1. LES DOMAINES D'APPLICATIONS DES METHODES D'IDENTIFICATION MOLECULAIRES

Les domaines dans lesquels les techniques moléculaires peuvent être utilisées pour la caractérisation d'organismes et/ou de gènes sont divers. Quel que soit ce domaine d'application, ces techniques offrent souvent une alternative pour accélérer ou préciser la détection d'espèces et de gènes.

1. 1. L'industrie agro-alimentaire

L'évolution des réglementations sur l'alimentation humaine et animale crée de nouveaux besoins et impose de nouvelles contraintes à l'industrie agro-alimentaire. L'identification moléculaire permet de tracer l'origine des aliments (OGM par exemple) ou de détecter la présence éventuelle d'agents pathogènes (par exemple, les Salmonelles sont couramment détectées par PCR (Van Kessel, Karns and Perdue, 2003).

1.2. L'industrie pharmacologique

Les programmes de recherche des industries pharmaceutiques demandent une forte capacité de traitement en criblage médicamenteux (Lacroix et al., 2002). Il est courant pour un laboratoire de recherche de tester de l'ordre de 10 000 molécules par jour. La disponibilité de puces à ADN performantes et peu coûteuses permet d'accélérer considérablement la vitesse de découverte des nouveaux médicaments, donc d'en réduire le coût.

1.3. La recherche en génomique

La génomique est l'étude de l'ensemble des gènes des organismes vivants, de leur disposition sur les chromosomes, de leur séquence et de leur fonction. L'objectif est de réaliser l'inventaire des gènes qui s'expriment dans un type cellulaire donné, à un instant donné et dans un environnement donné. La révolution des puces à ADN est fortement liée au projet à très grande échelle de décodage du génome humain, initié par le Human Genome Program (HGP) (Watson and Jordan, 1989; Grisolía, 1991).

1.4. Le bio-terrorisme

Le bio-terrorisme est défini comme étant l'usage de micro-organismes dans l'intention de causer la mort. Étant donné qu'il est plus facile aujourd'hui de se procurer des agents

biologiques ou l'information technique nécessaire pour les produire, le bio-terrorisme pourrait être une arme de choix. Si depuis 1972 l'usage d'armes biologiques est prohibé par la majorité de la communauté internationale, le risque est toujours présent comme peut en témoigner la dissémination intentionnelle de spores de la bactérie du charbon en 2001 aux États-Unis (Kottow, 2003). Dans le cas de bio-terrorisme la rapidité d'identification de l'organisme pathogène est alors primordiale.

1.5. L'environnement

Dans le domaine environnemental, la diversité spécifique est très importante. De plus, dans le cas de certains micro-organismes ou champignons, beaucoup d'espèces sont difficilement cultivables. Les techniques moléculaires sont alors particulièrement bien adaptées. L'environnement contient un très grand nombre d'espèces inconnues (plus de 90% pour les bactéries). Les applications dans ce domaine sont très nombreuses et diverses. On distingue deux grands types d'applications: soit une approche populationnelle afin de caractériser un milieu (O'brien et al., 2005; Poretsky et al., 2005; Wu et al., 2002) soit on recherche des organismes spécifiques dans un échantillon.

2. CHOIX DES GENES CIBLES

2.1. L'ARNr 16S

Des études sur *E. coli* ont montré que le choix de la cible était un facteur important d'une bonne identification des bactéries viables présentes dans un échantillon (Norton and Batt, 1999; Yaron and Matthews, 2002).

A l'origine le choix de la molécule ARNr 16S ou 23S pour des études phylogéniques fut déterminé par des raisons techniques. La taille plus importante et la présence de structures secondaires particulièrement marquées dans l'ARNr 23S ont longtemps rendu son clonage et son séquençage difficiles. Pour ces raisons, l'ARNr 16S a été choisi comme index phylogénique et en particulier pour la phylogénie des procaryotes. Ceci a permis la construction de la plus grande banque de données actuelle avec un rythme d'accroissement supérieur à celui de toutes autres molécules. C’est donc le gène "classique" pour ce genre de travail et pour les seules Bacteria, on dispose maintenant de plus de 174 000 séquences (Septembre 2005).

L'ARN 16S a une structure particulière faite d'une succession de domaines dont les vitesses d'évolution sont très variables, de relativement élevée à presque nulle. Chacun de ces domaines a son importance pour l'identification moléculaire des micro-organismes. Certaines parties sont identiques chez toutes les bactéries et sont donc utilisables comme sites d'hybridation pour des amorces universelles de séquences. La comparaison des domaines conservés permet de retracer les liens de parenté qui unissent des bactéries éloignées, tandis que les domaines à vitesse d'évolution plus rapide permettent l'étude des relations phylogénétiques d'espèces plus proches. D'autres parties de séquences sont propres à un groupe et permettent ainsi l'identification de séquences dites signatures caractéristiques d'ordres taxonomiques différents (espèce, genre, famille ou classe).

Une autre raison qui a motivé le choix de cette molécule est la quantité de séquences présentes dans les bases publiques. Elle est telle qu'il est maintenant quasiment impossible d'utiliser une autre molécule et d'obtenir des résultats équivalents. De plus, c'est un gène ubiquitaire, bien conservé, soumis à des contraintes fortes et stables et relativement non affecté par les changements du milieu extérieur (Woese, Kandler and Wheelis, 1990).

Malgré son usage extrêmement répandu dans les études des bactéries et ses avantages, ce gène possède quelques limitations. La principale est son manque de pouvoir résolutif entre espèces proches (Achenbach, Carey and Madigan, 2001).

2.2. Le facteur d'élongation alpha ou gène « tuf »

Le résultat d'une étude préalable sur la faisabilité du projet Aquachip a montré que l'ARNr 16S ne présente pas assez de variabilité par rapport à la diversité bactérienne pour assurer une identification précise de toutes les bactéries. Pour pallier à ce problème, il a été nécessaire de travailler sur un autre gène : le gène tuf.

Ce gène code pour le facteur d'élongation alpha, impliqué dans la synthèse des protéines. Il facilite l'élongation des chaînes polypeptidiques lors de la traduction. Il est présent dans toutes les bactéries (Sela et al., 1989). Ces caractéristiques font de ce gène un bon candidat pour les études phylogéniques (Ludwig et al., 1994).

