الجمهورية الجزائرية الديمقراطية الشعبية

République Algérienne Démocratique et Populaire

وزارة التعليم العالي والبحث العلمي

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

جامعة الإخوة منتوري قسنطينة 1
Frères Mentouri Constantine 1 University
Université Frères Mentouri Constantine 1

Faculté des Sciences de la Nature et de la Vie
Département de Biologie Appliquée

كلية علوم الطبيعة والحياة
قسم البيولوجيا التطبيقية

Mémoire présenté en vue de l'obtention du diplôme de Master

Domaine : Sciences de la Nature et de la Vie

Filière : Sciences biologiques

Spécialité : *Bio-informatique*

N° d'ordre :

N° de série :

Intitulé :

Une approche basée sur le traitement automatique du langage naturel (TALN) pour la classification taxonomique des séquences métagénomiques 16S rRNA.

Présenté par **:** GUECHTAL Loubna

OUELBANI Rania Nour

TALBI Yasmina

Le 18/06/2023

Jury d'évaluation :

Président : **Dr. KELLOU Kamel** (Université Frères Mentouri, Constantine 1).

Encadreur : **Dr. MATOUGUI Brahim** (Centre de recherches biotechnologie, Constantine).

Co-Encadreur : **Dr. GHERBOUDJ Amira** (Université Frères Mentouri, Constantine 1).

Examinateur 1 : **Dr. CHEHILI Hamza** (Université Frères Mentouri, Constantine 1).

Année universitaire
2022 – 2023

REMERCIMENT

Nous tenons tout d'abord a exprimer notre profonde gratitude envers notre professeur et encadrant, M. MATOUGUI BRAHIM. Sa guidance, son expertise et son devouement ont ete essentiels a la realisation de ce memoire. Sa disponibilite et sa volonte de partager ses connaissances nous ont grandement inspires tout au long de ce projet. Ses precieux conseils et ses encouragements constants nous ont pousses a donner le meilleur de nous-memes.

nous souhaitons egalement exprimer notre reconnaissance envers notre sous-encadrante, Mme. GHERBOUDJ AMIRA. Sa contribution significative et ses conseils eclaires ont grandement enrichi ce memoire. Sa passion pour le domaine et son soutien inebranlable ont ete des moteurs d'inspiration et d'excellence. A M. MATOUGUI BRAHIM et Mme GHERBOUDJ AMIRA, nous vous sommes profondement reconnaissants pour votre soutien indefectible et votre contribution precieuse a notre parcours academique. Votre confiance en nos capacites nous a encourages a repousser nos limites et a nous engager pleinement dans ce travail.

Nous souhaitons egalement remercier chaleureusement les membres du jury, M. KELLOU KAMEL ET M. CHEHILI HAMZA, pour avoir accepte d'evaluer notre memoire. Votre expertise et vos commentaires eclaires seront d'une valeur inestimable pour notre travail.

À tous, nous vous exprimons notre profonde gratitude et notre sincere appreciation pour votre contribution a notre reussite academique.

Dédicace

Avec l'expression de ma reconnaissance, je dédie ce modeste travail À ceux qui, quels que Soient les termes embrassés, je n'arriverais Jamais à leur exprimer mon amour sincère.

- *À l'homme, l'homme, mon précieux offre du dieu, qui doit ma vie, ma réussite Et tout Respect : mon cher père RACHID.*

- *À la femme qui a souffert sans me laisser souffrir, qui n'a jamais dit Non amés exigences et qui n'a épargné aucun effort pour me rendre Heureuse : mon adorable ma mère.*

- *À ma chère sœur FERIAL et mon frère ABDARAHMAN que n'ont pas Cessée de me conseiller, encourager et soutenir tout au long de mes Etudes. Que Dieu les protège et leurs offre la chance et le bonheur.*

- *À mon adorable petite sœur Racha qui sait toujours comment me réconfortée Procurer la joie et le bonheur pour tout la famille.*

- *. À camarade des files d'attente universitaires, à une sœur de l'école de la vie, à mon Amie Guechtal Loubna, merci d'être le meilleur soutien et l'aide dans les jours D'adversité.*

- *À mes amis Wissal et Meryem qui m'ont toujours encouragé, Merci pour leurs amours et leurs encouragements.*

- *Sans oublier mon binôme Loubna et Nour pour son soutien moral, sa patience Et sa compréhension tout au long de ce projet*

Yasmina

Dédicace

À ma chère mère Sabrina et à mon cher père Reda, mes premiers piliers et sources infinies d'amour et de soutien tout au long de mon parcours, je vous remercie du fond du cœur pour tout ce que vous avez fait pour moi. Votre présence inébranlable et vos conseils précieux ont été les clés de ma réussite.

À mon frère et ma sœur bien-aimés, Sofiane et Serine, mes compagnons de vie et mes alliés indéfectibles, je suis fière de vous avoir dans ma vie. Notre lien familial est une source de bonheur et de force.

À ma tendre mamie, un refuge de tendresse et de réconfort, je chéris chaque instant passé à tes côtés. Tes paroles douces, tes conseils avisés et ta bienveillance infinie ont été des bouées dans les moments difficiles. Je te porte dans mon cœur avec une affection éternelle.

À mon cher grand-père, qui repose en paix, mais dont l'héritage continue d'illuminer ma vie, je suis reconnaissante d'avoir eu la chance de te connaître et de t'avoir comme modèle. Ta sagesse, ton amour et ton exemple restent gravés en moi.

À mes tantes adorées, Dalila et Lynda, et à mes oncles Djamil, Salim et Nasser, je suis profondément reconnaissante pour votre soutien inconditionnel, votre amour sincère, votre générosité et vos conseils éclairés. Votre présence bienveillante a été une source d'inspiration et de réconfort.

Au Docteur Salah Eddine ALIOUANE, je voudrais exprimer ma profonde gratitude pour votre aide précieuse tout au long de notre mémoire. Votre expertise, votre patience et votre dévouement ont été essentiels à notre succès. Vos conseils éclairés et vos commentaires perspicaces ont été inestimables. Merci d'avoir été un mentor exceptionnel.

À mes chères amies Sérine, Malek, Rahma et Yasmine, je vous suis infiniment reconnaissante pour votre amitié sincère et votre soutien indéfectible. Votre présence dans ma vie apporte une joie et une complicité précieuses. Merci d'être là pour moi.

Merci du fond du cœur à tous ceux qui ont contribué à mon chemin, de près ou de loin, avec leur amour, leur soutien et leur présence. Votre impact dans ma vie est inestimable, et je vous serai éternellement reconnaissante pour tout ce que vous avez apporté à mon parcours.

RANIA NOUR

Dédicace

Mes chers parents, ma famille, mes amis et mes binômes, Je voudrais prendre un moment pour vous exprimer ma gratitude et dédier ces mots à vous tous, qui ont été mes piliers de soutien tout au long de mon parcours de rédaction de mémoire.

- *À mon père Bekouche et ma mère Wahiba, qui m'ont toujours encouragée à poursuivre mes rêves et à me dépasser. Votre amour inconditionnel, vos encouragements constants et votre présence réconfortante m'ont donné la force de persévérer.*
- *À ma famille, ma grand-mère Abida, mes tentes qui ont été ma source d'inspiration et ma motivation. Votre soutien indéfectible, vos encouragements et votre compréhension ont été essentiels pour moi. Vous avez été mes rochers, m'aidant à surmonter les hauts et les bas de ce projet exigeant.*
- *A ma sœur et amie Yasmina, qui m'a toujours accompagné dans mes moments de tristesse et de joie, à ma compagne de vie, merci pour ton amitié sincère et tes encouragements constants, ils ont été comme un vrai baume pour mon âme.*
- *A mes sœurs Ines, chaima, Chourouk et Darine et mon petit frère Iyed, merci d'être à mes cotes de toujours me soutenir, de me comprendre, de m'aimer comme je suis.*
- *À mes binômes Yasmina et Nour, avec qui j'ai partagé des heures de travail acharné, de brainstorming et de collaboration. Votre soutien, votre expertise et votre engagement envers notre projet commun ont été inestimables. Ensemble, nous avons relevé les défis et nous avons atteint des sommets que je n'aurais pu atteindre seule.*
- *A mes amies Ouissal et meryem, mes partenaires, mes confidentes, avec qui nous avons passé des moments inoubliables, merci à vous.*
- *À vous tous, je dédie cette réussite. Vos encouragements, vos conseils et votre présence m'ont donné la force de continuer lorsque les obstacles semblaient insurmontables. Votre confiance en moi m'a inspiré à donner le meilleur de moi-même et à ne jamais abandonner.*

Avec tout mon amour et ma reconnaissance.

GUECHTAL LOUBNA

RESUME

Cette étude vise à développer une approche basée sur les réseaux LSTM (Long Short-Term Memory) pour simuler les différentes étapes du processus de la métagénomique, en se concentrant spécifiquement sur l'analyse des données générées par les technologies de séquençage de nouvelle génération (NGS). Cette approche repose sur deux axes clés de l'intelligence artificielle, à savoir le traitement automatique du langage naturel (NLP) et l'apprentissage profond (DL). En utilisant un ensemble de données d'apprentissage composé de neuf bactéries, des tests ont été effectués et ont abouti à un taux de précision de 98%. Ces résultats démontrent l'efficacité de l'approche, notamment en ce qui concerne la phase de prédiction basée sur le TALAN et le DL. La combinaison de ces deux outils a permis de développer un modèle possédant une grande capacité d'extraction de connaissances à partir des données génomiques, permettant ainsi la prédiction et la classification taxonomique des génomes. Ce modèle a été entraîné de manière approfondie en exploitant les séquences génomiques. Les résultats de cette recherche mettent en évidence l'apport significatif de cette approche pour améliorer la précision de la classification des génomes.

Mot clés : Apprentissage ; prédiction ; Intelligence Artificielle ; TALAN ; NGS ; LSTM.

