalangium	Myxococcales
Plesiocystis_pacifica_SIR-1.fa	GCF_000170895	Plesiocystis pacifica SIR-1	0.0000	100.0%	0.0000	Plesiocystis	Myxococcales
Type strain True							
Sanderscinus_amyolyticus_DSM_53668.fa	GCF_000737325	Sanderscinus amyolyticus	0.0000	100.0%	0.0000	Sanderscinus	Myxococcales
Sorangium_cellulosum_So_ce56_So_ce_56.fa	GCF_000067165	Sorangium cellulosum So ce56	0.0000	100.0%	0.0000	Sorangium	Myxococcales
Myxococcus_xanthus_DK_1622.fa	GCF_800106535	Myxococcus xanthus	0.0004	100.0%	0.0000	Myxococcus	Myxococcales
Myxococcus_xanthus_DK_1622.fa	GCF_000012685	Myxococcus xanthus DK 1622	0.0000	100.0%	0.0000	Myxococcus	Myxococcales
Myxococcus_xanthus_DK_1622.fa	GCF_000340515	Myxococcus xanthus DZF1	0.0004	100.0%	0.0000	Myxococcus	Myxococcales
Myxococcus_xanthus_DK_1622.fa	GCF_000278585	Myxococcus xanthus DZ2	0.0005	100.0%	0.0000	Myxococcus	Myxococcales

En fonction du flux de travail sélectionné et des choix de l'utilisateur, un arbre généré à partir d'un alignement de séquences multi-locus, un arbre coalescent ou un arbre généré par l'ajout de séquences utilisateur à un arbre de référence est présenté à l'utilisateur.

Les séquences soumises par l'utilisateur sont préfixées par l'abréviation « QS » ; leurs nœuds respectifs sont colorés en bleu.

Les groupes externes sont préfixés par l'abréviation « OG » ; leurs nœuds respectifs sont colorés en rouge.

Les souches types sont préfixées par l'abréviation « TS » ; leurs nœuds respectifs sont colorés en vert.

Il est également possible de rechercher des organismes dans l'arbre ; les organismes correspondant au terme de recherche sont affichés avec une taille de police plus grande. Plusieurs détails concernant la visualisation de l'arborescence peuvent être modifiés par l'utilisateur.

Après la construction de l'arbre, l'utilisateur peut le télécharger sous différents formats, par exemple : le format Newick (sans couleurs), l'arbre au format .svg (avec couleurs). De plus, il peut récupérer un fichier compressé contenant les alignements à partir desquels il est construit la liste des gènes utilisés dans l'analyse MLST et la liste des organismes pour lesquels la liste complète des ANI a été estimée entre leurs organismes et les organismes de référence.

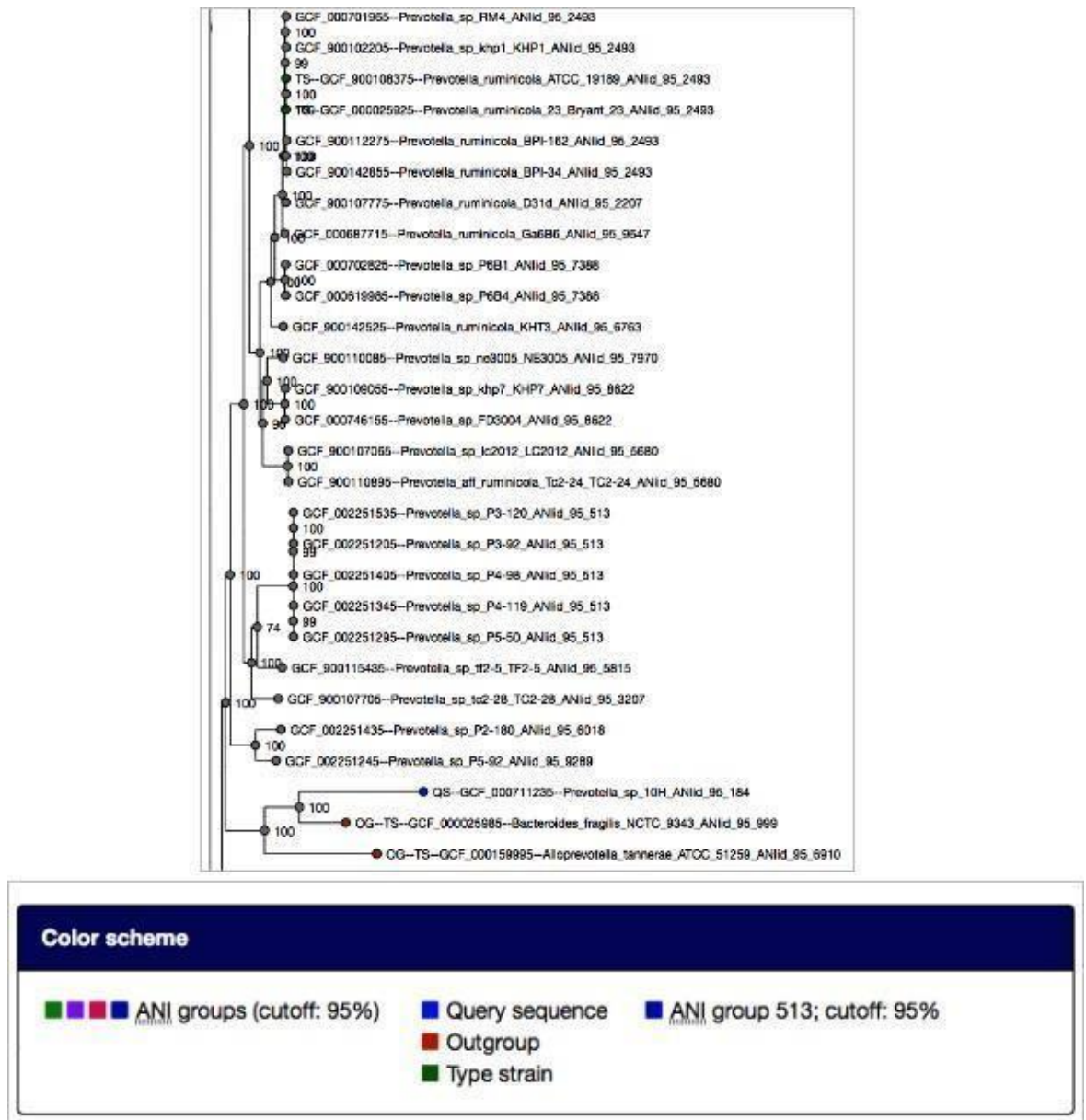


Figure 18 : Dendrogramme construit par autoMLST

Partie III : Les outils de la bio-informatique

1. Banques de données biologiques

La recherche d'informations biologiques via le Net est contemporaine. Sans une approche structurée, les chercheurs d'informations peuvent se sentir perdus dans cette toile d'araignée géante. Il en résulte que l'utilisation de méthodes organisées et structurées pour la recherche est devenue essentielle. Cela peut vous faire économiser énormément de temps tout en vous permettant d'entreprendre des recherches plus spécifiques.

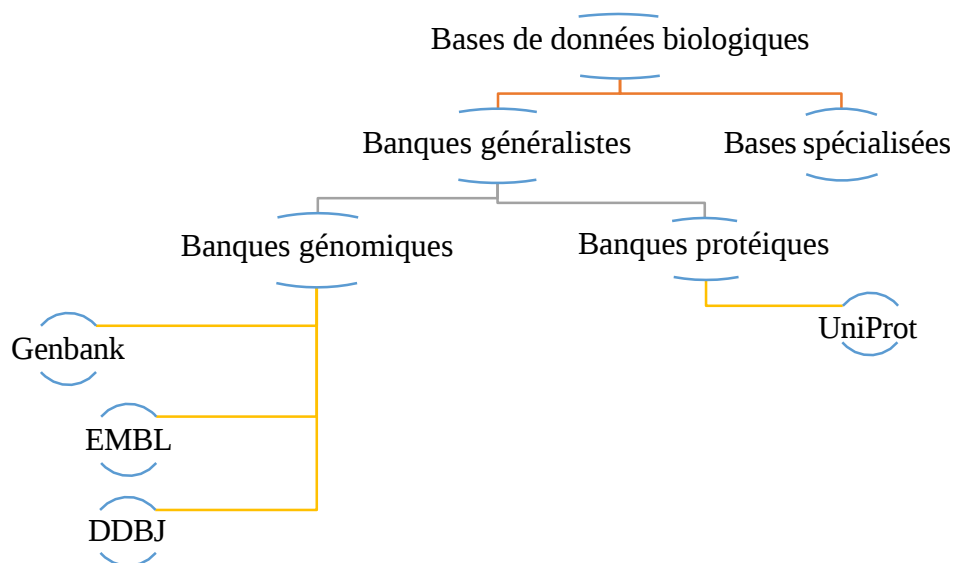
En bio-informatique, une base de données biologique peut contenir des informations sur les protéines, de même, elle peut contenir des informations sur les gènes, les plasmides...etc.

Il existe actuellement une variété de bases de données biologique disponibles comme référence.

1.1. Les bases de données et les banques de données biologiques

Ces bases fournissent une fiche descriptive pour une séquence d'acide nucléique ou de protéine (AND, ADNc, ARN, protéine). Ces fiches sont appelées 'Entrées'.

Une banque de données est une base de données (car c'est un tableau structuré), mais elle contient des informations biologiques **hétérogènes** (virus, bactéries, champignons, plantes, animaux), alors qu'une base de données est plus **spécialisée** (base de données spécifique aux bactéries : *Bacillus*, *E. coli*, etc.).



1.1.1. Les bases de données biologiques

- **Base de données MEDLINE** (Medical Literature Analysis and Retrieval System Online) : est une base de données bibliographiques, gérée par la bibliothèque nationale américaine (United States National Library of Medicine) qui couvre tous les domaines biomédicaux de l'année 1966 à nos jours.