Ce gène a été choisi aussi parce qu'il présente des domaines nucléotidiques ayant des vitesses d'évolution plus importantes que celles constatées dans l'ARNr 16S.

2.3. Les facteurs de pathogénicité

L'utilisation d'un marqueur universel comme l'ARNr 16S ou tuf n'est cependant pas suffisante. La principale raison est qu'il existe des bactéries qui ne sont pathogènes que si un facteur de pathogénicité est présent dans leur génome ou dans un plasmide. Nous devons donc utiliser ces facteurs de pathogénicité comme marqueurs (portés par les bactéries pathogènes), en plus de l'ARNr 16S ou de tuf.

Les gènes de pathogénicité sont spécifiques d'un groupe de bactéries et ont généralement un pouvoir résolutif plus important (Chang et al., 2001). Ils permettent également de mettre en évidence une activité physiologique particulière, parfois différente entre des espèces proches phylogénétiquement et permettent alors de résoudre des problèmes de diversité entre espèces proches (Bourne, Mcdonald and Murrell, 2001).

Malheureusement ces gènes de pathogénicité ne sont pas exempts de défauts. En effet, La mise au point d'amorces universelles d'amplification se heurte à deux problèmes :

- la vitesse d'évolution de ces gènes est élevée ce qui réduit le nombre de domaines dans lesquels il existe des parties conservées.
- le manque de données de séquences ne permet pas de calculer des amorces efficaces pour toutes les variantes possibles.

Ces problèmes seront résolus en partie au fur et à mesure que de nouveaux gènes seront clonés et séquencés.

3. Ressources Bioinformatiques et Bases de Données

L'European Molecular Biology Laboratory (EMBL – http://www.embl-heidelberg.de) entretient une base donnée nucléotidique qui est mise à jour quotidiennement mais aussi fournit d'autres sources bioinformatiques. L'European Bioinformatics Institute (EBI– http://www.ebi.ac.uk/ebi_home.html) maintient la base de donnée des séquences protéiques SwissProt et Sequence Retrieval System (SRS – http://srs.ebi.ac.uk/). Le Biomolecular Structure and Modeling (BSM) est supporté par l'Université Collège de London.

Le National Center for Biotechnology Information (NCBI – http://www.ncbi.nlm.nih.gov/) est la hôte du GenBank et la base de données des séquences d'ADN du National Institutes of

Health (NIH). Aussi, NCBI maintient le système ENTREZ (http://www.ncbi.nlm.nih.gov/Entrez/) qui donne accès aux données de biologie moléculaire et les articles. D'autre part, ENTREZ donne accès aux séquences d'ADN du GenBank, EMBL, DDBJ (DNA data base of Japan), aussi aux séquences protéiques du SWISS-PROT.

4. L'alignement : Pairwise Alignment

4.1. Motivation

Si une nouvelle séquence est obtenue à partir du séquençage génomique, la première étape est la recherche de similarités avec des séquences connues dans d'autres organismes. Si la fonction/structure des séquences similaires/protéines est connue, très probablement (highly likely) la nouvelle séquence correspond à une protéine avec la même fonction/structure. En effet, il a été trouvé que seulement à peu près 1% des gènes humains n'ont pas de contrepartie dans le génome de souris et que la moyenne de similarité entre les gènes de la souris et de l'homme est de 85%.

Les similarités existent parce que toutes les cellules possèdent une cellule ancêtre commune (a mother cell). Donc, dans les différents organismes il pourrait avoir des mutations d'acides aminés dans certaines protéines parce que les acides aminés ne sont pas tous importants pour la fonction et peuvent être remplacés par des acides aminés qui ont des caractéristiques chimiques semblables sans changer la structure. Parfois les mutations sont tellement nombreuses qu'il est difficile de trouver des similarités.

La méthode du calcul des fonctions des gènes par similarités est appelée la *génomique comparative* ou la *recherche d'homologie*. Deux séquences sont homologues lorsqu'ils ont comme racine un ancêtre commun.

4.2. Les similarités de séquences et score : La matrice d'identité

Pour les séquences biologiques, il est connu comment une séquence peut mutée en une autre. Premièrement, il y'a les *points de mutation* ou un nucléotide ou acide aminé est changé en un autre. Deuxièmement, il y'a les *suppressions* ou un élément (nucléotide ou acide aminé) ou

une subséquence entière d'un élément est supprimée de la séquence. Troisièmement, il y'a les *insertions* ou un élément ou une subséquence est insérée dans la séquence.

Pour notre première approche de similarité de deux séquences biologiques, elle peut être exprimée par le nombre minimal de mutations pour transformer une séquence en une autre, exemple :