ABSTRACT

This study aims to develop an approach based on LSTM (Long Short-Term Memory) networks to simulate the different stages of the metagenomics process, with a specific focus on analyzing data generated by Next-Generation Sequencing (NGS) technologies. This approach relies on two key aspects of artificial intelligence, namely Natural Language Processing (NLP) and Deep Learning (DL). Using a training dataset consisting of nine bacteria, tests were conducted, resulting in a precision rate of 98%. These results demonstrate the effectiveness of the approach, particularly in the prediction phase based on NLP and DL. The combination of these two tools has enabled the development of a model with a high capacity to extract knowledge from genomic data, thus enabling genome prediction and taxonomic classification. The model was trained extensively by exploiting genomic sequences. The findings of this research highlight the significant contribution of this approach in improving the precision of genome classification.

Keywords: Learning; Prediction; Artificial Intelligence; NLP; NGS؛LSTM.

الملخص

تهدف هذه الدراسة إلى تطوير نهج قائم على شبكات LSTM (الذاكرة طويلة المدى) لمحاكاة المراحل المختلفة لعملية الميتاجينوميك، مع التركيز بشكل خاص على تحليل البيانات الناتجة عن تقنيات تسلسل الجيل التالي (NGS). يعتمد هذا النهج على جانبين رئيسيين للذكاء الاصطناعي، وهما معالجة اللغة الطبيعية (NLP) والتعلم العميق (DL). باستخدام مجموعة بيانات تدريبية تتكون من تسع بكتيريا، تم إجراء الاختبارات، مما أدى إلى معدل دقة يبلغ 98٪. توضح هذه النتائج فعالية النهج، لا سيما في مرحلة التنبؤ بناءً على NLPوDL. أتاح الجمع بين هاتين الأداتين تطوير نموذج ذي قدرة عالية على استخراج المعرفة من البيانات الجينومية، وبالتالي تمكين التنبؤ الجينومي والتصنيف التصنيفي. تم تدريب النموذج على نطاق واسع من خلال استغلال التسلسلات الجينية. تسلط نتائج هذا البحث الضوء على المساهمة الكبيرة لهذا النهج في تحسين دقة تصنيف الجينوم.

الكلمات الرئيسية: التعلمالعميق؛ التنبؤ؛ الذكاء الاصطناعي؛ البرمجة اللغوية العصبية؛ تسلسل الجيل التالي؛ الذاكرة طويلة المدى

ACRONYMES

- 16S: Ribosomal 16S (16S ribosomal)
- ACP : Analyse en composantes principales (Principal Component Analysis)
- ADN : Acide désoxyribonucléique (DNA – Deoxy ribo nucleic Acid)
- ARN: Acide ribo nucléique (RNA - Ribonucleic Acid)
- ARNR 16S: ARN ribosomique 16S (16S Ribosomal RNA)
- BLAST: Basic local alignment search tool (Outil de recherche d'alignement local de base)
- CBOW: Continuous bag of words (Sac continu de mots)
- CNN : Convolutional Neural Network (Réseau neuronal convolutif)
- CSV : Comma-Separated Values (Valeurs séparées par des virgules)
- DL : Deep Learning (Apprentissage profond)
- EMG : European nucleotide archive (ENA) métagénomiques (Archives européennes de nucléotides - métagénomique)
- FASTA : Format for alignment and sequenceanalysis (Format pour l'alignement et l'analyse de séquences)
- FASTDNA : Fast DNA embedding (Incorporation rapide de l'ADN)
- GLVDNA : Global vectors for DNA representation (Vecteurs globaux pour la représentation de l'ADN)
- GLVPROTÉINE : Global vectors for protein representation (Vecteurs globaux pour la représentation des protéines)
- HMP : Human microbiome project (Projet du microbiome humain)
- IMG/MER : Integratedmicrobialgenomes and microbiome expert review (Examen d'experts des génomes microbiens intégrés et du microbiome)
- ITS : Internal transcribed spacer (Espace transcrit interne)
- LCA : Lowest common ancestor (Ancêtre commun le plus récent)
- LSHVEC : Locality-sensitive hashingvector (Vecteur de hachage sensible à la localité)
- LSTM : Long short-term memory (Mémoire à court terme longue)
- ML : Apprentissage automatique (Machine Learning)
- NGS : Next generation sequencing (Séquençage de nouvelle génération)
- NLP : Natural language processing (Traitement automatique du langage naturel)

- NLTK : Natural language toolkit (Boîte à outils pour le traitement du langage naturel)
- PACBIO : Pacific Biosciences (Pacifique Biosciences)
- PCA : Principal component analysis (Analyse en composantes principales)
- PCR : Polymerase chain reaction (Réaction en chaîne par polymérase)
- SEQ2VEC : Sequence to vector (Séquence vers vecteur)
- SILVA : Système d'information sur les séquences du domaine de la vie (Sequence information database for the domain of Life)
- TALN : Traitement automatique du langage naturel (Natural Language Processing)
- WORD2VEC: Word to vector (Mot versvecteur)
-

LISTE DES TABLEAUX

LISTE DES FIGURES

Table des matières

INTRODUCTION GÉNIRALE

Avec l'émergence des techniques de séquençage à haut débit, la métagénomique est devenue une discipline cruciale pour explorer la diversité et les fonctions des communautés microbiennes présentes dans divers écosystèmes. Cependant, l'analyse des données métagénomiques générées par ces techniques nécessite l'utilisation d'outils et de méthodes avancées afin de traiter et d'interpréter les informations complexes contenues dans les séquences d'acides nucléiques.

La bio-informatique joue un rôle crucial dans cette analyse en fournissant des solutions informatiques pour gérer, analyser et interpréter les données métagénomiques. L'une des tâches clés en métagénomique est la classification taxonomique des séquences génomiques, c'est-à-dire l'assignation des séquences à des groupes taxonomiques tels que les espèces, les genres ou les familles microbiennes. Cette classification permet de comprendre la composition et la diversité des communautés microbiennes dans un échantillon métagénomique donné. Traditionnellement, les méthodes de classification taxonomique se basent sur des techniques de comparaison de séquences ou d'alignement avec des bases de données de référence. Cependant, avec l'énorme quantité de données métagénomiques produites, ces méthodes traditionnelles peuvent devenir inefficaces en termes de rapidité et de précision.

C'est là que l'intelligence artificielle et plus spécifiquement le traitement automatique du langage naturel (NLP) avec l'utilisation de l'outil GloVe (Global Vectors for Word Representation) entre en jeu. L'intelligence artificielle, en particulier l'apprentissage profond, offre de nouvelles opportunités pour développer des approches de classification taxonomique plus rapides et plus précises.

Notre contribution consiste à proposer une nouvelle approche qui utilise l'apprentissage profond et le traitement automatique du langage naturel pour proposer un nouveau classifieur taxonomique. Cette approche vise à améliorer la classification taxonomique des séquences métagénomiques en exploitant les informations contenues dans les séquences sous forme de chaînes de

caractères En utilisant GloVe pour représenter les séquences génomiques sous forme vectorielle. Cette représentation permis de capturer les similitudes sémantiques et structurelles entre les séquences génomiques et les utiliser pour la classification taxonomique.

En utilisant cette approche novatrice, nous espérons obtenir des résultats prometteurs dans la classification taxonomique des échantillons métagénomiques, en offrant une solution plus rapide et plus précise par rapport aux méthodes traditionnelles. Cela pourrait contribuer à une meilleure compréhension de la diversité microbienne et de son rôle dans les écosystèmes, ainsi qu'à de nombreuses applications dans des domaines tels que la santé, l'agriculture et l'environnement [1].

CHAPITRE 1 : ETAT DE L'ART

CHAPITRE 1

1. Du Génomique à la Métagénomique

L'étude des génomes bactériens, y compris leur structure, leur évolution, la fonction de leurs gènes codifiés et de leur régulation, est devenue possible grâce à la génétique bactérienne, qui repose principalement sur l'isolement et la culture d'une bactérie particulière. Pour relier le phénotype unique d'une souche bactérienne à sa composition génétique, par exemple pour examiner en détail sa virulence, sa pathogenèse ou sa résistance, l'obtention d'une culture pure est une étape cruciale. Or, une très grande majorité de bactéries - plus de 99 % - ne peuvent être cultivées en laboratoire. Afin d'étudier une communauté bactérienne dans son ensemble, il est désormais possible de séquencer l'ADN de toutes les bactéries présentes dans un environnement donné (sol, eau, tubes digestifs humains et animaux, échantillons cliniques, etc.) Cette méthode, dite "Métagénomique". Les premières études métagénomiques remontent à 1985 et 1986 mais le terme “métagénomique” a été utilisé pour la première fois en 1998. [2].

2. Définition de la Métagénomique

La métagénomique est l'application de techniques génomiques contemporaines pour l'étude des communautés microbiennes directement dans leur environnement naturel, en évitant la nécessité d'isoler et de cultiver en laboratoire des espèces spécifiques. Grâce à ses méthodes, la recherche métagénomique va au-delà du génome individuel pour examiner le génome d'une communauté dans son ensemble et quantifier sa diversité en termes d'abondance d'espèces. L'étude de la métagénomique est basée sur deux approches : l'approche descriptive qui permet d'estimer les abondances relatives du microbiome en fonction de diverses conditions physiologiques et environnementales afin de révéler la structure communautaire et la variabilité du microbiome , et l'approche fonctionnelle, qui étudie les interactions hôte-microbe et microbe-microbe et construire des modèles d'écosystèmes dynamiques prédictifs pour refléter les liens entre les microbes ou les identités communautaires. Le développement conjoint de ces d’approches métaboliques, qui sont basées sur l'extraction directe de l'acide nucléique adénosine (ADN) bactérien, a le potentiel de faire progresser les connaissances dans un large éventail de domaines, y compris l’agriculture, les biocarburants, la biotechnologie, les sciences de l'environnement et la médecine. Par exemple dans la médecine où la métagénomique a permis de comprendre le rôle de l’environnement naturel et de découvrir ainsi de nouveaux médicaments pouvant être potentiellement utiles au développement des molécules à intérêt thérapeutique comme des an-

tibiotiques, un autre exemple dans le domaine de la marine, où les biomatériaux dérivés d'organismes marins sont utilisés dans un large éventail d'applications industrielles. Parmi les exemples courants les biopolymères, les acides aminés et les pigments naturels [3].