Plus de 21 millions de références issues de plus 5 000 périodiques, principalement en langue anglaise sont disponibles en octobre 2014.

- **Base de données PDB (Protein Data Bank)**

C'est une base de données de structure 3D des protéines, elle compile des agrégats moléculaires (protéines, ADN, ARN...) dont les structures sont connues. Au 18 février 2021, le PDB contient 174 826 entrées, toutes avec un code à quatre chiffres (un chiffre suivi de trois caractères alphanumériques ; par exemple, le code 1MBC, fait référence à la structure de la myoglobine).

Pour connaître la structure 3D d'une molécule, c'est-à-dire la position de ses atomes dans l'espace, nous utilisons des techniques expérimentales telles que la diffraction des rayons X, la diffraction des neutrons, la résonance magnétique nucléaire (RMN) ou la microscopie électronique.

1.2. Les banques de données

Une banque de données biologique, est une base de données généraliste. Elle regroupe des informations sur de nombreuses espèces et de nombreuses molécules. Leur utilisation touche plusieurs domaines et correspond à des ensembles de données complets contenant des informations hétérogènes.

Il faut savoir qu'il existe des banques de données nucléiques et une banque de données protéiques.

1.2.1. Les banques de données nucléiques

- **GenBank**

La banque de données GenBank qui est hébergée au portail NCBI, est une banque universelle contenant toutes les séquences nucléiques, quelle que soit leur nature (ADN génomique, ARN messager, ADN plasmidique, etc.). Des séquences de plus de 100 000 organismes différents produites dans des laboratoires du monde entier sont régulièrement soumises au NCBI (National Center for Biotechnology Information). Par conséquent, GenBank connaît

une croissance exponentielle, doublant sa taille tous les 10 mois. Pour y accéder nous utilisons l'URL suivant : <https://www.ncbi.nlm.nih.gov/genbank/>.

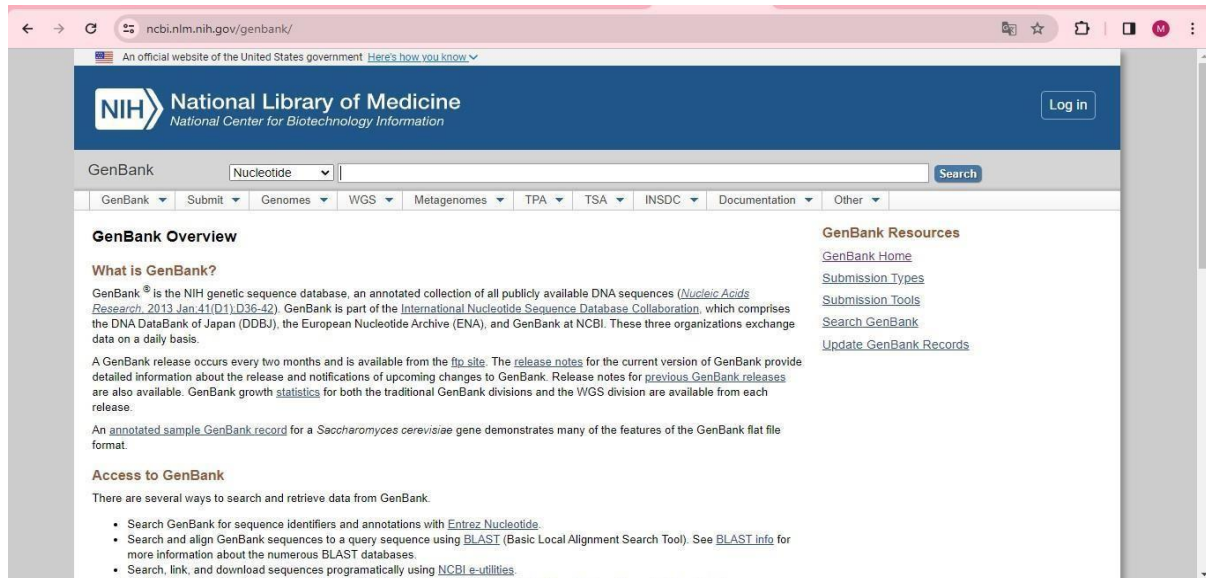


Figure 19 : L'interface de la banque de données GenBank.

- **EMBL**

La banque nucléique EMBL de EMBO (European Molecular Biology Organization) : EMBL Bank est établi et gérée à l'Institut européen de bio-informatique (EBI)(<https://www.ebi.ac.uk/>.) près de Cambridge, au Royaume-Uni.

En 12 mois, la taille est passée d'environ 6,7 millions d'entrées contenant 8,255 milliards de nucléotides (version 63, juin 2000) à plus de 12 millions d'entrées et 12,82 milliards de nucléotides (version 67, juin 2001).

Au cours de la même période, le nombre d'organismes inclus dans la base de données a augmenté de plus de 30 % pour atteindre plus de 75 000 espèces. (Source NCBI, consulté le 08/02/2023 : <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC99098/>).

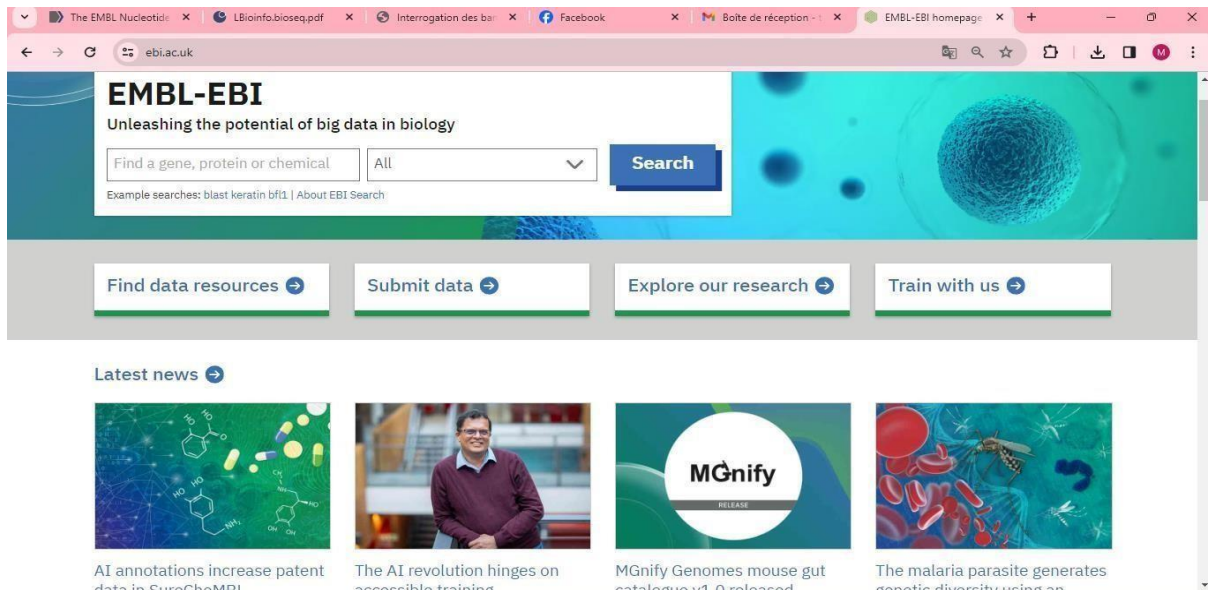


Figure 20 : L'interface de banque de données EMBL

- **DDBJ**

La banque de données DDBJ (<http://www.ddbj.nig.ac.jp>) est une banque de séquences nucléotidiques hébergée au Japon et créée par l'institut national de génétique (NIG).

La DDBJ a publié un total de 11909516 entrées WGS (1694 génomes), 1505087 entrées contig/construites (CON), 1313171 entrées TSA (18 projets), 786 entrées TPA, 6374 entrées TPA-WGS (un génome) et 1272 entrées TPA-CON au 27 mai 2016. (Source NCBI : <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5210514/#B1>).

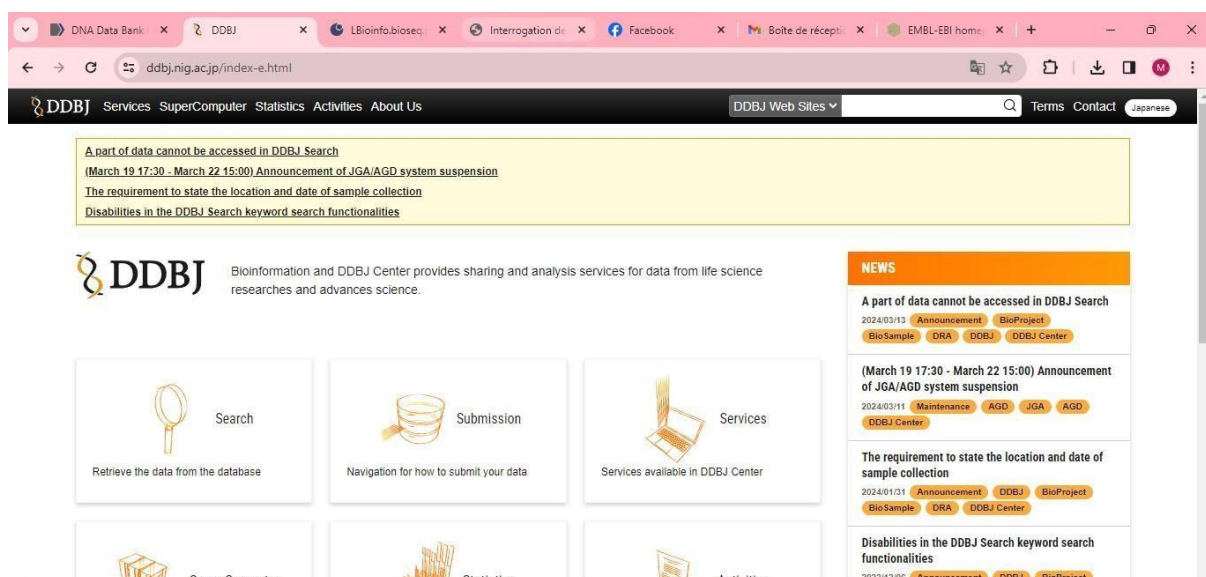


Figure 21 : L'interface de la banque de données DDBJ.