BIOINFORMATICS → BIOI-N-FORMATICS

BOILING FOR MANICS → B-OILINGFORMANICS

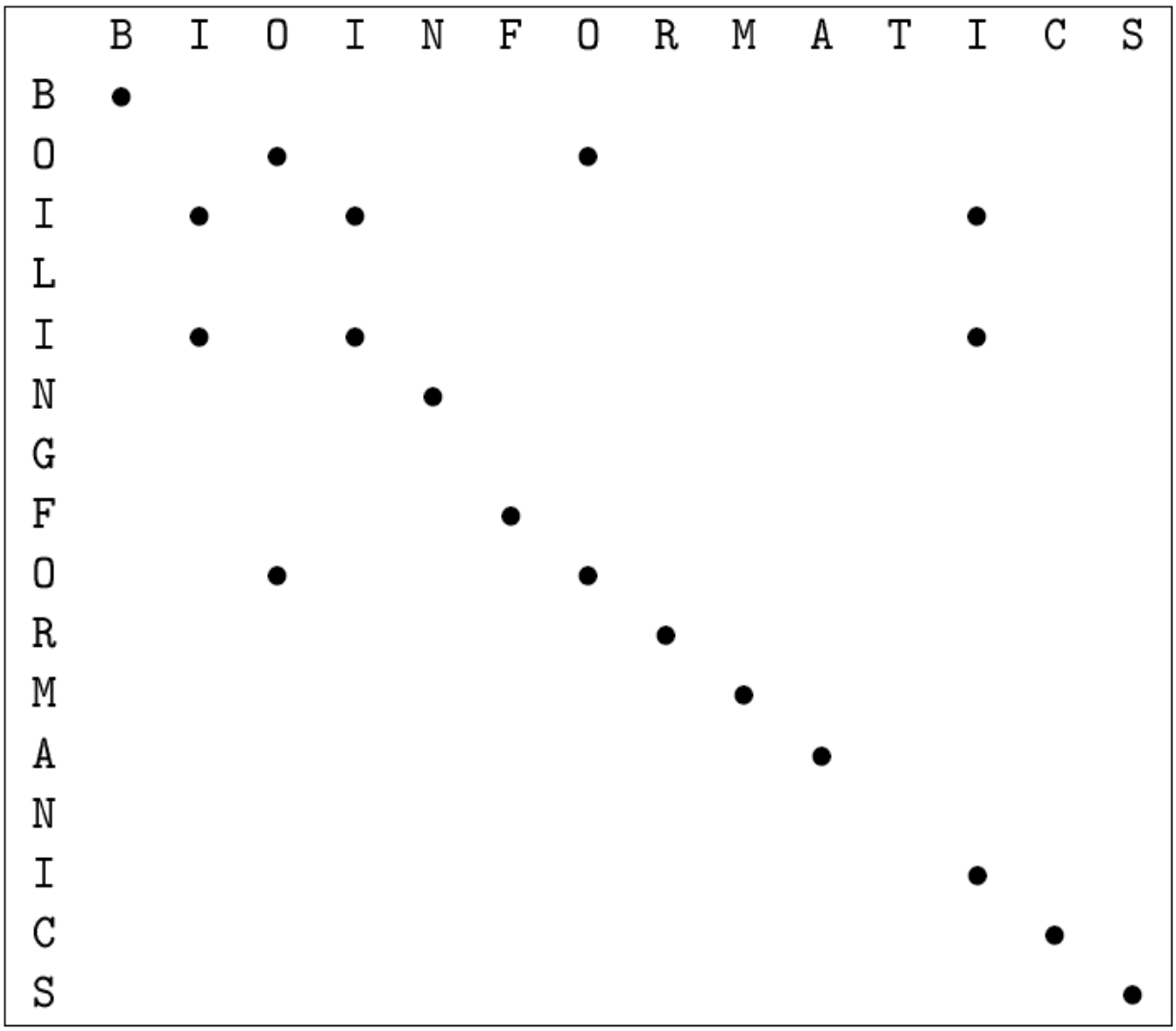
Le conte donne 12 lettres identiques sorties des 14 lettres de BIOINFORMATICS. Les mutations pourraient êtres :

(1) suppression I	BOINFORMATICS
(2) insertion LI	BOILINFORMATICS
(3) insertion G	BOILINGFORMATICS
(4) changement de T en N	BOILINGFORMANICS

Les deux textes semblent très similaires. Noter que l'insertion ou la suppression ne peuvent pas être distinguées si les deux séquences sont présentées (es que le I est supprimé de la première séquence ou inséré dans la seconde ?). Donc, les deux cas sont dénotées par "-".

La tache des algorithmes bioinformatiques est de trouver à partir de deux séries (la partie à gauche dans l'exemple au-dessus) l'alignement optimal (la partie à droite dans l'exemple au-dessus). L'alignement optimal est l'arrangement des deux séries d'une manière ou le nombre de mutations est minimal.

La matrice de dot (Dot Matrix) est un outil de représentation des alignements, ou une séquence est écrite horizontalement en haut et l'autre verticalement à gauche. Ce qui donne une matrice ou chaque lettre de la première séquence est couplée avec chaque lettre de la deuxième séquence. Pour chaque correspondance de lettres un point (dot) est inscrit dans la position concordante dans la matrice. Quelles paires apparaissent dans l'alignement optimal ? On va voir ci-après que chaque chemin à travers la matrice correspond à un alignement.



Règles : vous pouvez bouger horizontalement "→", verticalement "↓", et vous pouvez bouger seulement diagonalement "↘" si vous êtes dans la position de dot.

Tache : faite le plus possible de mouvements diagonaux quand vous bougez du coint le plus haut à gauche au coint le plus bas à droite.