2.1. Type des études Métagénomique

Presque tous les micro-organismes présents sur Terre aujourd'hui sont étudiés par la métagénomiques dans divers environnements notamment le sol, les nuages, l'air, l'eau des lacs et des océans, ainsi que les communautés de microbes liées aux règnes végétal et animal (telles que les communautés rhizosphériques et photos sphériques). Elle étudie également les micro-organismes associés à l'homme, tels que les intestins, qui font l'objet d'études de grands projets métagénomiques [4].

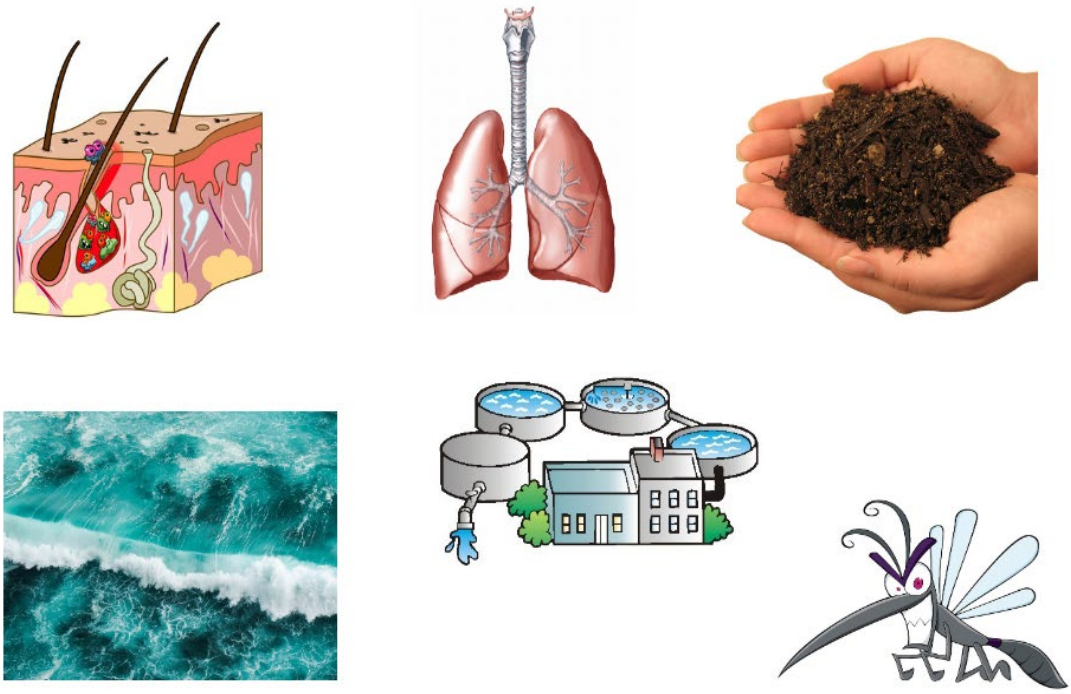

Figure 1: Les types des études Métagénomiques [5].

Il y a eu plusieurs types des études Métagénomique En ce qui concerne la santé, la métagénomique est de plus en plus utilisée pour étudier la diversité microbienne présente dans le microbiote intestinal et son rôle dans de nombreuses maladies, notamment les maladies inflammatoires de l'intestin, le syndrome du côlon irritable, l'obésité, le diabète et même certaines maladies neurodégénératives [4], un exemple de projet dans ce domaine d'étude :

2.1.1. La Métagénomique santé

En ce qui concerne la santé, la métagénomique est de plus en plus utilisée pour étudier la diversité microbienne présente dans le microbiote intestinal et son rôle dans de nombreuses maladies, notamment les maladies inflammatoires de l'intestin, le syndrome du côlon irritable, l'obésité, le diabète et même certaines maladies neurodégénératives [4], un exemple de projet dans ce domaine d'étude :

Human Microbiome Project (HMP) : c'est un projet lancé en 2008 et qui s'est terminé en 2013. Le but du projet était de cartographier le microbiote humain en identifiant les microbes présents dans le corps humain, en étudiant leur fonction et leur interaction avec l'hôte [6].

2.1.2. La Métagénomique marine

La métagénomique marine implique la collecte d'échantillons d'eau de mer, de sédiments ou d'autres matrices environnementales pour extraire l'ADN des communautés microbiennes. Elle a des applications importantes dans la surveillance de la qualité de l'eau, la surveillance des maladies et la découverte de nouveaux produits naturels d'origine marine [4], voici un exemple de projet de grande envergure dans ce domaine :

Tara Océans Expédition : c'est un projet de recherche lancé en 2009 pour étudier la biodiversité marine à l'échelle mondiale. Il a utilisé des approches métagénomiques pour séquencer l'ADN de la communauté microbienne présente dans les échantillons d'eau de mer prélevés tout au long de l'expédition. Ce projet a permis de découvrir de nouvelles espèces de micro-organismes et d'identifier des processus métaboliques clés dans l'océan [7].

2.1.3. La Métagénomique environnemental

Permet d'obtenir des informations sur la diversité des micro-organismes, leur distribution spatiale, leur dynamique temporelle et leur fonctionnement écologique dans les écosystèmes naturels [4]. Un exemple de projet :

Earth Microbiome Project (EMP) : c'est un projet de recherche lancé en 2010 pour cartographier la diversité microbienne de la planète et pour étudier les interactions entre les microbes et leur environnement. Ce projet a utilisé des approches métagénomiques pour séquencer l'ADN des communautés microbiennes dans différents écosystèmes terrestres et aquatiques. [8]

3. Le Séquençage

L'étude des communautés microbiennes existe depuis longtemps, mais elle a longtemps été limitée à l'utilisation de techniques d'imagerie qui ne permettaient qu'une observation des caractéristiques morphologiques. Seuls les organismes pouvant être cultivés pouvaient être étudiés dans ce contexte. Ainsi, avant le développent des technologies de la biologie moléculaire, seule une étude à faible résolution d'un petit sous-ensemble de bactéries connues était possible. Le développement des outils de biologie moléculaire a permis de contourner ces obstacles et a révolutionné la microbiologie. Cette évolution a été rendue possible par le séquençage Sanger, créée en 1977, qui a permis d'accéder à la structure et à la fonction des génomes bactériens, en donnant leur séquence sous forme de fragments appelés lectures et mesurant quelques centaines de bases [9].

Tableau 1: comparaison des différentes technologies de séquençage [9].

	Technologie	Longueur de la lecture (Bases)	Taux d'erreur	Cout moyen par giga base
Premières Génération	Sanger	400-900	<0,1%	NA
Second		150	<0,1%	$7
	Illumina	300 (lectures paires)		
Génération	Roche 454	400	1%	$9,500
	ABI SOLID	75	<0,1%	$70
Troisiéme	pacific bioscience	10K en moyenne	~5%	$1,000
Génération	Oxford nanopore	10K en moyenne	~5%	$750

3.1. Le Séquençage de la deuxième génération

La technologie Sanger est toutefois limitée par le faible nombre de lectures qu'elle produit et, par conséquent, par la nature coûteuse d'un grand projet de séquençage. À la fin des années

1990, de nouvelles technologies connues sous le nom de séquençage à haut débit de nouvelle génération (ou NGS pour Next Generation Sequencing) ont commencé à voir le jour. Ces technologies automatisées et hautement parallèles permettent la production à faible coût de millions de lectures à chaque run de séquençage (Figure 2). Une expérience de séquençage haut-débit commence par l'extraction du matériel génétique d'un échantillon biologique [9]. Un exemple populaire de séquenceur NGS est le système Illumina. Il utilise la technologie de séquençage par synthèse dite "sequencing by synthesis". Dans ce processus, de courtes séquences d'ADN sont amplifiées et fixées sur une surface. Ensuite, des amorces spécifiques sont ajoutées et des nucléotides marqués fluorescent sont incorporés un par un. Chaque nucléotide incorporé est détecté par un scanner laser, et cette information est utilisée pour reconstruire la séquence d'ADN originale. Grâce à sa capacité à produire des lectures massivement parallèles, les séquenceurs Illumina sont largement utilisés dans la recherche génomique, la médecine personnalisée et d'autres domaines liés à la génomique [10].

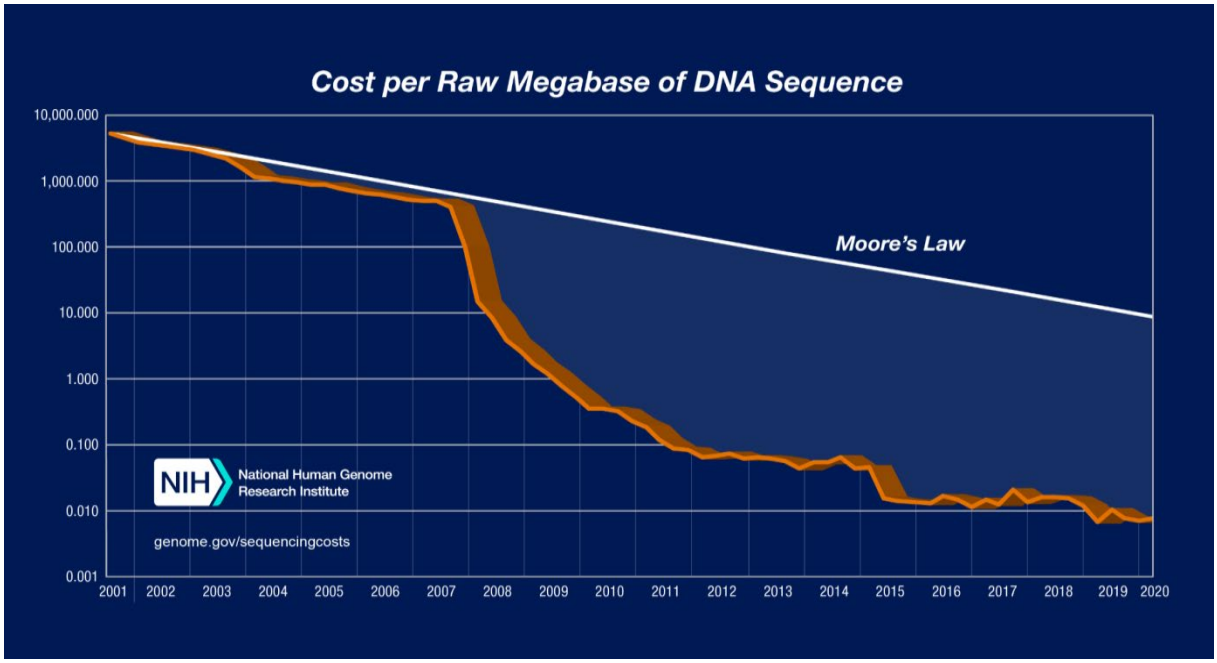

Figure 2: Évolution du coût de séquençage (en dollar) d'une méga base d'ADN, en échelle logarithmique [11].