1.2.2. Les banques de données protéiques

La traduction des gènes d'ADN donne des protéines ayant différentes fonctions, ces biomolécules possèdent des structures des propriétés importantes, les chercheurs du monde les caractérisent et les mettent sur des banques de données seulement pour les protéines, ici en cite quelques une les plus connues.

UniProt : est une banque de données des séquences protéiques accessible en ligne, fournie des données robustes, complètes sur les séquences protéiques et les annotations fonctionnelles.

Cette banque de données est régulièrement mise à jour et combine les données de deux bases de données : Swiss-Prot et TrEMBL.

Le consortium UniProt est une collaboration entre l'Institut européen de bio-informatique (EBI), l'institut suisse de bio-informatique (SIB). UniProt comprend de quatre composants principaux, où chacun est optimisé pour un objectif différent. Base de connaissances UniProt, clusters de référence UniProt (**UniRef**), archives UniProt (**UniParc**) et base de données de séquences métagénomiques et environnementales UniProt (**UniMES**).

Elle est mise à jour chaque trois semaines.

Les activités clés de cette banque de données comprennent :

- La conservation manuelle des séquences de protéines à l'aide d'une analyse informatique.
- L'archivage des séquences.
- L'élaboration d'un site web UniProt intuitif et la mise à disposition d'informations enrichissantes par le biais de références diffusées avec d'autres bases de données.

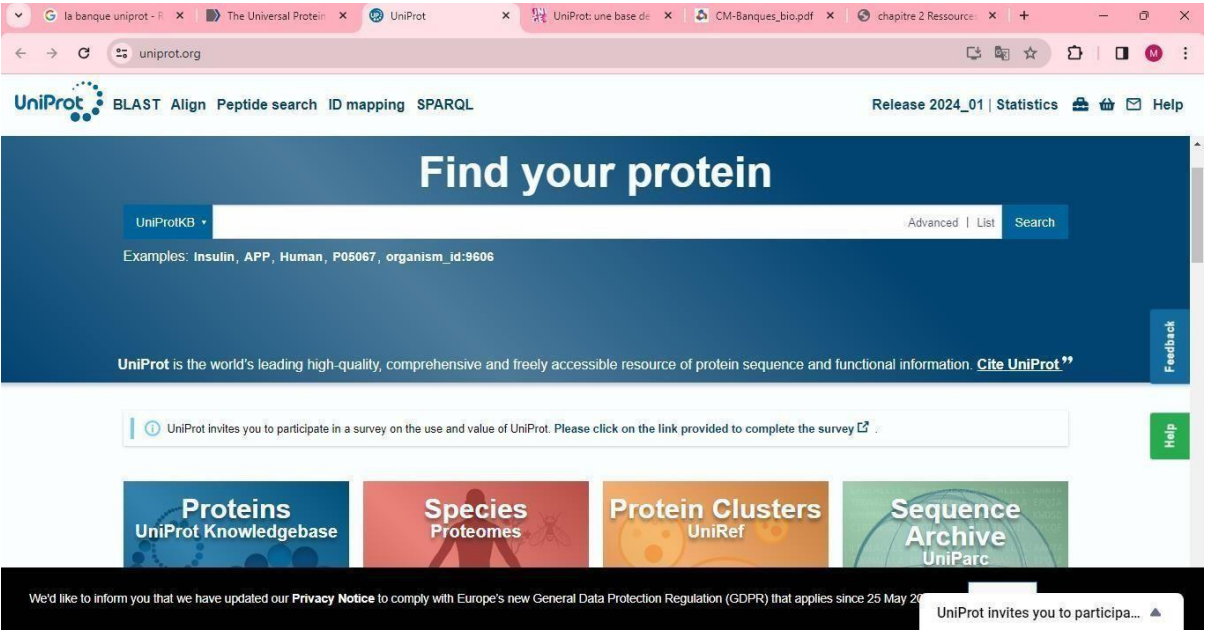


Figure 22 : Interface de la banque de données protéiques UniProt.

1.2.3. Les bases de données spécialisées

Nom	Site	Description
ENZYME	https://enzyme.expasy.org/	La base de données ‘ENZYME’ est un référentiel d'informations relatives à la nomenclature des enzymes, les réactions et les classifications Enzymatiques.
PFAM d’InterPro	https://www.ebi.ac.uk/interpro/	La base de données ‘InterPro’ est une ressource qui permet l'analyse fonctionnelle des séquences protéiques en les classant en familles et en prédisant la présence de domaines et d'emplacements importants.
Prosite	https://prosite.expasy.org/	La base de données ‘PROSITE’ peut être considérée comme un dictionnaire qui recense des motifs protéiques ayant une signification biologique.
UniLectin	https://unilectin.unige.ch/	La base de données ‘UniLectin’ pour les protéines, appelées Lectines ou agglutinines qui sont des

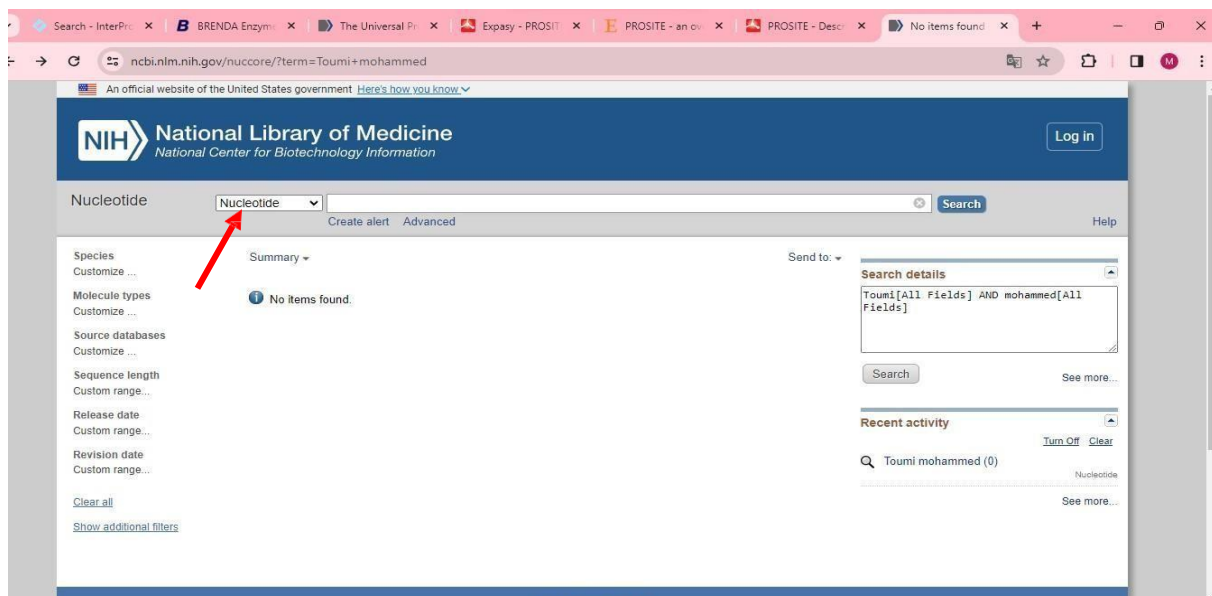
		molécules de reconnaissance et d'adhésion cellulaire.
BRENDA	https://www.brenda-enzymes.org/	Spécifique aux enzymes, leurs propriétés et la prédiction de leurs mécanismes de fonction.

1.2.4. La structure d'une entrée des banques de données

Le résultat d'une recherche ou une interrogation sur une banque de données est appelé une **‘Entrée’**, cette dernière est une fiche descriptive comportant toutes les informations nécessaires sur la séquence génomique ou protéique recherchée de façon de présentation différentes entre les banques de données, la recherche s'effectue en utilisant des mots clés en anglais.

- **Entrée de la banque ‘NUCLEOTIDE’ (GenBank)**

Avant de lancer une requête, il faut aller au menu déroulant (flèche rouge) pour choisir la banque NUCLEOTIDE, puis insérer le nom ou le numéro d'accèsion de la séquence cible.



Il y aura l'apparition d'une entrée comporte un ensemble des mots (champs) qui ont les significations suivantes les éléments suivants :

- Locus : code d'accèsion du gène dans la figure est MT553101, la taille : 734 bp, type ADN ou ARN (ici c'est un ADN),
- La date de publication : 08- JUN-2020.
- Définition : le nom du gène.

- Source : l'organisme source du gène, leur classification taxonomique.
- Authors : les auteurs qui fournissent ce travail et leurs affiliations.
- Origin: la séquence du gène sous format GenBank.

The screenshot displays a GenBank entry for the Cyclocybe cylindracea strain MEST42. The entry is titled "Cyclocybe cylindracea strain MEST42 small subunit ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and large subunit ribosomal RNA gene, partial sequence". The accession number is MT553101.1. The entry includes a detailed definition, accession number, version, keywords, source, organism, reference, authors, title, journal, and a comment. The 'ORIGIN' section shows the nucleotide sequence in FASTA format, which is highlighted with an orange box.