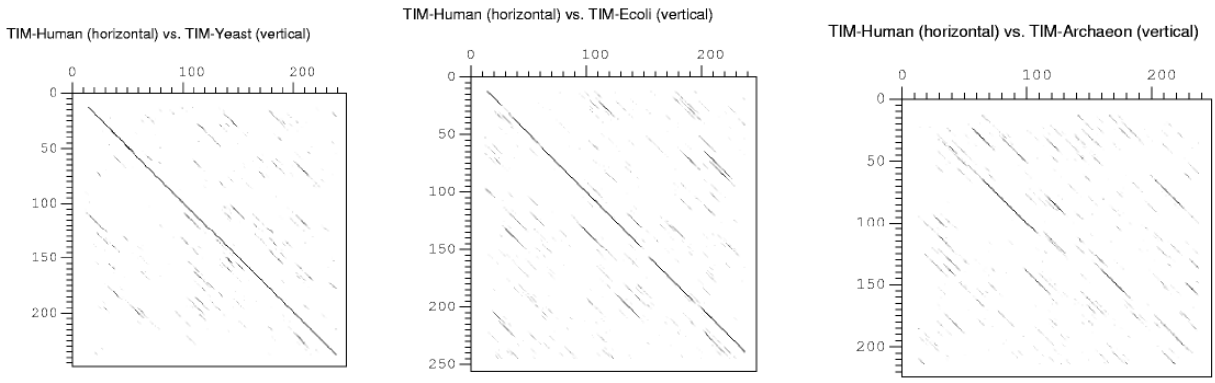
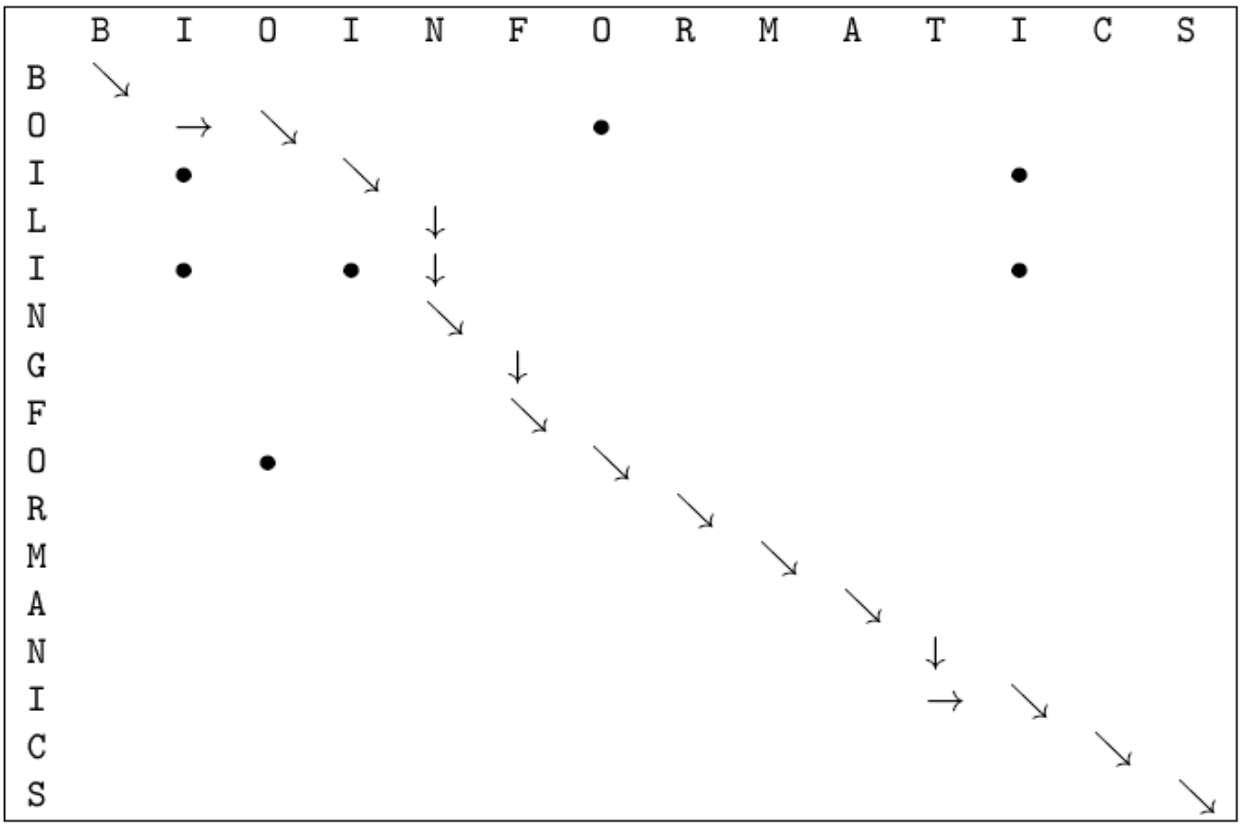


Figure : Matrice de dot de la triosephosphate isomérase humaine avec la même protéine dans la levure, *E. coli* et *Archaeon.* La levure donne la meilleure correspondance car la diagonale est presque complète. *E. coli* a quelques fractures dans la diagonale. *Archaeon* montre la similarité la plus faible. Cependant, la structure 3D et la fonction est la même pour toutes les protéines.



Le nombre de mouvements diagonaux“↘ ” représente les correspondances et le nombre de scores, “→” correspond à “-” dans la séquence verticale, “↓” à “-” dans la séquence horizontale et la combinaison “→↓” ou “↓→” correspond à une divergence. Donc, chaque chemin à travers la matrice correspond à un alignement et chaque alignement peut être exprimé par un chemin dans la matrice.

Dans l'exemple au-dessus les dot sur les diagonales correspondent aux régions de correspondances (similarités). Dans la figure des Matrices Dot pour la comparaison de la protéine triosephosphate isomérase (TIM) humaine avec celle de la levure, *E. coli* et *Archaeon.* Pour la levure la diagonale est complète et pour *E. coli* de petits trous « gaps » sont visibles, mais *Archaeon* ne montre pas une diagonale étendue. Donc, la TIM humaine correspond le plus avec la TIM de la levure, suivie par la TIM d'*E. coli* et possède la similarité la plus faible avec la TIM d'*Archaeon.*

5. Exemple d'Analyse bioinformatique des séquences : cas de gène ARNr 16S

L'analyse bioinformatique des séquences ADNr 16S est réalisée sur le site miroir du NCBI qui offre la possibilité d'aligner nos séquences ADNr 16S avec celles de la banque nucléique (GenBank en l'occurrence).

BLAST, un outil d'alignement local du NCBI, est un programme en ligne libre de droits permettant de rechercher des séquences similaires dans une banque de séquences à partir d'une séquence requête (query) d'ADN. Le programme BLAST contient six programmes différents d'alignement des séquences (appelés parfums de BLAST). BLASTN (N pour nucleotide) est l'un des différents parfums de Blast et est utilisé pour comparer une séquence requête nucléique à une banque de séquences nucléiques.

L'utilisation des différents types de parfums de Blast dépond de la nature de la séquence requête ; en effet :

- Blastn : compare une séquence requête nucléique contre une base de séquences nucléiques,
- Blastp : compare une séquence requête protéique contre une base de séquences protéiques (SwissProt par exemple),
- Blastx : compare une séquence requête nucléique traduite selon six phases de lectures (ORF : Open Reading Frame : cadres de lecture ouverts) contre une banque de séquences protéiques,

- Tblastn : Compare une séquence requête protéique contre une banque de séquences nucléiques traduites selon les six phases de lecture,
- Tbalstx : compare une séquence requête nucléique traduite en six phases de lecture contre une banques de séquences nucléiques traduites en six phases de lecture.

Après alignement de chacune de nos séquences ADNr 16S avec l'outil BLASTN, seules les séquences de GenBank ayant une similarité (avec notre séquence) supérieure ou égale 99% et une E-value nulle ont été retenues car la définition moléculaire du genre stipule que les homologies des séquences des ADNr 16S doivent être supérieures ou égale à 97%. Une homologie supérieure ou égale à 99% traduit l'appartenance à une même espèce. Alors qu'un score d'homologie inférieur à 97% ne permet pas l'identification.