3.2. Le séquençage de la troisième génération

La principale limite du séquençage de la deuxième génération est la longueur limitée des lectures, ce qui rend difficile, voire impossible, la résolution de certains problèmes liés à l'assemblage. Les technologies de séquençage les plus récentes peuvent être classées dans la catégorie "longue portée". Elles permettent, par exemple, le séquençage des lectures plus longues qui peuvent couvrir jusqu'à un million de bases ou même l'enchaînement de de courtes lectures provenant d'une même région génomique. Malgré les points positifs que la troisième génération nous offre qui sont la longueur des lectures qui est plus élevée que celui de la 2eme génération mais elle contient des points négatifs et là on parle du taux d'erreur qui est presque égale 5% Comparé à celui de la 2eme génération qui est pratiquement inexistante qui varie entre 0.1% et1% (Tableau 1) [9].

Figure 3: séquenceur « MinION » [12].

4. Types de la Métagénomique

La métagénomique essaye de répondre principalement sur ces trois questions (Figure 4) Qui est là ? Pour découvrir la composition taxonomique de l'échantillon. Tandis que les deux autres questions (quoi ? et comment ?) Essayent de révéler le potentiel fonctionnel de l'échantillon.

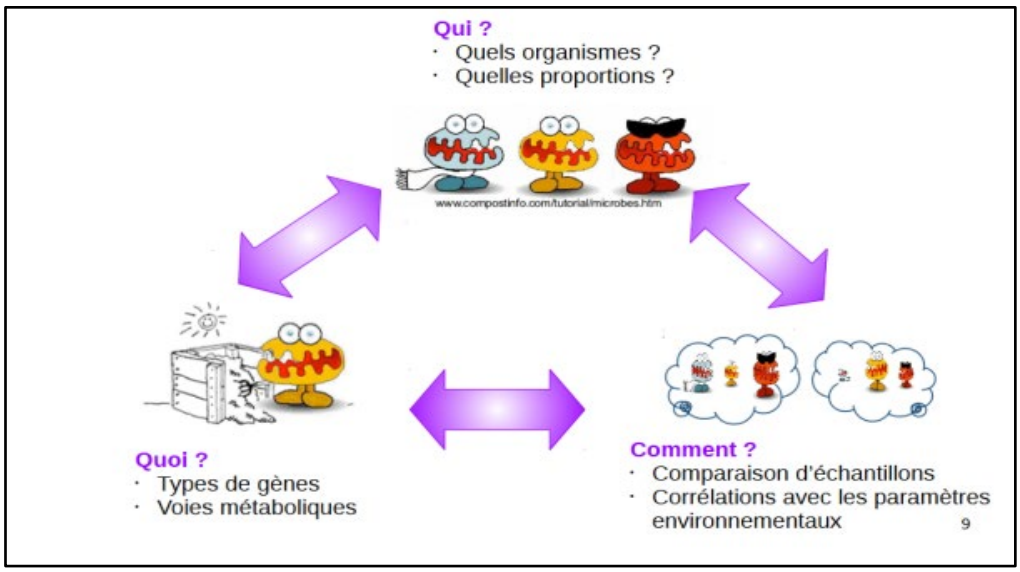

Figure 4: les questions posées par la métagénomique [5].

Suivant les questions, il existe deux types d'approches de la métagénomique :

4.1. La métagénomique ciblée (Amplicon)

La métagénomique ciblée, également connue sous le nom de metabarcoding, est une approche qui implique l'amplification et le séquençage d'une région spécifique du génome, en particulier la région de l'ADN ribosomique 16S des bactéries [13]. Cette méthode est couramment utilisée pour caractériser taxonomiquement un échantillon, en répondant à la question "Qui est là ?" Elle permet d'identifier les organismes présents dans une communauté microbienne. La méta-génomique ciblée est moins coûteuse que la métagénomique globale et permet d'identifier des organismes plus rares avec un effort de séquençage comparable. Elle est souvent utilisée dans les phases initiales de l'analyse métagénomique et dans le cadre de projets de catalogage de la diversité bactérienne [9].

Cette approche consiste à localiser les bactéries d'une communauté complexe en utilisant des cibles universelles et bien établies. Plusieurs loci se sont imposés comme marqueurs de référence pour divers règnes, fréquemment trouvés dans l'opéron ribosomique (ADNr 16S pour les bactéries, ITS (Internal Transcribed Spacers) pour les champignons, ADNr 18S pour les eucaryotes). Par exemple, l'ARNr 16S est un marqueur génétique sur lequel les biologistes se sont appuyés pour identifier et classifier les différentes espèces bactériennes. Il s'agit d'un ARN non codant composé d'environ 1500 nucléotides, présent dans la petite sous-unité ribosomique des bactéries. Cette région contient à la fois des régions conservées, qui permettent la conception de sondes spécifiques, et des régions hypervariables, qui fournissent des informations phylogénétiques pour la classification des organismes [5].

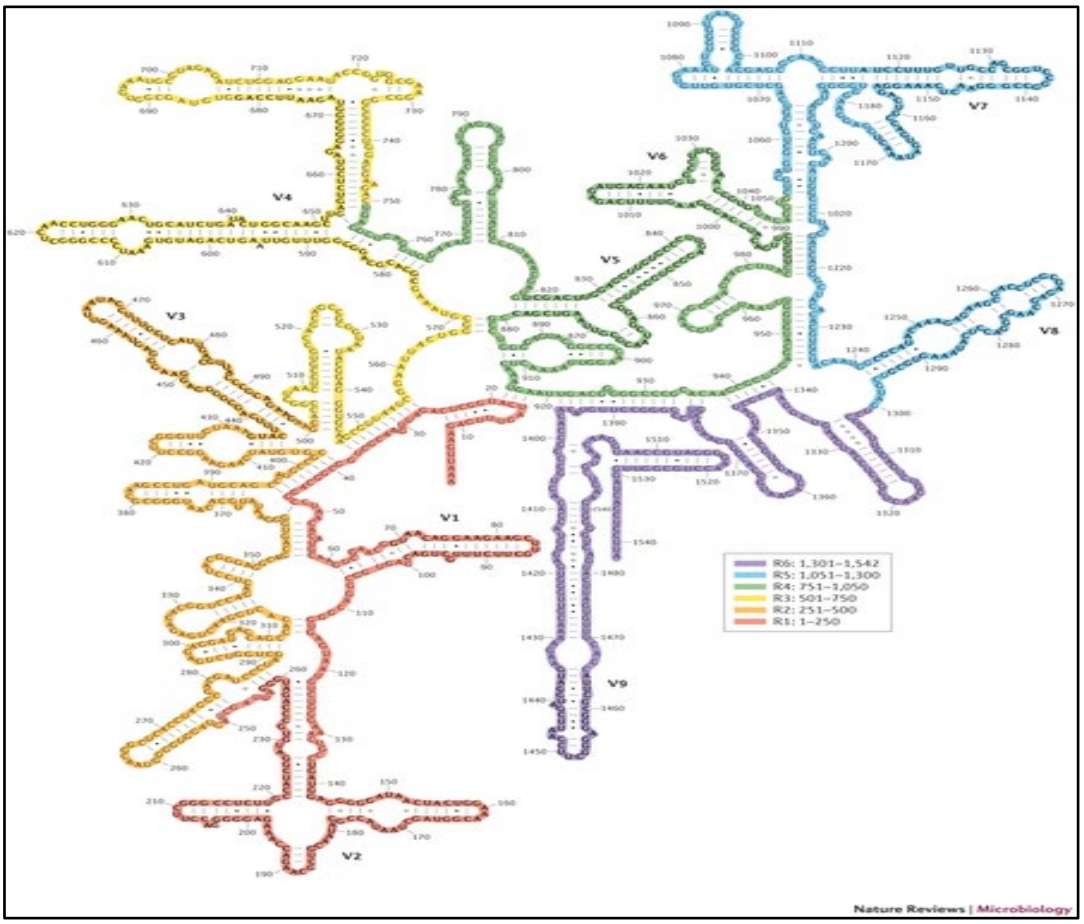

Figure 5: Structure secondaire de l'ARNr 16S [5].

Ce gène est en effet le plus conservé des trois gènes de l'opéron ribosomique au sein d'une même espèce, et contient des régions hypervariables qui permettent de ségréguer les espèces bactériennes en se basant sur sa séquence (Figure 8). Il a ainsi été proposé comme marqueur évolutif de référence pour le règne bactérien. Il peut être amplifié dans de nombreuses bactéries différentes d'un même échantillon en une seule réaction, grâce à ses régions hautement conservées entre taxon [14].