FASTA Format:

```

1 tcttcttggg tccatttaga ggaagtaaaa gtctgaacaa ggtttccgta ggtgaacctg
61 cgaagagctg attactgaat aaaactcgtt gggcttgatg ctggctttct ttggaaggca
121 tgtgcacgcg ctgtctcttt tatttctttt ccactgtgac accctttgta ggctgaacc
181 gctttttctt ttggcctttg ggtcagtggt ggggactgct gaccgctgtt ctttaccgt
241 tctgagctat gcttttacac tatacaccat tgttaaacct tagaatgta aaggctctac
301 gtgtgcctat cttaaacact atacaacctt cagcaacgga tctcttggtc ctgcctcga
361 tgaagaacgc agcgaatgc gataagtaat gtgaattgca gaattcagtg aatcctcga
421 tcttgaacgc caacttgctc tcttggatgt tctgaaggag atgcttggtt ggtgtcatt
481 acattctcaa ccttggaatt tgtttcgtat ggttgacct tggggatgag tggcggcct
541 aaaggtcgcg tcccttaaaa tgcattagct ggttgacct ctgcgctgaa cttggtgtga
601 taattcttac ttgacgcgt tttgacgtg ggtgtctttt ggtgactgc ttttaacgg
661 tcccttgagg acaacgataa aatccctctt tcggagggtc tttttgacc tctctgact
721 caaatcaggt aggt

```

Figure 23 : Structure d'une entrée de la banque Nucleotide (GenBank).

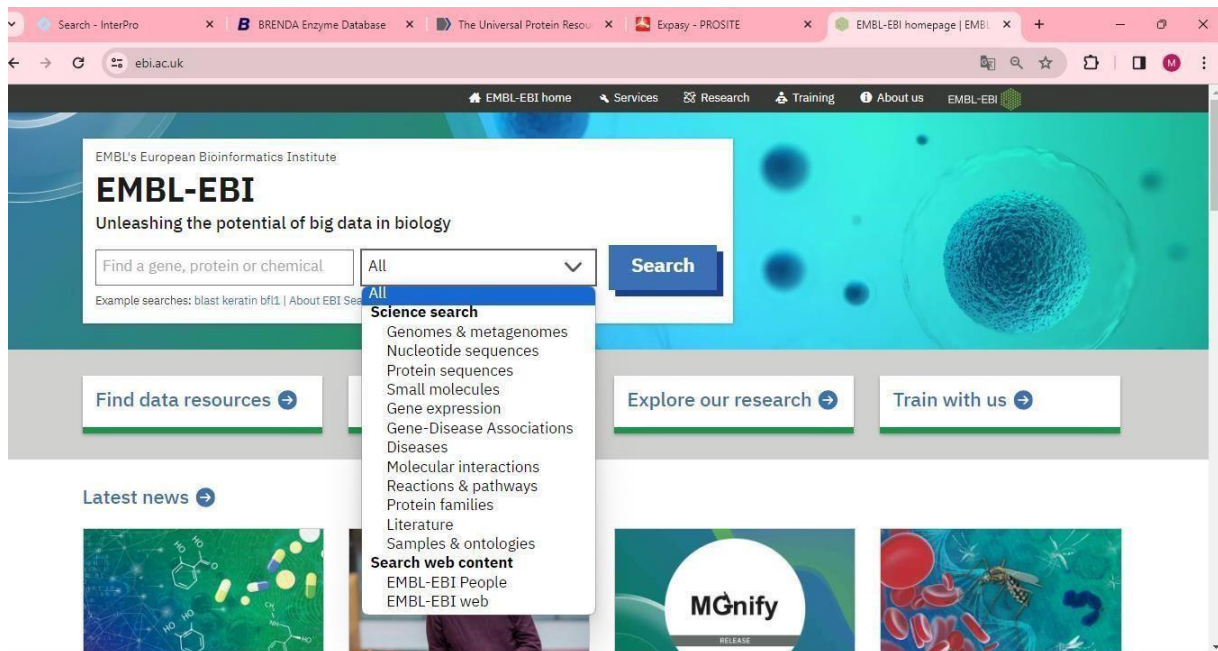
Remarque : le mot **FASTA** en bleu représente un autre format de la séquence en mode texte sans numérotation.

The screenshot shows the NCBI GenBank website displaying a nucleotide sequence in FASTA format. The browser tabs at the top include Search - InterPro, BRENDA Enzymes, The Universal Protein, ExPASy - PROSITE, PROSITE - analysis, PROSITE - description, and Cyclocybe cylindracea. The URL in the address bar is ncbi.nlm.nih.gov/nuccore/MT553101.1?report=fasta. The page features a search bar at the top with the text 'Nucleotide' and a 'Search' button. Below the search bar, the sequence is displayed in FASTA format. The header line is: >MT553101.1 Cyclocybe cylindracea strain MEST42 small subunit ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and large subunit ribosomal RNA gene, partial sequence. The sequence itself is a long string of nucleotides: TCTTCTTGGTTCATTTAGAGGAAGTAAAGTGTGTAACAAGGTTTCCGTAGGTGAACCTGCGGAAGGATCATTACTGAATAAACTCGGTGGGCTTGAGTGTGCTTTCTTTTGAAGGCATGTGACCGCTGTGCTTTTAAATCTTTCCACCTGTGACCGCTTTGTAGGCTTGAACCGCTTTTCTTTTGGCTTTGGTCAAGTGTGGGACTGTGACCGGCTGTCTTCTACCGTTCGAGTCTATGCTTTTACACTATACACATTGTTAAACCTTAGAATGTTAAAGGCTCTACGTGTGCTTCTTAAACACTATACACCTTCAGCAACGGATCTCTTGGCTCTCGCATCGATGAAGAAGCGAGGAAATGCGATAAGTAATGTGAATTGCAAGATTCAGTGAATCATCGAACTTTTGAACGACCTTGCCTCTTGGTATTCCGAGGAGCATGCTGTTTGAAGTGCATTACATCTCAAACATTGGATTTGTTTTCGATGGGTTGGACTTGGGGGTGTGTGCGGCGCTAAAGGTGCGCTCCCTTAAATGTCATTAGCTGGTTGCCCTCTCGCTGCTGAGTGGTGTGATAAATCTTACTTTGCGCCGTTTGTGACGTGATGCTCTTTGGGTGGCTGCTTCTAACCGTCCCTTGTGGAACGATAAAATCCCTCTTCTGGAGGGCTTTTGTGACCTCTCTGACCTCAATCAGGTAGGT.

Figure 24 : Format FASTA d'une séquence nucléique

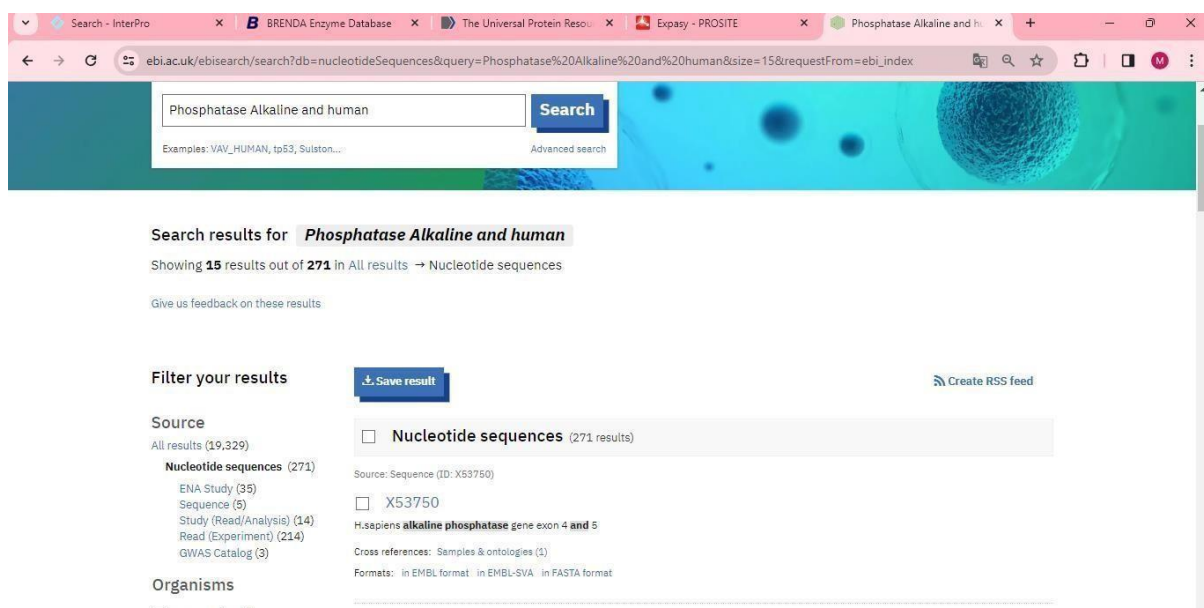
- **Entrée de la banque EMBL**

- Après avoir accéder à la banque EMBL (URL : <https://www.ebi.ac.uk/>).
- Une fenêtre est apparue, elle représente l'interface générale de la banque.
- Dans la banque, on peut faire des recherches et importer des données, et aussi soumissionner de nouvelles séquences nucléiques.
- Il est important d'utiliser des mots clés et l'un des opérateurs booléens (AND, OR, NOT) au moment de recherche pour mieux cibler le résultat de requête.



Exemple sur un résultat de recherche sur EMBL :

- Dans la case de recherche, nous écrivons la requête suivante : **“Alkaline Phosphatase AND homo sapiens”**.
- En choisissant **“Nucleotides sequences”**, puis on lance la recherche.
- Le résultat apparu montre qu’il y aura 271 séquences nucléotidiques pour cette enzyme.