Dans le cas où plusieurs séquences sont proposées par la banque, et qui ont la même E-value et le même pourcentage d'identité, nous avons tranché pour la séquence ayant présenté le meilleur score d'alignement avec notre séquence requête.

Dix souches ont été séquencées. BlastN a permis de comparer nos séquences à l'ensemble des séquences existantes sur GenBank. Les résultats du blasting sont les suivants :

Tableau 1 : Confrontation et correspondance biomoléculaire avec GenBank.

Séquence inconnue	Souche proposée par GenBank	Score	E-value	% identité
1	*Enterococcus faecalis strain JCM 5803*	1223	0	99
2	*Enterococcus faecalis strain JCM 5803*	1201	0	99
3	*Enterococcus faecalis strain JCM 5803*	1225	0	99
4	*Enterococcus faecalis strain JCM 5803*	1199	0	99
5	*Enterococcus camelliae strain FP15-1*	752	0	99
6	*Acinetobacter calcoaceticus NCCB 22016*	1131	0	100
7	*Aeromonas veronii*	1186	0	99
8	*Pseudomonas aeruginosa strain DSM 50071*	1074	0	98
9	*Bacillus mojavensis strain IFO15718*	1105	0	99
10	*Stenotrophomonas maltophilia*	1452	0	99

Nous constatons que pour toutes les séquences inconnues qui sont les nôtres, GenBank, via le programme BlastN réalise un alignement en utilisant ses propres séquences et prose celle qui présente la meilleure identité avec la nôtre en calculant un score qui correspond au

nombre de nucléotides identiques chez les deux séquences. Ce score peut être traduit sous forme de pourcentage d'identité (%id). La valeur calculée de E-value indique la probabilité que le résultat de cet alignement a eu lieu par hasard. Donc plus cette valeur est proche du zéro et mieux c'est. Or tous les alignements ont abouti à des valeurs nulles de la E-value ; ce qui exprime que les identités retrouvées entre nos séquences et celles proposées par GenBank ne sont pas dues au hasard.

Sur les 10 souches alignées, quatre d'entre elles sont *Enterococcus faecalis strain JCM 5803* et une seule correspond à l'espèce *Enterococcus camelliae strain FP15-1*. Les autres souches ont été identifiées à *Acinetobacter calcoaceticus NCCB 22016, Aeromonas veronii, Pseudomonas aeruginosa strain DSM 5007, Bacillus mojavensis strain IFO15718* et *Stenotrophomonas maltophilia.*

La figure 1 représente un alignement de la séquence partiel du gène ARNr 16S d'*Aeromonas veronii* obtenue sur GenBank, via le programme BlastN.

>*Aeromonas veronii*

TACTTTTGCCGGCGAGCGGCGGACGGGTGAGTAATGCCTGGGGATCTGCCCAGTCGAGGGGGATAACTACTGGAA
ACGGTAGCTAATACCGCATACGCCCTACGGGGGAAAGCAGGGGACCTTCGGGCCTTGCGCGATTGGATGAACCCA
GGTGGGATTARCTAGTTGGTGAGGTAATGGCTCACCAAGGCGACGATCCCTARCTGGTCTGAGAGGATGATCAGC
CACACTGGAACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATATTGCACAATGGGGGAAACCC
TGATGCMGCCATGCCGCGTGTGTGAAGAAGGCCTTCGGGTTGTAAAGCACTTTCAGCGAGGAGGAAAGGTTGGTA
GCTAATAACTGCCAGCTGTGACGTTACTCGCAGAAGAAGCACCGGCTAACTCCGTGCCAGCAGCCGCGGTAATAC
GGAGGGTGCAAGCGTTAATCGGAATTACTGGGCGTAAAGCGCACGCAGGCGGTTGGATAAGTTAGATGTGAAAGC
CCCGGGCTCAACCTGGGAATTGCATTTAAAACTGTCCAGCTAGAGTCTTGTAGAGGGGGGTAGAATTCCAGGTGT
AGCGGTGAAATGCGTAGAGATCTGGAGGAATACCGGTGGCGAAGGCGGCCCCC

***Aeromonas veronii* bv. sobria strain ER.1.24 16S ribosomal RNA**
gene, partial sequence, Length=1029 Score = 1195 bits (647), Expect = 0.0 Identities = 650/653 (99%), Gaps = 0/653 (0%), Strand=Plus/Plus

```
Query  1    TACTTTTGCCGGCGAGCGGCGGACGGGTGAGTAATGCCTGGGGATCTGCCCAGTCGAGGG  60
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  61   TACTTTTGCCGGCGAGCGGCGGACGGGTGAGTAATGCCTGGGGATCTGCCCAGTCGAGGG  120

Query  61   GGATAACTACTGGAAACGGTAGCTAATACCGCATACGCCCTACGGGGGAAAGCAGGGGAC  120
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  121  GGATAACTACTGGAAACGGTAGCTAATACCGCATACGCCCTACGGGGGAAAGCAGGGGAC  180

Query  121  CTTCGGGCCTTGCGCGATTGGATGAACCCAGGTGGGATTARCTAGTTGGTGAGGTAATGG  180
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  181  CTTCGGGCCTTGCGCGATTGGATGAACCCAGGTGGGATTAGCTAGTTGGTGAGGTAATGG  240

Query  181  CTCACCAAGGCGACGATCCCTARCTGGTCTGAGAGGATGATCAGCCACACTGGAACTGAG  240
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  241  CTCACCAAGGCGACGATCCCTAGCTGGTCTGAGAGGATGATCAGCCACACTGGAACTGAG  300

Query  241  ACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATATTGCACAATGGGGGAAACCC  300
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  301  ACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATATTGCACAATGGGGGAAACCC  360

Query  301  TGATGCMGCCATGCCGCGTGTGTGAAGAAGGCCTTCGGGTTGTAAAGCACTTTCAGCGAG  360
            |||||| ||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  361  TGATGCAGCCATGCCGCGTGTGTGAAGAAGGCCTTCGGGTTGTAAAGCACTTTCAGCGAG  420
```

```
Query  361  GAGGAAAGGTTGGTAGCTAATAACTGCCAGCTGTGACGTTACTCGCAGAAGAAGCACCGG  420
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  421  GAGGAAAGGTTGGTAGCTAATAACTGCCAGCTGTGACGTTACTCGCAGAAGAAGCACCGG  480

Query  421  CTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGTGCAAGCGTTAATCGGAATTACTG  480
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  481  CTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGTGCAAGCGTTAATCGGAATTACTG  540

Query  481  GGCGTAAAGCGCACGCAGGCGGTTGGATAAGTTAGATGTGAAAGCCCCGGGCTCAACCTG  540
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  541  GGCGTAAAGCGCACGCAGGCGGTTGGATAAGTTAGATGTGAAAGCCCCGGGCTCAACCTG  600

Query  541  GGAATTGCATTTAAAACTGTCCAGCTAGAGTCTTGTAGAGGGGGGTAGAATTCCAGGTGT  600
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  601  GGAATTGCATTTAAAACTGTCCAGCTAGAGTCTTGTAGAGGGGGGTAGAATTCCAGGTGT  660

Query  601  AGCGGTGAAATGCGTAGAGATCTGGAGGAATACCGGTGGCGAAGGCGGCCCCC  653
            |||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  661  AGCGGTGAAATGCGTAGAGATCTGGAGGAATACCGGTGGCGAAGGCGGCCCCC  713
```