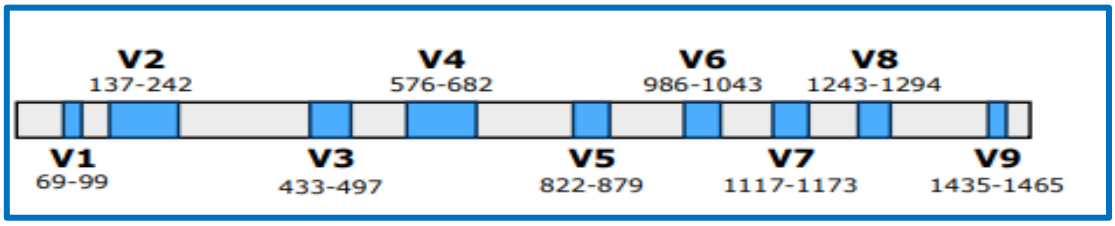

Figure 6: Région conservées et Hypervariable de l'ARNr [14].

Les régions conservées : il est possible de construire des amorces PCR pour choisir la région désirée et capturer tout l'ARNr 16S. Les régions variables n'ont pas de rôle fonctionnel significatif et peuvent diverger dans le temps à la suite de mutations neutres. Celles-ci permettront de distinguer les taxons bactériens. [5]

4.2. La métagénomique globale (Shotgun)

La métagénomique Shotgun, également connue sous le nom de métagénomique globale, est une approche qui consiste à séquencer tout l'ADN présent dans un échantillon, sans amplification préalable d'une région spécifique du génome. Cette méthode permet de répondre à des questions sur la composition taxonomique des communautés microbiennes ("Qui est là ?"), ainsi que sur leur potentiel fonctionnel à travers l'analyse du répertoire de gènes des membres de la communauté ("Quoi ?" et "Comment ?») [5].

Le séquençage de l'ADNr 16S microbien est considéré comme l'étalon-or pour la caractérisation des communautés microbiennes, mais il peut présenter certaines limitations. Par exemple, le séquençage basé sur le gène de l'ARNr peut détecter les membres prédominants de la communauté, mais il peut ne pas détecter les membres rares avec des séquences cibles divergentes. De

plus, le biais des amorces et la faible profondeur d'échantillonnage peuvent limiter sa sensibilité. Ces limitations peuvent être surmontées par le séquençage de génomes microbiens entiers. Les approches basées sur le génome entier offrent la promesse d'une couverture plus complète grâce aux plateformes de séquençage de l'ADN à haut débit. Elles ne sont pas limitées par la conservation de la séquence ou la variation du site de liaison de l'amorce au sein d'une cible spécifique. Cela permet d'obtenir des informations plus riches et détaillées sur la diversité et les fonctions microbiennes.

Cependant, le séquençage d'un échantillon par métagénomique Shotgun est plus coûteux que la métagénomique ciblée, car il nécessite une grande quantité de lectures pour obtenir une couverture adéquate des génomes collectifs. Malgré ces coûts plus élevés, la métagénomique Shotgun est une approche puissante pour explorer la diversité et le fonctionnement des communautés microbiennes à un niveau plus profond [9].

5. Défis bio-informatiques pour la métagénomique

Les données métagénomiques nous permettent principalement de jeter un regard neuf sur des communautés précédemment sous-estimées. Ces données présentent également des caractéristiques uniques qui nécessitent le développement de méthodes informatiques spécialisées [9].

5.1. Volume de données

Les séquences de données métagénomiques ont le potentiel de fournir d'énormes quantités de données. Pour caractériser des organismes rares dans des écosystèmes complexes tels que le sol ou l'eau de mer, un important travail de séquençage est nécessaire. De plus, de nombreuses études nécessitent le séquençage et l'analyse collaborative de plusieurs dizaines ou centaines d'échantillons afin de comparer des écosystèmes différents [9].

5.2. Diversité génomique

Les communautés bactériennes présentent généralement un continuum de diversité, décomposé en différents niveaux taxonomiques. En règle générale, ces communautés contiennent une variété d'espèces de microbes, dont certaines peuvent avoir des régions homologues dans leur génome en raison d'un transfert horizontal de gènes, par exemple. L'abondance de ces espèces,

qui peut parfois être très déséquilibrée, est déterminée par la mesure dans laquelle elles sont couvertes par les lectures métagénomiques.

En outre, chaque espèce est représentée par un nombre variable d'individus, dont chacun peut présenter une variété de génotypes, tels que des variantes courtes et des différences structurelles. Les données métagénomiques offrent donc un très grand nombre de variations, contrairement à la ploïdie connue des données génomiques. Ces nombreuses versions peuvent également être mesurées en mesurant leur couverture. Selon le niveau d'effort de séquençage utilisé, certaines variantes rares peuvent être confondues avec des erreurs de séquençage ou des régions appartenant à une autre espèce de la communauté. Il est donc difficile de comprendre cette diversité, en particulier lorsqu'il s'agit de comparer différentes communautés qui peuvent abriter différentes espèces. Ce polymorphisme rend difficile les tâches génomiques standards, comme l'assemblage où la recherche de variantes est effectuée, ce qui nécessite le développement d'algorithmes dédiés [9].

6. Flux de travail dans la métagénomique (Workflow)

Que ce soit dans l'établissement du plan d'expérience, l'utilisation de différents protocoles techniques ou solutions analytiques est important, chaque choix du biologiste a un impact sur la figure qu'il obtiendra de la composition de ses échantillons après analyse (Figure 7) [5].

Le protocole principal d'une étude métagénomique :

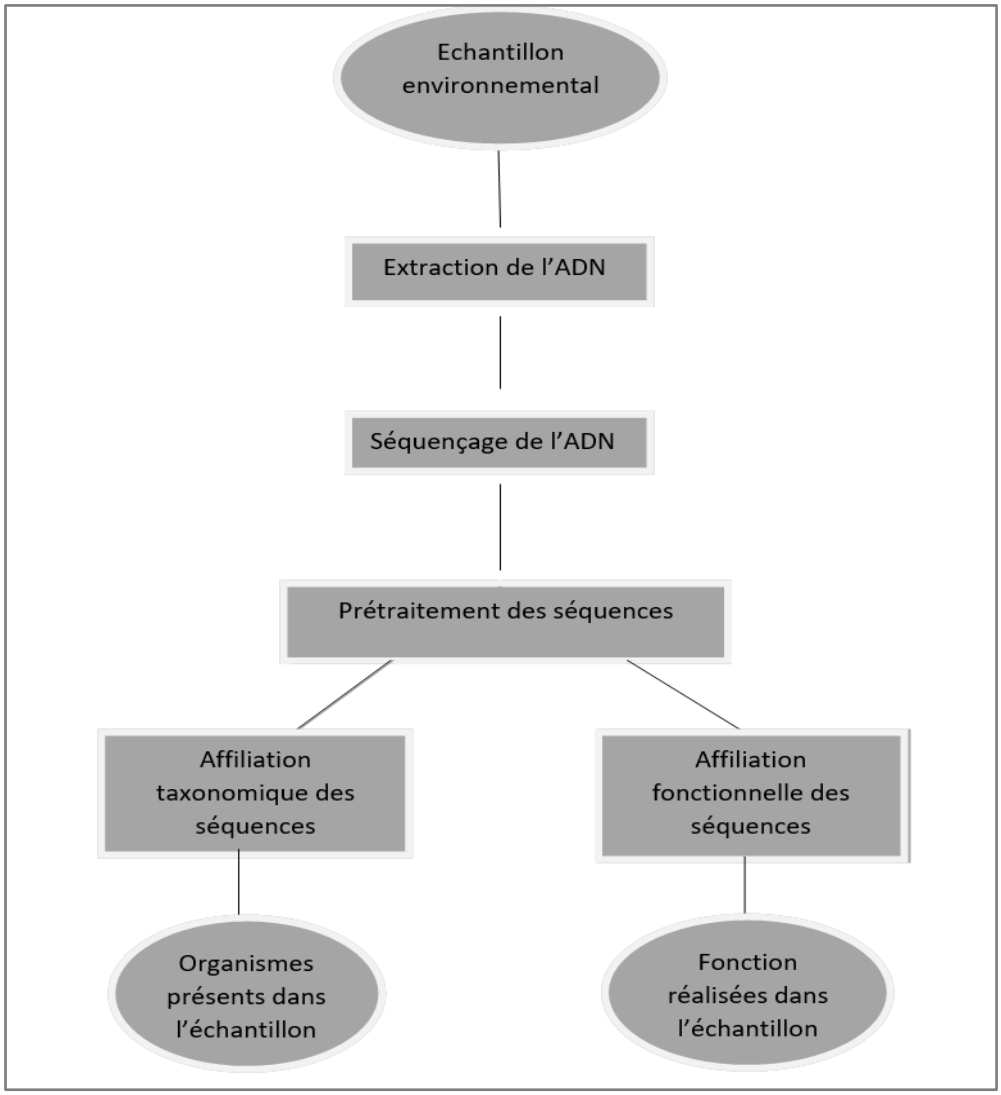

Figure 7: Schématisation des étapes de construction et d'analyse des librairies Métagénomiques.

6.1. Plan d'expérience

Le plan d'expérience est une étape fondamentale de toute étude métagénomique. Définir une question biologique claire est indispensable pour monter un plan d'expérience robuste qui permettra au biologiste d'y répondre. Le nombre d'échantillons à séquencer, les variables à évaluer, la profondeur de séquençage, le nombre de réplicas sont autant de facteurs à prendre en compte ; ces derniers pourront guider le choix des technologies, les protocoles techniques et les méthodes d'analyse les plus adaptées [5].

6.2. Collecte d'échantillons

La première étape consiste à prélever et traiter les échantillons à séquencer. Une composition microbienne peut évoluer très rapidement en passant de son milieu chaud anaérobie d'origine à un milieu à température ambiante. L'idéal est ainsi d'extraire l'ADN sitôt de l'échantillon prélevé. Toutefois, la nécessité de transport ou de stockage lors d'études à grandes échelles impose souvent une nécessité de préservation pour figer la composition du microbiote. Ces méthodes de préservation peuvent tout de même altérer la composition des échantillons, certaines études décrivant une composition différente entre un échantillon congelé et un échantillon frais. En outre, effectuer des prélèvements dans des conditions non stériles peut augmenter le risque d'introduction d'ADN exogène contaminant dans les échantillons d'intérêt [5].