Pour obtenir le résultat de recherche, nous devons cliquer sur le code **“X53750”**, ensuite, il faudra aller à droite et cliquer sur EMBL :

Sequence: X53750.1
H. sapiens alkaline phosphatase gene exon 4 and 5

Organism:	Homo sapiens (human)
Mol Type:	genomic DNA
Topology:	linear
Base Count:	1795
Dataclass:	STD
Tax Division:	HUM
Accession:	X53750
Keywords:	alkaline phosphatase
Chromosome:	1
Md5 Checksum:	f04374285d1afe8ba3c4212b03c75148
Dev Stage:	fetal
Map:	p34-p36.1
Tissue Type:	liver

Show Less

View: EMBL FASTA
Download: EMBL FASTA
Navigation: Show
Publications: Show
Sequence Versions: View

Dans l'entrée la banque de données EMBL, les code à deux lettres, situés dans la partie gauche sont dénommés, **Etiquettes**, ces dernières ont les significations suivantes :

ID : identifiant ou le code d'accension avec la taille, la nature de la séquence et sa source biologique.

AC : le numéro d'accension (équivalent au champ *locus* dans l'entrée de la banque GenBank).

XX : ligne vide

DE : définition.

OS : organisme source.

OC : organisme classification.

FT : caractéristiques de la séquence

SQ : séquence du gène.

```

ID   X53750; SV 1; linear; genomic DNA; STD; HUM; 1795 BP.
XX
AC   X53750;
XX
DT   13-APR-1992 (Rel. 31, Created)
DT   14-NOV-2006 (Rel. 89, Last updated, Version 2)
XX
DE   H.sapiens alkaline phosphatase gene exon 4 and 5
XX
KW   alkaline phosphatase.
XX
OS   Homo sapiens (human)
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
OC   Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae;
OC   Homo.
XX
RN   [1]
RP   1-1795
RA   Hsu H.H.T.;
RT   ;
RL   Submitted (05-JUL-1990) to the INSDC.
RL   Hsu H.H.T., University of Kansas Medical Centre, Pathology Department, 39th
RL   and Rainbow Blvd, Kansas City Kansas 66103, U S A.
XX
RN   [2]
RA   Hsu H.H.T.;
RT   "The partial sequencing of human non-specific (bone liver kidney) alkaline

```

```

FT   source      1..1795
FT               /organism="Homo sapiens"
FT               /chromosome="1"
FT               /map="p34-p36.1"
FT               /mol_type="genomic DNA"
FT               /dev_stage="fetal"
FT               /clone_lib="Charon 4A phages"
FT               /clone="CPSTC4M1"
FT               /tissue_type="liver"
FT               /db_xref="taxon:9606"
FT   exon        7..181
FT               /number=4
FT   exon        926..1101
FT               /number=5
XX
SQ   Sequence 1795 BP; 395 A; 516 C; 532 G; 352 T; 0 other;
ctgcagacgt acaacaccaa tgcccagggt cctgacagcg ccggcaccgc caccgcctac    60
ctgtgtgggg tgaaggccaa tgagggcacc gtgggggtaa gcgcagccac tgagcgttcc    120
cgggtgcaaca ccacccaggg gaacgagggt acctccatcc tgcctggggc caaggacgct    180
ggtgagtcgg gggagcagtg gggagcaggg ccagcttcgt gccctgtcca ggctcagtct    240
ttctgactgt gtcattggaa ggggctagaa aaggcttctc tgtggggctc ccagctctga    300
tatgctggga gtctgtaata tccatgggat tcctccgtg acaggggaag ctgagggttt    360
ggggttcact cagtgggtgg ttctggtccc agttcatttc tctgagtcga ttcttcattt    420
gactataaac ggggtggaac tagccctcca taggttgtgt gagaggagaa gtcagccagg    480
ctagcctgtg ggtgcagtc atgccacgcc ccttccctcc tcagaattgg gagccagccc    540
tgtgccccca gaagcttag ggtcagggcc aaggccaaaa cacagggtgac aaggccaaca    600
tacagggtgac agagccacag tccacttgag atccctcagg ggtcctgtgg gcgccttgat    660
-----

```

2. Alignement des séquences biologiques (séquences génomiques et protéiques)

L'alignement est le processus par lequel deux séquences sont comparées afin d'obtenir le plus de correspondances possibles entre les lettres qui les composent. Il permet de déterminer le degré de similitude ou identité entre ces séquences. Néanmoins, Comparer des séquences biologiques ainsi que leur alignement nécessite la mise en œuvre de procédures informatiques et de modélisation biologique pour pouvoir quantifier la notion de similarité. Une similarité entre les séquences alignées indique une fonction biologique proche et une origine commune.

L'alignement des séquences a pour objectif de découvrir des régions fonctionnelles afin de déterminer la fonction et la structure d'une séquence ou étudier le lien évolutif à l'échelle moléculaire pour établir une étude phylogénétique des espèces.

En effet, il existe différents type d'alignement :

- Alignement global : aligner deux séquences sur toute leur longueur. Cela permet de déterminer à quel point deux séquences sont similaires.

- Alignement multiple : alignement appliqué à plusieurs séquences à la fois. Il permet de rechercher des motifs conservés, de prédire des structures, de réaliser des études phylogénétiques.
- Alignement local : comparer deux séquences seulement sur une partie (blocs) de leur longueur. Il permet de comparer une chaîne inconnue avec une série de chaînes connues.
- Alignement paire : correspond à la comparaison de deux séquences seulement.

Un alignement repose essentiellement sur une entité numérique appelée le score (s) qu'on attribue à chaque paire de nucléotides des deux séquences à aligner. Lors d'une comparaison, ce score prend la valeur 1 lorsque les deux nucléotides des deux séquences sont identiques et la valeur 0 s'ils ne sont pas.

Le score global (S) d'alignement correspond à la somme des scores élémentaires (s)

$$S = \sum_{i=1}^n s_i$$

Exemple :

Séquence S1	G	A	A	T	-	G	G	C	C	G
			*		*					
Séquence S2	G	A	-	T	T	C	G	C	-	G
Score élémentaire (s)	1	1	0	1	0	0	1	1	0	0

Donc le score global de cette comparaison est égal à la somme des scores élémentaires $(1+1+0+1+0+0+1+1+0+0) = 5$

En bio-informatique, nous utilisons différentes matrices et méthodes pour réaliser un alignement entre les séquences

2.1. Les méthodes d'alignements des séquences nucléiques et protéiques

Ce sont des méthodes qui prennent en considération la séquence de façon entière et aboutissent à une comparaison globale de la première séquence avec la deuxième.

2.1.1. Dotplot

Le dotplot est l'une des méthodes les plus anciennes pour comparer deux séquences

Le principe de cette méthode est basé sur la comparaison de fenêtres de longueur fixe que l'on déplace le long des séquences.

Les diagonales représentent chacun une séquence et lui-même montre une comparaison de ces deux séquences par un score calculé pour chaque position de la séquence. Si une fenêtre de taille fixe sur une séquence (un axe) correspond à l'autre séquence, un point est dessiné sur le plan.

La ressemblance des séquences est "lue" dans les diagonales.

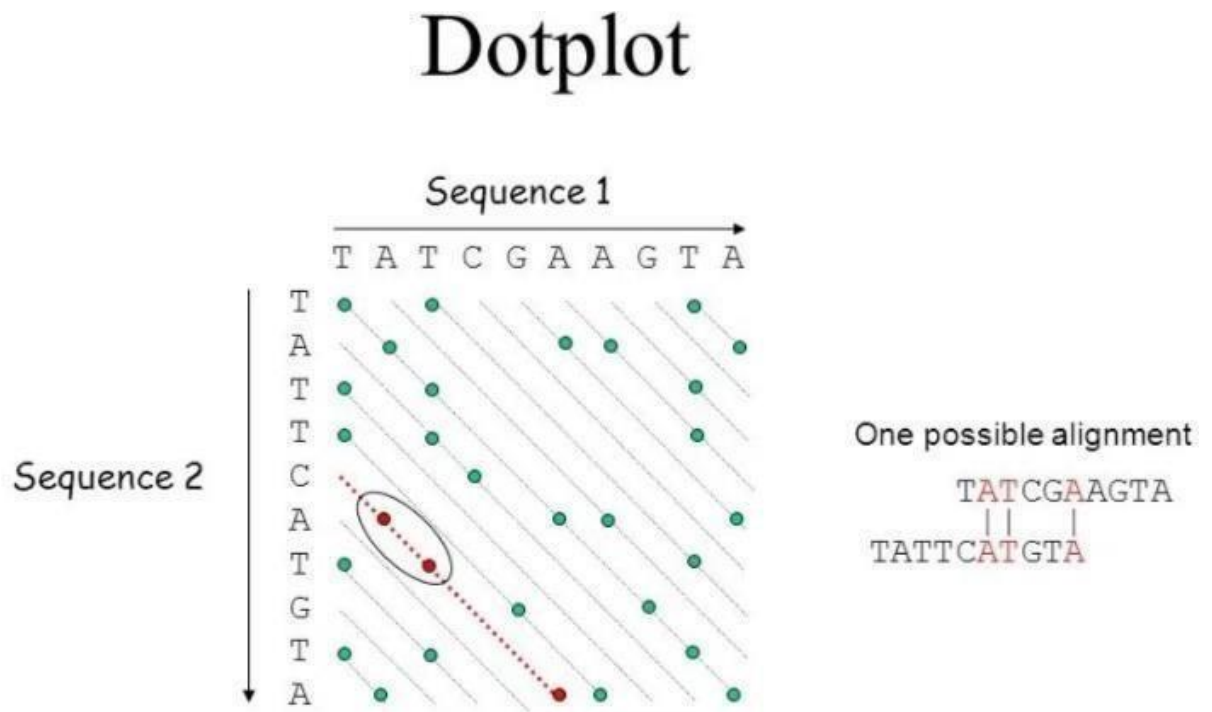


Figure 25 : Représentation graphique des points communs entre les séquences selon la méthode de Dotplot.