Figure 1. Analyse bioinformatique des séquences d'ADNr16s sur GenBank, via le programme BlastN.

TP 1 : Lecture et Visualisation de Séquences d'ADN: Logiciel Sequence Scanner

L'intérêt de l'étude de la bioinformatique

- L'identification
- La taxonomie
- L'Evolution

Les étapes de l'identification moléculaire :

Extraction, PCR, Electrophorèse, Séquençage et Analyse Bioinformatique

Le choix du gène ARNr16S pour l'identification moléculaire

- La stabilité des extrémités du gène permet la synthèse d'amorces universelles.
- La grande base de données disponible sur internet utile pour la comparaison.
- C'est un gène universel présent chez tous les êtres vivants.
- Il contient des régions stables à vitesse d'évolution faible et des régions instables à vitesse d'évolution élevée.

Les alternatives du gène ARNr16S dans l'identification moléculaire ?

- Gène tuf
- Gènes codants pour: Enzyme, toxine, récepteur, hormone.

Le rôle du logiciel Sequence Scanner

Il permet de lire les fichiers AB machine du séquenceur

TP 2 : Recherche d'Alignement sur NCBI par le Programme BLAST

Les bases de données (ressources) les plus importantes en Bioinformatique :

NCBI JDDP EMBL PDB

Alignement

Homologie Similitude Correspondance Identité Blast

Le principe fondamental de l'alignement

Le maximum d'alignements pour le minimum de mutations

Définitions

-**Score** : Nombre de correspondance de nucléotides.

-**E value** : la probabilité que l'alignement a eu lieu par hasard.

-**GAP** : mutation (*indel)*

Identification et pourcentage d'homologie :

- L'espèce ≥ 99%
- Le genre ≥ 97%
- Le pourcentage qui ne permet l'identification < 97%

TP 3 : Introduction à l'Analyse Phylogénétique : Logiciel MEGA6

Les étapes et les programmes utilisés pour la construction d'un arbre phylogénétique par le logiciel MEGA6

- Alignement multiple ⟶ Alignement Explorer
- Matrice de distance ⟶ Matrix Distance Explorer
- Topologie de l'arbre ⟶ Tree Explorer

L'intérêt de la comparaison des séquences d'ADN dans l'Alignement Multiple

-Régions stables : comparaison des espèces éloignées.

-Régions instables : comparaison des espèces proches.

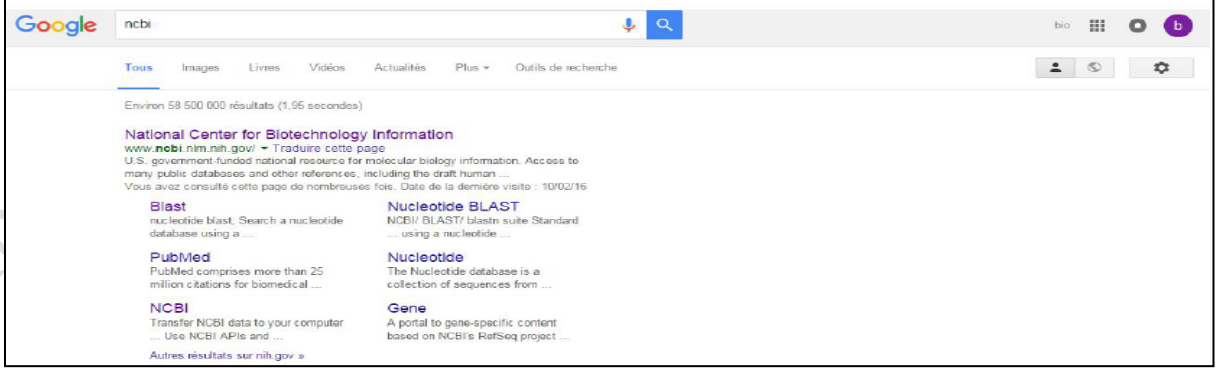
Recherche d'Alignement sur NCBI par le programme BLAST