6.3. Extraction d'ADN

Une fois tous les échantillons disponibles, une étape d'extraction et de purification permettra d'accéder à l'ADN des organismes présents, idéalement sans contaminer l'hôte de l'ADN en cas de prélèvement organique. L'ADN doit être extrait dans une quantité suffisante pour la préparation de la librairie de séquençage, et dans des proportions respectant celles des organismes dans le milieu. Comme il est déjà mentionné auparavant, il existe deux approches pour analyser un échantillon métagénomique [5].

6.4. Séquençage de l'échantillon

Le séquençage de l'ADN est devenu un outil clé en biologie moléculaire, utilisé en médecine et dans de nombreux autres domaines des sciences de la vie. Cette procédure est une technique standard utilisée dans les laboratoires de biologie moléculaire. La capacité de séquencer de grands génomes et les connaissances acquises grâce à cette méthode ont encouragé les chercheurs à développer de plus en plus de techniques de séquençage. L'une de ces technologies est le séquençage de deuxième génération, une approche révolutionnaire décrite dans la section (3.1). Cette technologie utilise également des méthodes de séquençage à haute résolution actuellement disponibles pour les laboratoires. Il existe deux grandes catégories de méthodes de séquençage optimisées pour la lecture de fragments courts d'ADN (50 à 300 Pb, également appelé séquençage à lecture courte). La première catégorie est le séquençage par synthèse, utilisé par des technologies telles qu'Illumina et Ion Torrent (Figure 10). La deuxième catégorie

est le séquençage par épissage, utilisé par la technologie SOLiD. Le séquençage est basé sur l'amplification clonale par réaction en chaîne par polymérase (PCR) d'une bibliothèque de fragments d'ADN à séquencer. Le résultat de cette amplification clonale est un modèle de séquence [5].

Figure 8: Séquenceurs haut-débit [5].

6.5 Prétraitement des séquences

Le prétraitement des données de séquençage métagénomique est une étape cruciale pour garantir la qualité des données et assurer une analyse précise des communautés microbiennes présentes dans l'échantillon. Les principales étapes du prétraitement comprennent :

Débruitage: Eliminer le bruit de fond (erreurs de séquençage) et améliorer la qualité des données en éliminant les reads qui ne sont pas représentatifs de la communauté microbienne étudiée tout en préservant les reads informatifs.

Filtrage par qualité : Eliminer les reads de mauvaise qualité en fonction d'un seuil de qualité fixé.

Élimination des reads dupliqués : Eliminer les reads en double qui peuvent être générés pendant la préparation de la bibliothèque.

Élimination des reads chimériques : Eliminer les reads qui sont le résultat de la fusion artificielle de deux ou plusieurs séquences distinctes pendant la réaction de PCR.

Trim des séquences : Eliminer les bases de mauvaise qualité et les adaptateurs présents aux extrémités des reads [5].

6.6. Affiliation fonctionnelle de séquences

Consiste à assigner des fonctions biologiques à des séquences d'ADN ou d'ARN issues de communautés microbiennes présentes dans un échantillon environnemental. Cette étape est Essentielle pour comprendre les fonctions écologiques et métaboliques des micro-organismes présents dans l'échantillon.

6.7. Affiliation taxonomique de séquences

Est une étape importante dans l'analyse des données métagénomiques. Elle permet d'identifier les organismes présents dans l'échantillon et de déterminer leur abondance relative [15].

7. Taxonomie

Les chercheurs découvrent continuellement de nouvelles espèces et les estimations actuelles suggèrent qu'il pourrait y avoir jusqu'à 100 millions d'espèces différentes sur terre. Pour garder une trace de la relation entre les espèces, un système de classement biologique est utilisé et à ce jour, environ 2 millions d'espèces ont été décrites et cataloguées.

La science de la classification des espèces s'appelle la *taxonomie*. L'arbre taxonomique est divisé en plusieurs niveaux. Il existe de nombreuses façons de définir l'arbre taxonomique, cependant, il est généralement divisé en sept niveaux illustrés à la (figure 11). Ces niveaux sont organisés comme des "boîtes dans une boîte", ce qui signifie que les catégories plus larges sont successivement divisées en catégories plus étroites créant une hiérarchie taxonomique. Ici le royaume (Kingdom ou Domaine) est la catégorie la plus large et chacun des royaumes est divisé en plusieurs groupes au niveau du phylum. De plus, chaque phylum est divisé en plusieurs classes, et ainsi de suite jusqu'à atteindre les niveaux plus spécifiques genre (Genus) et enfin espèces. Traditionnellement, les organismes sont classés et ordonnés dans l'arbre taxonomique selon leurs caractéristiques. Cependant, pour les procaryotes (bactéries et archées), il est difficile de distinguer avec précision les différentes espèces uniquement en fonction de leurs caractéristiques et il est donc nécessaire de considérer d'autres caractéristiques telles que la structure génétique [16].

8. Principes de la taxonomie

8.1. Rangs taxonomiques

Il existe sept niveaux de hiérarchie pour les êtres vivants : Espèce, Genre, Famille, Ordre, Classe, Phylum (Embranchement) et Royaume dont l'unité de base de la classification est l'espèce. Contrairement aux Eucaryotes, où plusieurs phylums constituent un règne, les Archées et les Bactéries n'ont pas de règnes reconnus. Des niveaux intermédiaires sont parfois utilisés, comme le sous-embranchement, la famille (tribu) et l'espèce [17].

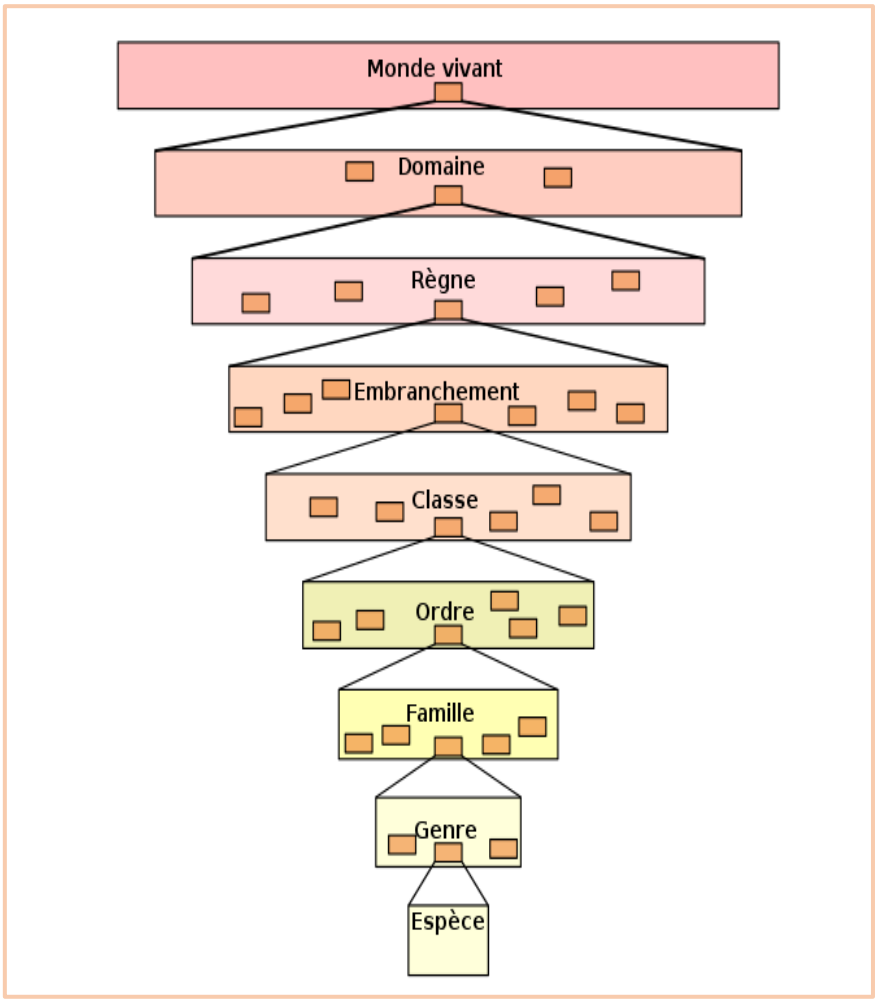

Figure 9: Pyramide des rangs taxonomiques utilisés dans la classification du vivant, du plus large, le domaine, au plus précis, l'espèce [18].

8.2. Classification taxonomique

La classification taxonomique est le processus de classification et de nommage des organismes vivants en groupes hiérarchiques en fonction de leurs caractéristiques et de leur relation évolutive. Elle vise à établir une hiérarchie de groupes d'organismes qui reflètent leur parenté évolutive. La taxonomie comprend plusieurs niveaux hiérarchiques, allant des catégories les plus générales (royaume, règne) aux plus spécifiques (espèce, souche). Le mieux est d'avoir un classifier taxonomique capable d'assigner une séquence au rang taxonomique le plus bas possible (espèce). La classification taxonomique est utilisée pour organiser et comprendre la diversité du monde vivant. Principalement il existe deux approches pour la classification taxonomique [19].

A/ Approche basé sur la similarité

La première approche pour la classification taxonomique est basée sur l'alignement de chaque fragment sur une base de données de séquences de référence (exemple : **SILVA, Greengenes**). Cet alignement est réalisé par des algorithmes d'alignements le plus populaire BLAST. Le choix d'une banque de séquences de référence est crucial : cette banque doit être adaptée au locus cible d'intérêt, correctement annotée, aussi exhaustive que possible et doit suivre une taxonomie standardisée. Il existe trois banques principales de référence pour l'ADNr 16S bactérien dont les caractéristiques sont résumées dans le Tableau 2 [14].