Pour réaliser un alignement entre 2 séquences par la méthode de Dotplot
 Nous utilisons le logiciel de Dotlet JS interactif qui est disponible sur le site :
<https://myhits.sib.swiss/cgi-bin/dotlet>

Dotlet JS *beta*

SEQUENCE 1 SEQUENCE 2 SAVED SEQUENCES Window size 15 Scoring matrix Identity

Sequence 2

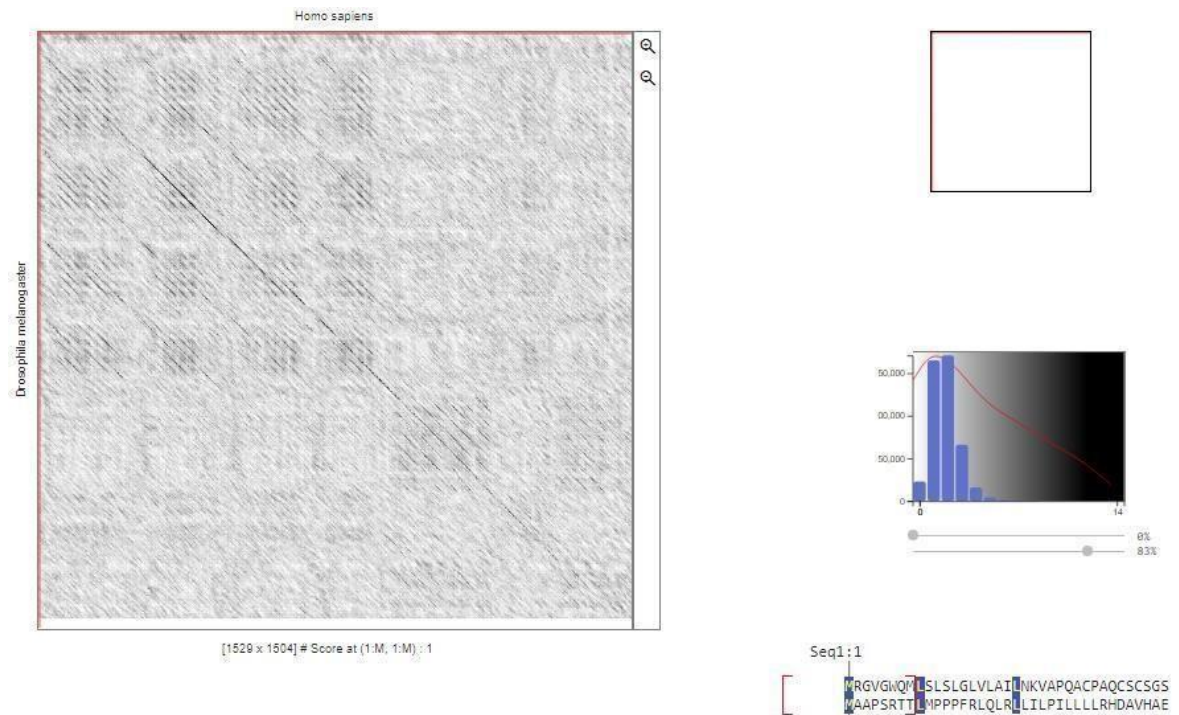
Sequence 2:

Sequence 1

Sequence 2

00,000 00,000

Il suffit d'insérer les deux séquences pour obtenir le résultat suivant :



2.1.2. L'alignement par programmation dynamique (Needleman et Wunsch)

La comparaison de deux séquences A et B (formés des lettres A, C,G,T) est possible grâce à l'algorithme de Needleman Wunsch datant de 1969. Cet algorithme permet de réaliser un alignement global entre deux séquences nucléiques. Son expression est de la forme :

Nous pouvons abouti à séquence optimale par l'algorithme ci-dessous. Nous commençons à parcourir la dernière ligne, puis la dernière colonne de la matrice, et pour chaque ligne colonne après colonne (en partant de la dernière). L'expression de cet algorithme est :

$$S(i,j) = \text{Max} \begin{cases} S(i+1,j+1) + s(i,j) \\ s(i+1,j) \\ s(i,j+1) \end{cases}$$

La similarité entre deux séquences est égale à la valeur $S(1,1)$ et l'alignement est un graphe d'origine $S(1,1)$ et parcourt la matrice croissante i et j pour trouver le voisin maximum de l'élément.

Dans l'exemple suivant, nous désirons calculer un alignement global des deux séquences suivantes de taille m et n respectivement :

$S1 = \text{GAATGGCCGC}$ $m=10$ et $S2 = \text{GATTGGCGC}$ $n=9$ ($S1$ à position $S1$ horizontale et $S2$ à la verticale de la matrice).

-La première étape consiste à calculer la matrice initiale en affectant la valeur 1 en cas d'identité entre deux nucléotides et la valeur 0 au cas contraire.

	G	A	A	T	G	G	C	C	G	C
G	1	1	0	0	1	1	0	0	1	0
A	0	0	1	0	0	0	0	0	0	0
T	0	0	0	1	0	0	0	0	0	0
T	0	0	0	1	0	1	0	0	0	0
G	1	1	0	0	1	1	0	0	1	0
G	1	1	0	0	1	1	0	0	1	0
C	0	0	0	0	0	0	1	1	0	1
G	1	1	0	0	1	1	0	0	1	0
C	0	0	0	0	0	0	1	1	0	1

-Dans la deuxième étape, nous calculons la matrice transformée dans laquelle la 1ère ligne et la 1ère colonne seront initialisées à zéro :

		G	A	A	T	G	G	C	C	G	C
		0	0	0	0	0	0	0	0	0	0
G	0										
A	0										
T	0										
T	0										
G	0										
G	0										
C	0										
G	0										
C	0										

-L'étape suivante correspond à l'exécution de l'algorithme de Needleman Wunsch pour remplir les cases, on obtient alors le tableau suivant :

		G	A	A	T	G	G	C	C	G	C
		0	0	0	0	0	0	0	0	0	0
G	0	1	1	1	1	1	1	1	1	1	1
A	0	1	2	2	2	2	2	2	2	2	2
T	0	1	2	2	3	3	3	3	3	3	3
T	0	1	2	2	3	3	3	3	3	3	3
G	0	1	2	2	3	4	4	4	4	4	4
G	0	1	2	2	3	4	5	5	5	5	5
C	0	1	2	2	3	4	5	6	6	6	6
G	0	1	2	2	3	4	5	6	6	6	6
C	0	1	2	2	3	4	5	6	7	7	8

On va parcourir alors du graphe en partant de la case maximale qui comporte le plus haut score calculé (ici $s=8$) jusqu'au score le plus petit ($s=1$).

		G	A	A	T	G	G	C	C	G	C
	0	0	0	0	0	0	0	0	0	0	0
G	0	1	1	1	1	1	1	1	1	1	1
A	0	1	2	2	2	2	2	2	2	2	2
T	0	1	2	2	3	3	3	3	3	3	3
T	0	1	2	2	3	3	3	3	3	3	3
G	0	1	2	2	3	4	4	4	4	4	4
G	0	1	2	2	3	4	5	5	5	5	5
C	0	1	2	2	3	4	5	6	6	6	6
G	0	1	2	2	3	4	5	6	6	6	6
C	0	1	2	2	3	4	5	6	7	7	8

Dans la dernière étape, nous allons calculer le score global et le pourcentage d'identité

Séquence S1	G	A	A	T	-	G	G	C	C	G	-	C
			*		*						*	
Séquence S2	G	A	-	T	T	G	G	C	-	G	G	C

- Le score global de cet alignement est de 9.
- Le pourcentage de l'identité entre les deux séquences S1 et S2 est égal à 75%
- ($\%id = (9/12) * 100 = 75\%$)

D'après le résultat de cet alignement, un Gap (InDel) a été retrouvé entre les nucléotides A et T de la séquence 2, ce qui signifie qu'à ce moment-là, la séquence S2 a subi une mutation par délétion dans laquelle le nucléotide A a été perdu au cours de l'évolution pour l'adaptation. En même temps, il est conservé dans la séquence S1 (à la 3^{ème} position), de même, on peut supposer que c'est la séquence S1 qui a subi une mutation par l'insertion du nucléotide A en raison de besoins adaptatifs. Dans un cas ou un autre, une des deux séquences a été mutée (par insertion ou délétion). Le constat concerne les deux Gaps en 5^{ème} et 1^{ème} position.

2.1.3. L'alignement des séquences protéiques

En effet, les systèmes nucléiques basés sur l'identité ne conviennent pas dans le cas des systèmes protéiques. Cela est dû au fait que certains acides aminés peuvent être remplacés par d'autres

acides aminés (notamment en raison de leurs propriétés physico-chimiques) sans modifier le rôle biologique et la fonction de la protéine.

Tenant compte tout d'abord que les matrices protéiques utilisées pour réaliser des alignements sont totalement différentes de celles des acides nucléiques, en raison des ressemblances entre acides aminés. Ces ressemblances sont de plusieurs types : - de même charge - hydrophiles / hydrophobes - type de groupement latéral : polaire (groupe I), polaire non chargé (groupe II), chargé (groupe III).

Les matrices les plus utilisées et reconnues comme les plus performantes sont : la matrice PAM (Dayhoff 1978) et la matrice BLOSUM (Henikoff et Henikoff 1992).

- **La matrice PAM 250**

Crée par Margareth Dayhoff en 1970. Elle provient d'un alignement à la main de 3000 protéines regroupées en 71 familles avec un total de 1600 mutations et les séquences d'une même famille ont au maximum 15% de différence, afin de déterminer les mutations les plus fréquentes d'un acide aminé à un autre.