Etapes du travail

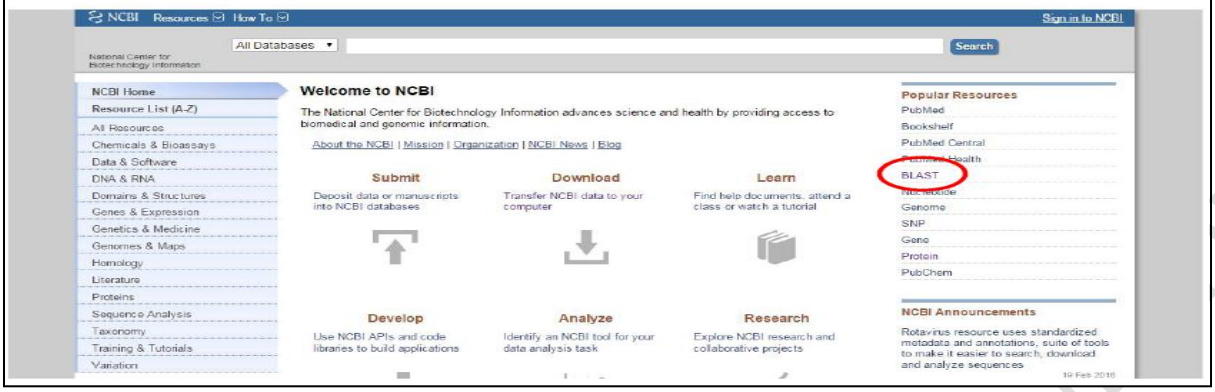
1. Ouverture du lien **NCBI** sur internet par l'utilisation du moteur de recherche google
2. Choix du programme **BLAST**
3. Choix de l'outil nucleotide **BLASTn**
4. Insertion de la **séquence** ADN ou le **Numéro d'Accès** sur Gene Bank et activation de l'outil BLAST
5. Lecture de la liste des résultats de l'Alignement
6. Lecture du détail des résultats de l'Alignement
7. Récolte des informations sur l'individu par le numéro d'accès sur Gene Bank

Auteur, affiliation, publication, séquence, etc.

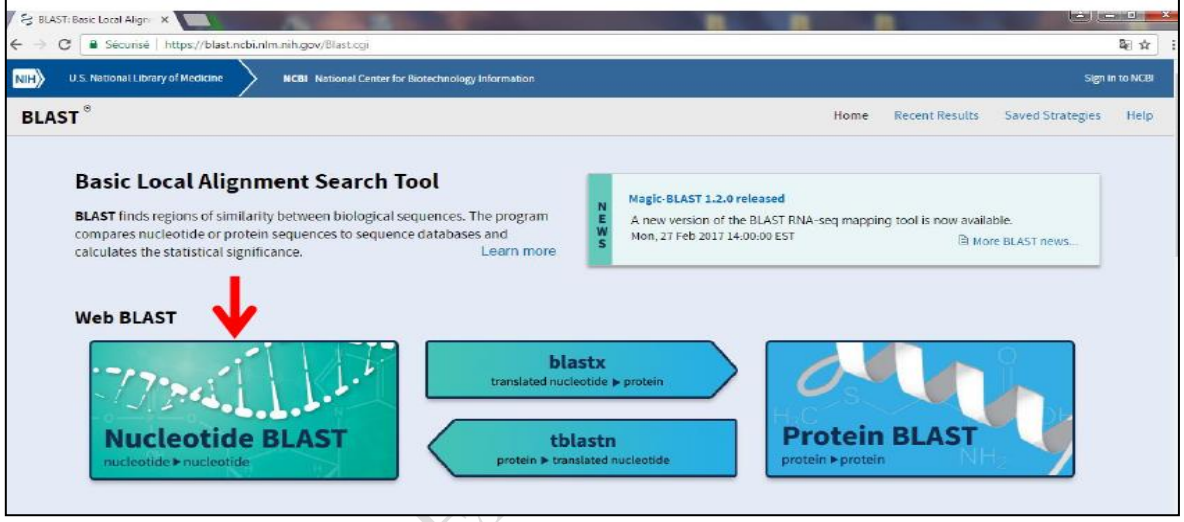
1. Ouverture du lien NCBI sur internet par l'utilisation du moteur recherche google