Tableau 2: Comparaison des trois principales banques de séquences d'ADN ribosomique [5].

	SILVA SSU Parc	SILVA SSU Ref	Greengenes	RDP
version actuelle	138(Avril 2023)	138(Avril 2023)	13.5 (mai 2013)	11.5 (septembre 2016)
Organismes	Bactéries, archées, eucaryotes	Bactéries, archées, eucaryotes	Bactéries,archées	11.5 (septembre 2016)
Origine des séquences	European, Nucleotide Archive	European, Nucleotide Archive	genbank	European Nucleotide Archive
Nombre de séquences	5616941	1922213	1262986	3356809
Taille minimale des séquences	300	1200 (Bactéries/archées) 900 (eucaryotes)	1250	500
Sélection et validation des séquences	Alignement >= 50 % d'identité avec au moinsune autre séquence de la banque	Alignement >= 70 % d'identité avec au moins une autre séquence de la banque	Score d'alignement positif avec au moins une autre séquence de la banque + élimination des séquences chimériques	Au moins 30 % de 7 - mers partagés avec une autre séquence de la banque + score d'alignement positif sur un alignement de référence
Taxonomie	SILVA [Yilmaz et al. 2013]	Au moins 30% de k-mers partagés avec au moinsune autre séquence de la banque+score d'alignement positif sur un alignement de référence	Greengenes [McDonald et al. 2012]	RDP [Cole et al. 2014]
licence	Utilisation gratuite académique/non-commerciale Licence payante non académique/commerciale	RDP [Cole et al. 2014]	Créative Commons BY-SA 3.0	Créative Commons BY - SA 3
Référence	[Quast et al. 2013]	Créative Commons BY-SA 3.0	[DeSantis et al. 2006]	[Cole et al. 2014]

Cette approche ne prend pas en compte la possibilité d'un décalage entre une lecture (d'un génome inconnu par exemple) et les génomes présents dans la banque de référence, ce qui pourrait généraliser de fausses affectations taxonomiques trop précises. Par exemple, en regardant l'apparence des alignements sur la Figure 12. À, la lecture pourrait tout aussi bien être attribuée à l'espèce A2 - le score inférieur de l'alignement entre la lecture et A2 pourrait simplement être la conséquence d'une séquence A2 tronquée dans la base de données qui ne couvre pas le début de la lecture.

Une interprétation plus fine des alignements a été rendue possible par l'algorithme LCA (Lowest Common Ancestor), introduit dans MEGAN, puis intégré dans de nombreux pipelines. Cet algorithme interprète, pour chaque lecture, une sélection de plusieurs hits BLAST (A1, A2, A3) validés comme significatifs sur la base de leur score. L'algorithme LCA attribue la lecture au taxon qui est l'ancêtre commun le plus bas parmi les résultats significatifs (A).

Cette approche, automatisant les alignements et leur interprétation, souffre toutefois d'un délai important entre soumission des lectures et réception des résultats : en 2015, le temps d'attente médian était entre 7 et 10 jours pour un échantillon Shotgun, et 24h pour un échantillon Amplicon .Ce délai est dû au temps d'analyse est démultiplié par la popularité du pipeline, recevant actuellement 4 téra paires de base de séquences à analyser par mois, ce qui impose une file d'attente de plus en plus longue à ses utilisateurs [14].

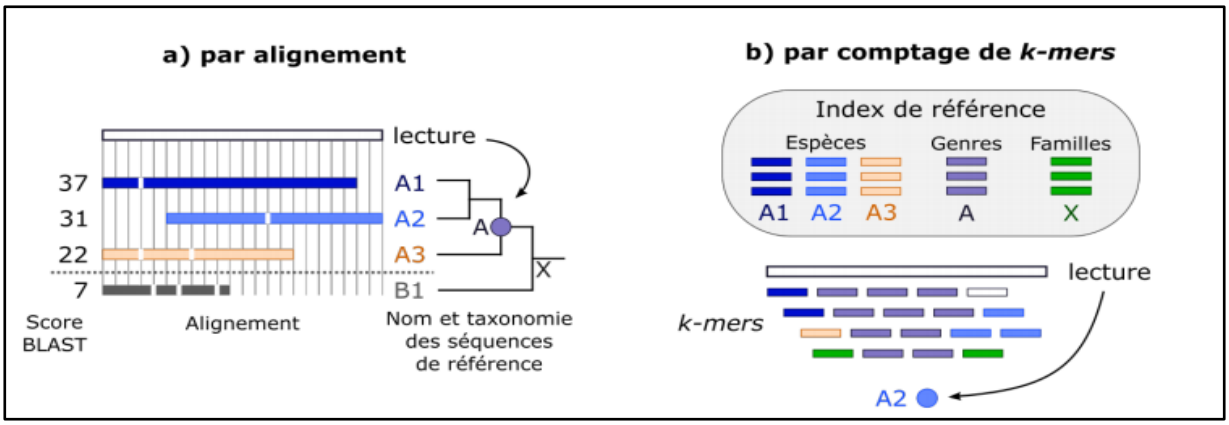

Figure 10: Méthodes taxonomiques utilisables en métagénomique [14].

B/ Approche basé sur la composition

Dans cette approche, les séquences qui partagent les mêmes signatures génomiques telles que les fréquences de k-mers ou le pourcentage guanine-cytosine (GC), pourraient être regroupées dans le même groupe taxonomique. En ce qui concerne les méthodes basées sur la fréquence des k-mers l'identification des espèces présentes dans un échantillon se fait en comparant leur profil de k-mers avec une base de données de référence contenant des profils de k-mers pour un ensemble connu d'espèces.

Les k-mers sont des sous-séquences de longueur k d'un génome ou d'un fragment d'ADN. Par exemple, pour un k = 4 et pour la séquence ACGTACG, les k-mers possibles sont ACGT, CGTA, GTAC, TACG, etc. La fréquence d'occurrence de chaque k-mer dans un échantillon est calculée et utilisée pour générer un profil de k-mers unique pour chaque organisme. Ces profils de k-mers sont ensuite comparés à une base de données de référence contenant des profils de k-mers pour des espèces connues afin d'identifier les organismes présents dans l'échantillon.

L'approche basée sur la composition a plusieurs avantages par rapport à aux méthodes de classification basées sur l'alignement des séquences, telles que la vitesse de traitement et la sensibilité à la détection des espèces rares ou peu représentées dans l'échantillon. Cependant, elle peut être influencée par la qualité des données de séquençage et le taux d'erreur de séquençage, ainsi que par le choix de la valeur de k utilisée pour l'analyse.

Cette approche est largement utilisée dans la communauté de la génomique comparative et de la métagénomique pour classer les communautés microbiennes et pour étudier la biodiversité des écosystèmes. Des outils tels que Kraken, Kaiju et DIAMOND sont couramment utilisés pour effectuer la classification basée sur la composition taxonomique k-mers [20].

CHAPITRE 2 : APPRENTISSAGE PROFOND (DEEP LEARNING)

1. Introduction

Le cerveau humain est capable de traiter de grandes quantités d'informations et de s'adapter à de nouvelles situations grâce à des processus cognitifs complexes. Ces processus sont la base de notre intelligence et de notre capacité à apprendre de nouvelles choses. L'intelligence artificielle a pour objectif de créer des systèmes informatiques qui imitent ces processus cognitifs pour résoudre des problèmes complexes. L'apprentissage automatique (machine Learning) est l'une des méthodes clés de l'intelligence artificielle qui permet aux ordinateurs d'apprendre à partir de données, tout comme le cerveau humain apprend de l'expérience [21].

2. L'apprentissage automatique (machine Learning)

L'apprentissage automatique (Machine Learning) est un sous-domaine de l'intelligence artificielle (IA) qui utilise des algorithmes pour apprendre à partir de données. Il s'appuie sur la théorie de l'apprentissage computationnel et la reconnaissance de modèles pour entraîner ces algorithmes à classer et prédire des données futures. Les programmes informatiques sont capables de s'améliorer et d'évoluer par eux-mêmes en apprenant à partir de nouvelles données. Contrairement aux programmes classiques, l'apprentissage automatique ne suit pas des instructions statiques, mais construit un modèle à partir d'exemples d'entrées pour faire des prédictions ou des choix basés sur les données. La précision des prédictions dépend de la quantité et de la qualité des données utilisées pour entraîner les algorithmes [22].

2.1. Types d'apprentissages automatiques

L'apprentissage automatique est divisé en :

2.1.1. Apprentissage supervisé

Les données d'apprentissage fournies à l'algorithme comprennent les solutions, appelées étiquettes (labels). Une tâche d'apprentissage supervisée typique est la classification. Le filtre anti-spam en est un bon exemple : il est formé avec de nombreux exemples d'e-mails avec leur classe (spam ou ham), et il doit apprendre à classer les nouveaux e-mails. Voici quelques-uns des algorithmes d'apprentissage supervisé les plus importants : Knearst Neighbors, Régression linéaire, Régression logistique, Machines à vecteurs de support (SVM), Arbres de décision et forêts aléatoires [22].

2.1.2. Apprentissage non supervisé

Les données d'apprentissage ne sont pas étiquetées. Le modèle n'a pas de « réponses» dont il peut tirer des données en entrées; il doit donner un sens aux données en fonction des observations elles-mêmes. L'apprentissage non supervisé permet d'aborder les problèmes avec peu ou pas d'idée de ce à quoi les résultats devraient ressembler. La possibilité d'obtenir une structure à partir de données dont l'effet des variables n'est pas nécessairement connue.

Voici quelques-uns des algorithmes d'apprentissage non supervisé les plus importants :

-Clustering : K-Means, Analyse des clusters hiérarchiques (HCA), Maximisation des attentes.

-Visualisation et réduction de la dimensionnalité : Analyse en composantes principales (ACP), Kernel PCA, local linéoca laire embedding (LLE),

T-distribued Stochastic Neighbor Embedding (t-SNE).