La matrice obtenue, PAM1, (respectivement PAM250), est construite en considérant que la probabilité d'une mutation est de 1 pour 1000 acides aminés. on obtient une matrice XPAM en la multipliant par elle-même. Les matrices de type PAM résultent d'un alignement global de protéines voisines et représentent un remplacement possible et acceptable d'un acide aminé avec un autre au cours de l'évolution des protéines.

[illegible]

Figure 26 : La matrice PAM 250

- **La matrice BLUSOM 62 Henikoff et Henikoff (1992)**

Cette matrice a été élaborée en examinant les substitutions d'acides aminés dans des blocs qui représentent des alignements multiples de segments de séquences sans Gaps, ainsi que dans les zones les plus conservées des familles de protéines. Ces auteurs ont utilisé plus d'une centaine de familles de protéines pour un nombre total de blocs supérieur à 2000. Les hypothèses de calcul de la matrice sont identiques à celles de calcul du PAM.

De nombreux programmes d'alignement des séquences adoptent les matrices de substitution et offre un choix pour leur utilisation. La plupart des programmes d'alignement ou de similarité de séquences qui utilisent des matrices de substitution offrent le choix pour cette dernière matrice.

	E	V	L	V	M	D	H	L	F
D									-6
L									2
L									2
V									-1
E									-5
R									-4
L									-1
F	-7	-6	-2	-6	-4	-7	-3	-2	0

On obtient la matrice transformée suivante :

	E	V	L	V	M	D	H	L	F
D	3	-2	-4	-2	-3	-4	1	-4	-6
L	-3	2	6	2	4	-4	-2	6	2
L	-3	2	6	2	4	-4	-2	6	2
V	-2	4	2	4	2	-2	-2	2	-1
E	4	-2	-3	-2	-2	3	1	-3	-5
R	-1	-2	-3	-2	2	-1	2	-3	-4
L	-3	2	6	2	4	-4	-2	6	-5
F	-7	-6	-2	-6	-4	-7	-3	-2	9

Dans une troisième étape, on va parcourir de la matrice transformée du plus haut score vers le plus petit.

	E	V	L	V	M	D	H	L	F
D	35	-2	-4	-2	-3	-4	1	-4	-6
L	-3	32	30	26	24	14	13	15	2
L	-3	26	30	24	24	14	13	15	2
V	-2	24	22	24	22	16	13	7	-1
E	4	15	14	15	15	20	18	6	-5
R	-1	15	12	13	17	16	17	6	-4
L	6	11	15	11	13	5	6	15	-1
F	-5	-1	2	-1	0	-6	-2	2	9

La quatrième étape : alignement de deux séquences

Séquence S1	E	V	L	V	M	D	H	L	F
					*				
Séquence S2	D	L	L	V	-	E	R	L	F

Le score global de cet alignement est : $s = 182$

D'après l'alignement précédent, on peut relever 4 identités et 4 substitutions (similarités). La substitution observée à la première position d'alignement, représente un remplacement de l'acide aminé E par D par nécessité adaptative. Le même constat pour les autres substitutions.

Remarque :

La matrice BLOSUM est devenue une matrice de référence, plus que PAM, notamment BLOSUM62 qui était utilisé avant BLAST.

Les programme BLAST, hébergé sur le serveur NCBI, permettent des recherches de similarité de choix de protéines entre BLOSUM 80, 62, 45, PAM 70, 30 et propose BLOSUM 62.

2.1.4. Le programme BLAST

Le BLAST (Basic Local Alignment Tool) est un algorithme de comparaison de séquences, met en un alignement local entre une séquence requête (query sequence) et chacune des séquences d'une banque de données (Nucleotide, UniProtKB, PDB).

Pour pouvoir effectuer cette tâche énorme dans un temps raisonnable, BLAST se base sur une approche heuristique: les séquences de la base de données sont préalablement indexées dans un "dictionnaire de mots", qui dresse la liste des séquences de la base de données contenant chaque oligomère (oligopeptide pour les bases de données de protéines, oligonucléotides pour les séquences nucléiques) d'une taille donnée.

Quand la recherche est lancée, le BLAST commence par analyser la séquence requête en dressant des séquences présentes dans la base de données. Il consulte ensuite le dictionnaire pour extraire la liste des séquences de la base de données qui contiennent ces mots, et lance un alignement par paire avec ce sous-ensemble des séquences.

Cette heuristique est plus rapide que les méthodes d'alignement par paire par programmation dynamique (Needleman- Wunsch en alignement global, Smith-Waterman en alignement local), mais elle présente un certain risque de louper des similarités.

- **L'utilisation du programme Blast**

Nous pouvons utiliser ce programme directement en consultant le portail NCBI : <https://blast.ncbi.nlm.nih.gov/Blast.cgi>.

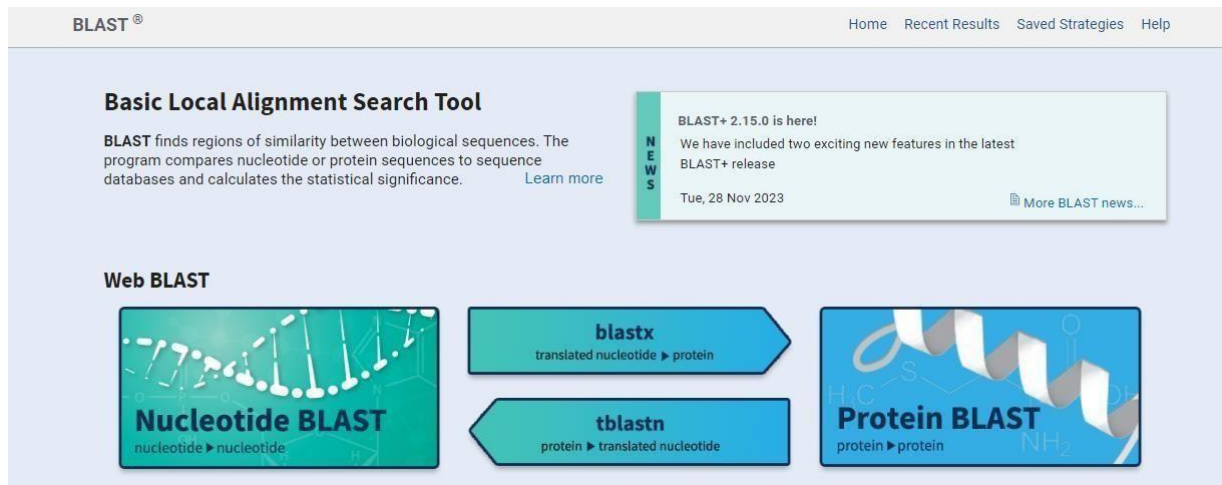


Figure 28 : L'interface du programme BLAST

- Les étapes d'alignement par le programme BLAST

Pour pouvoir lancer le programme, il suffit de choisir le type comparaison désirée, il s'agit d'un alignement entre des séquences de nature nucléiques ou protéiques.

NCBI/BLAST/blastn suite

blastn blastp blastx tblastn tblastx

Enter Query Sequence

Enter accession number, gi, or FASTA sequence

BM173041.1

1 : Séquence à étudier

Or, upload file

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)

2 : Choix de la base de comparaison

Organism Optional

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude Optional

Entrez Query Optional

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

3 : Lancer l'algorithme d'alignement

BLAST

Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly simila

Ensuite, l'utilisateur doit insérer la séquence requête.

Après avoir lancé l’alignement, nous obtiendrons les résultats illustrés dans le tableau suivant :