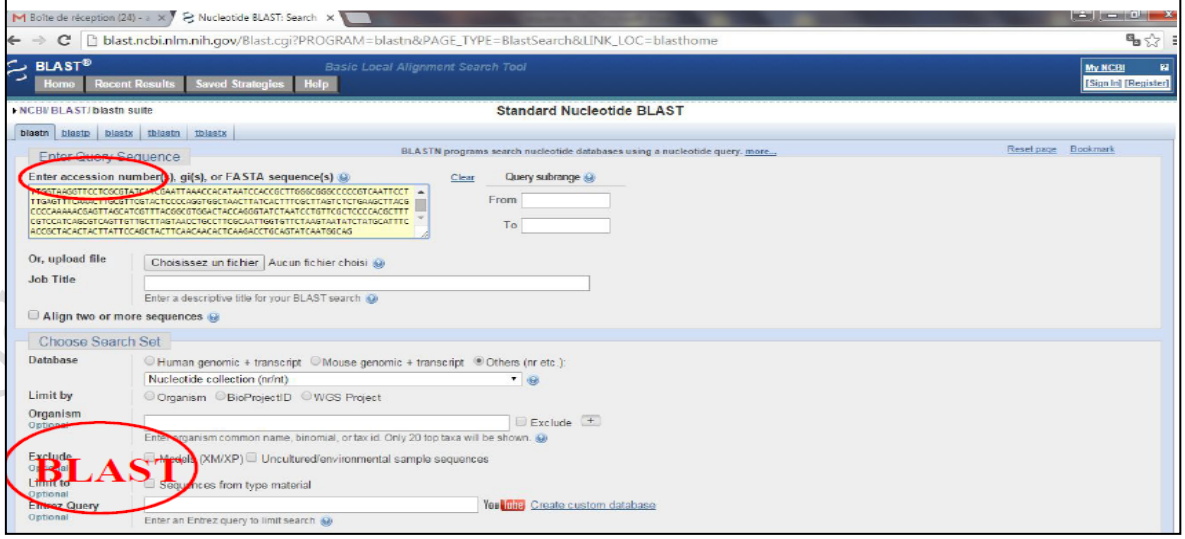
2. Choix du programme BLAST



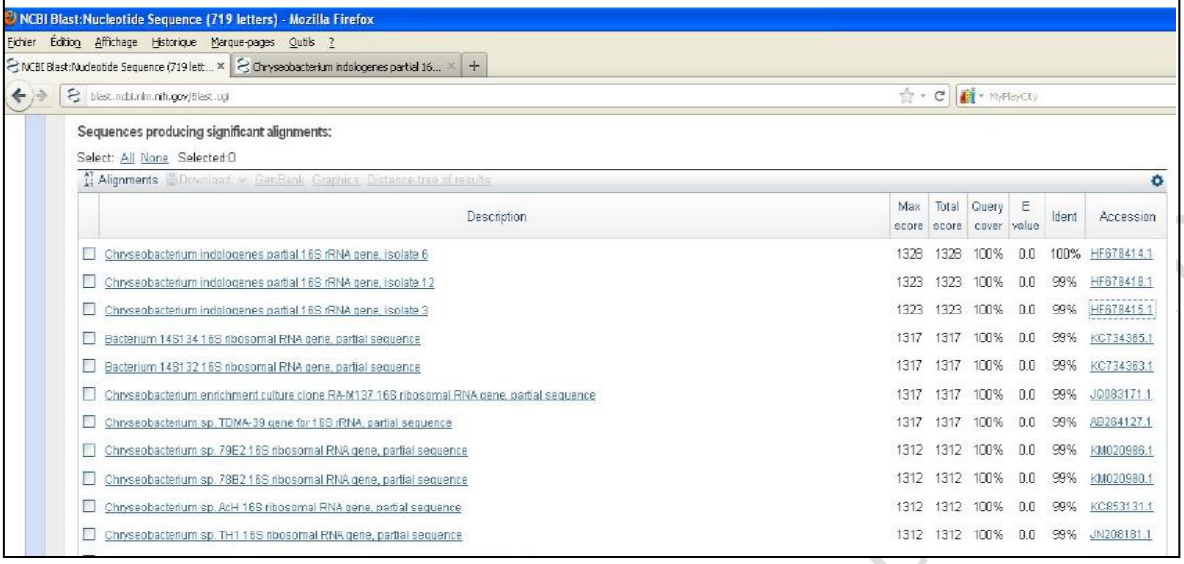
3. Choix de l'outil nucleotide BLASTn



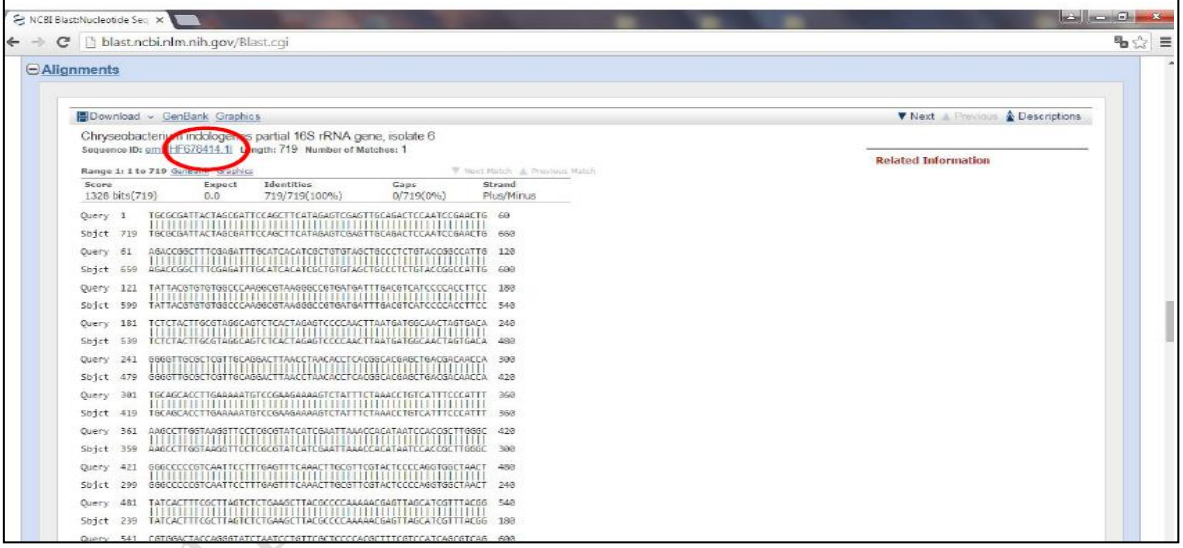
4. Insertion de la séquence ADN ou le numéro d'accès sur Gene Bank et activation de l'outil BLAST



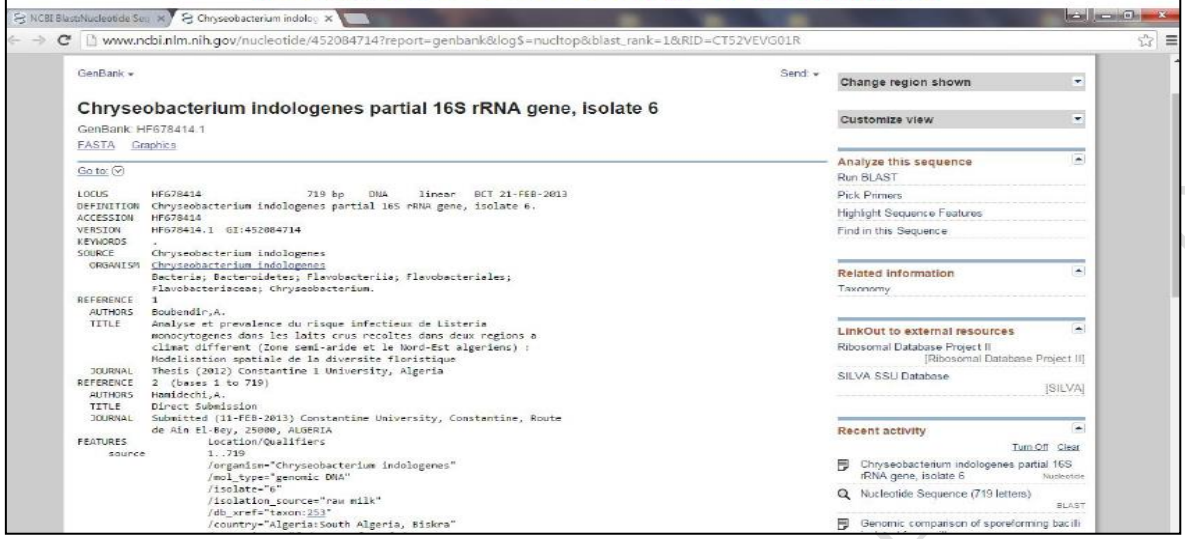
5. Lecture de la liste des résultats de l'Alignement



6. Lecture du détail des résultats de l'Alignement



7. Récolte des informations sur l'individu par le Numéro d'Accès sur Gene Bank



Suite Récolte des informations : Format FASTA