Tableau 2 : Récapitulation des résultats de BLAST

Résultats	Description																																																																																										
Distribution of 100 Blast Hits on the Query Sequence Mouse over to see the define, click to show alignments Color key for alignment scores <40 40-50 50-80 80-200 >=200 Query 1 30 60 90 120 150	Les traits de couleur correspondent à un alignement entre la séquence soumise et une séquence de la banque. La couleur correspond au score. La longueur correspond à la taille de l'alignement.																																																																																										
Sequences producing significant alignments Download Select columns Show 100 ☐ select all 0 sequences selected <table><thead><tr><th>Description</th><th>Scientific Name</th><th>Max Score</th><th>Total Score</th><th>Query Cover</th><th>E value</th><th>Per Ident</th><th>Acc. Len</th><th>Accession</th></tr></thead><tbody><tr><td>☐ Adylosia californica insulin precursor (PIN) mRNA</td><td>Adylosia californica</td><td>1788</td><td>1788</td><td>100%</td><td>0.0</td><td>100.00%</td><td>968</td><td>NM_001204686.1</td></tr><tr><td>☐ Halotis corrugata insulin mRNA, partial cds</td><td>Halotis corrugata</td><td>505</td><td>505</td><td>28%</td><td>1e-137</td><td>100.00%</td><td>273</td><td>FJ536845.1</td></tr><tr><td>☐ Pammea aurita genome assembly, chromosome 23</td><td>Pammea aurita</td><td>65.8</td><td>65.8</td><td>4%</td><td>2e-05</td><td>97.44%</td><td>24241859</td><td>OJ352315.1</td></tr><tr><td>☐ Yoonomeuta clumbellus genome assembly, chromosome 25</td><td>Yoonomeuta clumbellus</td><td>60.2</td><td>60.2</td><td>3%</td><td>0.001</td><td>100.00%</td><td>14468000</td><td>OJ371216.1</td></tr><tr><td>☐ Pammea aurita genome assembly, chromosome 2</td><td>Pammea aurita</td><td>60.2</td><td>120</td><td>3%</td><td>0.001</td><td>100.00%</td><td>49731725</td><td>OJ352281.1</td></tr><tr><td>☐ Pammea aurita genome assembly, chromosome 6</td><td>Pammea aurita</td><td>58.4</td><td>58.4</td><td>3%</td><td>0.004</td><td>100.00%</td><td>37834380</td><td>OJ352289.1</td></tr><tr><td>☐ Pammea aurita genome assembly, chromosome 1</td><td>Pammea aurita</td><td>58.4</td><td>58.4</td><td>3%</td><td>0.004</td><td>100.00%</td><td>55376203</td><td>OJ352280.1</td></tr><tr><td>☐ Eilema depressum genome assembly, chromosome 7</td><td>Eilema depressum</td><td>58.4</td><td>58.4</td><td>3%</td><td>0.004</td><td>100.00%</td><td>23088918</td><td>OJ812019.1</td></tr><tr><td>☐ Pandemis conyana genome assembly, chromosome 2</td><td>Pandemis conyana</td><td>56.5</td><td>56.5</td><td>3%</td><td>0.014</td><td>100.00%</td><td>37479866</td><td>OJ421850.1</td></tr></tbody></table>	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per Ident	Acc. Len	Accession	☐ Adylosia californica insulin precursor (PIN) mRNA	Adylosia californica	1788	1788	100%	0.0	100.00%	968	NM_001204686.1	☐ Halotis corrugata insulin mRNA, partial cds	Halotis corrugata	505	505	28%	1e-137	100.00%	273	FJ536845.1	☐ Pammea aurita genome assembly, chromosome 23	Pammea aurita	65.8	65.8	4%	2e-05	97.44%	24241859	OJ352315.1	☐ Yoonomeuta clumbellus genome assembly, chromosome 25	Yoonomeuta clumbellus	60.2	60.2	3%	0.001	100.00%	14468000	OJ371216.1	☐ Pammea aurita genome assembly, chromosome 2	Pammea aurita	60.2	120	3%	0.001	100.00%	49731725	OJ352281.1	☐ Pammea aurita genome assembly, chromosome 6	Pammea aurita	58.4	58.4	3%	0.004	100.00%	37834380	OJ352289.1	☐ Pammea aurita genome assembly, chromosome 1	Pammea aurita	58.4	58.4	3%	0.004	100.00%	55376203	OJ352280.1	☐ Eilema depressum genome assembly, chromosome 7	Eilema depressum	58.4	58.4	3%	0.004	100.00%	23088918	OJ812019.1	☐ Pandemis conyana genome assembly, chromosome 2	Pandemis conyana	56.5	56.5	3%	0.014	100.00%	37479866	OJ421850.1	Ici chaque ligne correspond à trait coloré de la représentation graphique. On trouve dans ce tableau, plusieurs paramètres pour juger la crédibilité de l'alignement, tels que, le % d'identité, l'E-value et le score total.
Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per Ident	Acc. Len	Accession																																																																																			
☐ Adylosia californica insulin precursor (PIN) mRNA	Adylosia californica	1788	1788	100%	0.0	100.00%	968	NM_001204686.1																																																																																			
☐ Halotis corrugata insulin mRNA, partial cds	Halotis corrugata	505	505	28%	1e-137	100.00%	273	FJ536845.1																																																																																			
☐ Pammea aurita genome assembly, chromosome 23	Pammea aurita	65.8	65.8	4%	2e-05	97.44%	24241859	OJ352315.1																																																																																			
☐ Yoonomeuta clumbellus genome assembly, chromosome 25	Yoonomeuta clumbellus	60.2	60.2	3%	0.001	100.00%	14468000	OJ371216.1																																																																																			
☐ Pammea aurita genome assembly, chromosome 2	Pammea aurita	60.2	120	3%	0.001	100.00%	49731725	OJ352281.1																																																																																			
☐ Pammea aurita genome assembly, chromosome 6	Pammea aurita	58.4	58.4	3%	0.004	100.00%	37834380	OJ352289.1																																																																																			
☐ Pammea aurita genome assembly, chromosome 1	Pammea aurita	58.4	58.4	3%	0.004	100.00%	55376203	OJ352280.1																																																																																			
☐ Eilema depressum genome assembly, chromosome 7	Eilema depressum	58.4	58.4	3%	0.004	100.00%	23088918	OJ812019.1																																																																																			
☐ Pandemis conyana genome assembly, chromosome 2	Pandemis conyana	56.5	56.5	3%	0.014	100.00%	37479866	OJ421850.1																																																																																			

```
>[gb|ADN04687.1| RNA polymerase beta I subunit [Panax ginseng]
gb|ADN04688.1| RNA polymerase beta I subunit [Panax japonicus var. bipinnatifidus]
gb|ADN04690.1| RNA polymerase beta I subunit [Panax japonicus var. bipinnatifidus]
>16 more sequence titles
Length=186
```

```
Score = 338 bits (866), Expect = 2e-91, Method: Compositional matrix adjust.
Identities = 165/170 (97%), Positives = 169/170 (99%), Gaps = 0/170 (0%)
```

```
Query 1 EGRFRETMLGKRVDSGRSVIVVGPSLSLHRCGLPREIAIELFQTFVIRGLIRQHLASNI 60
EGRFRET+LGKRVDSGRSVIVVGPSLSLHRCGLPREIAIELFQTFVIRGLIRQHLASNI
Sbjct 17 EGRFRETLLGKRVDSGRSVIVVGPSLSLHRCGLPREIAIELFQTFVIRGLIRQHLASNI 76

Query 61 GVAKSKIREKEPIVWGILQEVMRGHPILLNRAPTLHRLGIQAFQPIIVEGRAICLHPLVR 120
GVAKSKIREKEPIVW ILQEVW+GHP+LLNRAPTLHRLGIQAFQPI+LVEGRAICLHPLVR
Sbjct 77 GVAKSKIREKEPIVWEILQEVMQGHPVLLNRAPTLHRLGIQAFQPIVVEGRAICLHPLVR 136

Query 121 KGFNADFDGDQMAVHVPLSLEAQAEARLLMFSHMNLSPAIGDPISVPTQ 170
KGFNADFDGDQMAVHVPLSLEAQAEARLLMFSHMNLSPAIGDPISVPTQ
Sbjct 137 KGFNADFDGDQMAVHVPLSLEAQAEARLLMFSHMNLSPAIGDPISVPTQ 186
```

```
>[gb|ACB88372.1| RNA polymerase C [Populus alba]
Length=167
```

```
Score = 338 bits (866), Expect = 2e-91, Method: Compositional matrix adjust.
Identities = 167/167 (100%), Positives = 167/167 (100%), Gaps = 0/167 (0%)
```

```
Query 5 RETMLGKRVDSGRSVIVVGPSLSLHRCGLPREIAIELFQTFVIRGLIRQHLASNIGVAK 64
RETM LGKRVDSGRSVIVVGPSLSLHRCGLPREIAIELFQTFVIRGLIRQHLASNIGVAK
Sbjct 1 RETMLGKRVDSGRSVIVVGPSLSLHRCGLPREIAIELFQTFVIRGLIRQHLASNIGVAK 60

Query 65 SKIREKEPIVWGILQEVMRGHPILLNRAPTLHRLGIQAFQPIIVEGRAICLHPLVRKGFN 124
SKIREKEPIVWGILQEVMRGHPILLNRAPTLHRLGIQAFQPIIVEGRAICLHPLVRKGFN
Sbjct 61 SKIREKEPIVWGILQEVMRGHPILLNRAPTLHRLGIQAFQPIIVEGRAICLHPLVRKGFN 120

Query 125 ADFDGDQMAVHVPLSLEAQAEARLLMFSHMNLSPAIGDPISVPTQE 171
ADFDGDQMAVHVPLSLEAQAEARLLMFSHMNLSPAIGDPISVPTQE
Sbjct 121 ADFDGDQMAVHVPLSLEAQAEARLLMFSHMNLSPAIGDPISVPTQE 167
```

-Query : la séquence soumise (requête)

-Subject : la séquence de la base de données

-Les paramètres propres à chaque alignement : (score, E-value, % d'identité et nombre de Gaps.

Bibliographie

Bibliographie

- Acques van Helden. (2015). Cours en ligne. Introduction à la bio-informatique. Université de Marseille. http://pedagogix-tagc.univ-mrs.fr/courses/bioinfo_intro.
- Carbone, A. "Algorithmes sur les arbres et les graphes en bio-informatique.". Université Pierre et Marie Curie. Cours en ligne, consulté le 20.04.2020. http://www.ihes.fr/~carbone/L3_AAGB_genome_rearrangement.pdf.
- Cours de Bioinformatique pour les 3ème année LMD Génomique et Biotechnologies Végétales, Pr. DJEKOUN Abdelhamid et le Dr. HAMIDECHI Mohamed Abdelhafid (https://www.umc.edu.dz/index.php/ar/vf/images/cours/cours1_bioinformatique1.pdf).
- Deléage, G & Manolo, G. (2015). Bio-informatique-2e édition: Cours et applications. Dunod, 216 pages.
- Hachez, D & Pavel, A. (2007). Bio-informatique moléculaire: Une approche algorithmique. Springer, 329 pages.
- Heusser, S & Dupuy, H. (2008). Atlas de biologie animale, Dunod, 152 pages.
- Nguyen, K., Guo, X., & Pan, Y. (2016). Multiple biological sequence alignment: scoring functions, algorithms and evaluation. John Wiley & Sons. Georgia state universitu. 172 pages.
- Pavlopoulou, A., & Michalopoulos, I. (2011). State-of-the-art bioinformatics protein structure prediction tools. International journal of molecular medicine, 28(3), 295-310.
- Rincé, A et La carbone, S. (2010). Cours en ligne. Méthodes de dépistage des sources de pollution microbienne, les marqueurs bactériens. Université de Caen 14032 CAEN Cedex. 24p.
- Ruchi Singh. (2015). Bioinformatics: Genomics and Proteomics, VIKAS., ISBN: 978-93-259-7855-3. 277 pages.