- Apprentissage des règles d'association : APriori [23].

2.1.3. Apprentissage par renforcement

L'apprentissage par renforcement est une méthode d'apprentissage automatique qui permet à une machine, appelée agent, d'apprendre à prendre des décisions de manière autonome en interagissant avec son environnement. Contrairement à d'autres approches d'apprentissage, l'apprentissage par renforcement ne nécessite pas de supervision humaine ou de données pré-étiquetées. L'agent peut observer l'environnement, sélectionner et effectuer des actions, et recevoir des récompenses ou des pénalités en fonction des résultats de ses actions. En apprenant à maximiser ces récompenses au fil du temps. Une politique définit l'action que l'agent devrait choisir lorsqu'il est dans une situation donnée. Cette autonomie rend l'apprentissage par renforcement particulièrement intéressant pour les applications dans lesquelles les données ne sont pas facilement disponibles ou que les humains ne peuvent pas superviser directement [24].

2.2. Algorithme d'apprentissage automatique

L'algorithme d'apprentissage automatique est une évolution de l'algorithme régulier. Il rend les programmes « plus intelligents », en leur permettant d'apprendre automatiquement des données fournies. L'algorithme est principalement divisé en :

- Phase d'entraînement (d'apprentissage).
- Phase de test.

2.2.1. Phase d'entraînement

Consiste à sélectionner une partie des données d'entrée (appelée données d'entraînement) et à créer un tableau contenant toutes les caractéristiques pertinentes pour chaque entrée. Ces caractéristiques peuvent inclure des informations telles que la couleur, la taille, la forme, la longueur, le poids, etc. Ces données sont ensuite transmises à un algorithme d'apprentissage automatique (classification / régression) afin de trouver le modèle mathématique le plus approprié pour établir une corrélation entre les différentes caractéristiques.

2.2.2. Phase de test

Ou phase de vérification elle consiste à utiliser le modèle qui a été construit lors de la phase d'entraînement pour prédire les résultats pour une seconde partie de données, appelée données de test. L'objectif est de mesurer les performances du modèle en termes de précision et de qua-

lité de la prédiction. Toutefois, il est important de noter que les performances peuvent être impactées par l'over fitting, c'est-à-dire que le modèle peut avoir été surentraîné sur les données d'entraînement et ne pas généraliser correctement sur les données de test. Par conséquent,

Une etape d'optimisation est souvent nécessaire pour réduire les erreurs et améliorer les performances du modèle sur des données inconnues.

3. Apprentissage profond (Deep Learning)

Est un sous-domaine de l'apprentissage automatique, qui a été introduit pour la première fois par Dechter en 1986. La Figure 13 illustre que l'apprentissage profond est le résultat d'interactions entre les data science et la machine learning. Leurs algorithmes ont été influencés par la structure et le fonctionnement du cerveau humain [25].

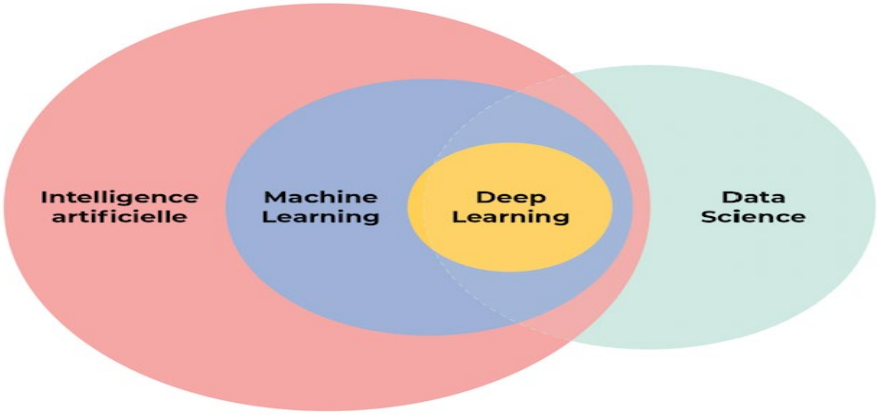

Figure 11: les domaines impliqués dans l'apprentissage profond [26].

L'apprentissage profond est une forme d'intelligence artificielle dérivée du Machine Learning. Il a été créé en s'inspirant des réseaux neuronaux (Neural Networks) qui se trouvent dans le cerveau humain. L'apprentissage profond est constitué d'un grand nombre de couches de neurones artificiels interconnectés. Plus le nombre de neurones est élevé, plus le réseau est qualifié de « profond » et délivre des performances exceptionnelles. Le réseau se compose d'une couche d'entrée, d'une ou plusieurs couches cachées et d'une couche de sortie. Dans chaque couche, il y a plusieurs nœuds, ou neurones, et les nœuds de chaque couche utilisent les sorties de tous les nœuds de la couche précédente comme entrées, de sorte que tous les neurones sont interconnectés les uns avec les autres à travers les différentes couches. Normalement, chaque neurone

se voit attribuer un poids qui s'ajuste (Figure 14) pendant le processus d'apprentissage, la diminution ou l'augmentation du poids modifie la force du signal de ce neurone. Les réseaux de neurones peuvent être utilisés pour l'apprentissage supervisé (classification, régression) et l'apprentissage non supervisé (reconnaissance de formes, regroupement) [27].

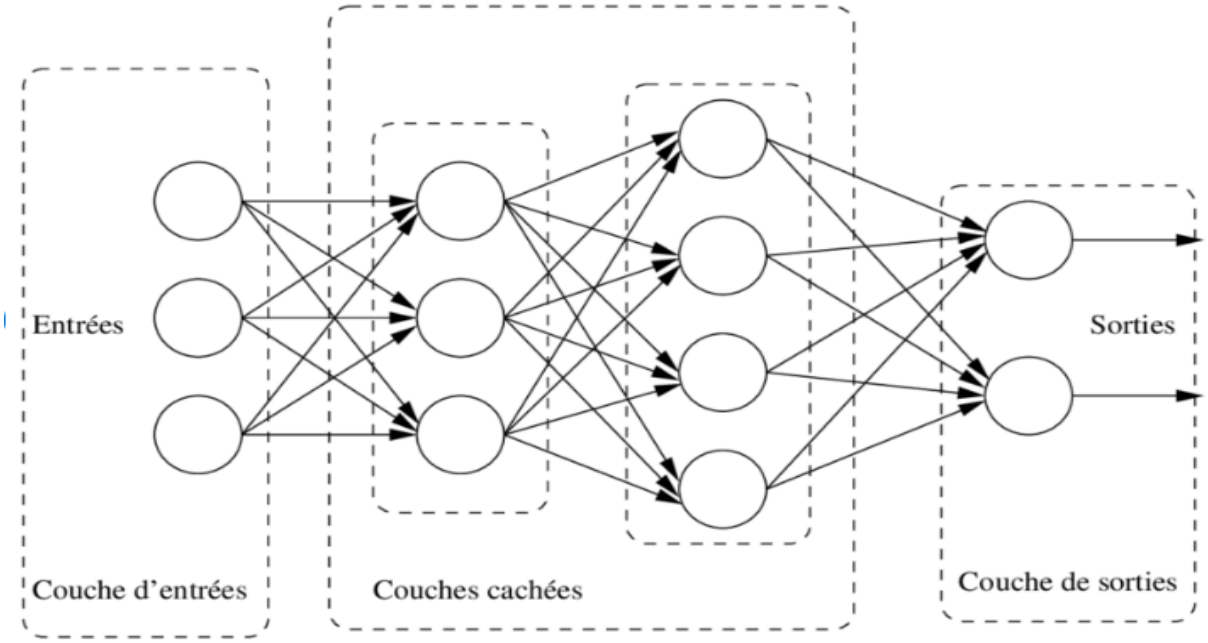

Figure 12: Architecture de réseau neuronal typique **[28].**

3.1. Modèles d'apprentissage profond

Les modèles d'apprentissage profond font référence à un type d'algorithme d'apprentissage automatique qui utilise des réseaux de neurones profonds pour apprendre des représentations de données complexes. Voici quelques exemples de modèles d'apprentissage profond [29] :

3.1.1. Perceptron Multicouche (Multi Layer Perceptron)

Il se compose d'une couche d'entrée, d'une ou plusieurs couches cachées et d'une couche de sortie (un réseau de neurones peu profond typique) [30].

3.1.2. Réseau de neurones convolutif (CNN)

Une architecture de réseau neuronal profond est largement appliquée au traitement d'image et comporte des couches convolutifs qui transmettent des fenêtres à travers l'entrée avec des nœuds qui partagent des poids, en supprimant l'entrée (généralement l'image) pour présenter des cartes [31].

3.1.3. Réseau de neurones récurrent (RNN)

Une architecture de réseau de neurones avec des boucles de rétroaction qui modélise les dépendances séquentielles en entrée, telles que les séries chronologiques, les capteurs et les données textuelles ; Le type le plus courant de RNN est un réseau de mémoire à long terme (LSTM) [32].

3.1.4. Modèles basés sur le traitement automatique du langage naturel (TALN)

Les modèles d'apprentissage en profondeur sont une classe de modèles de TALN qui utilisent des réseaux de neurones profonds pour apprendre à traiter le langage naturel. Ils sont capables d'apprendre des représentations de haute qualité du langage naturel à partir de données non étiquetées et étiquetées, ce qui leur permet d'effectuer de nombreuses tâches de TALN avec une grande précision. Ces modèles sont utilisés dans de nombreuses applications de TALN, mais nécessitent de grandes quantités de données et de ressources de calcul pour être entraînés correctement [33**].** Cette partie sera plus détaillée dans la section du Traitement automatique du Langage Naturel (TALN).

4. Domaines d'applications de Deep Learning

De nos jours, l'apprentissage profond sert à développer de nombreuses technologies révolutionnaires. Celles-ci sont très utiles dans la vie de tous les jours.

4.1. Dans le domaine de la médecine

L'IA de Deep Learning permet de distinguer les tumeurs cancéreuses de celles qui ne le sont pas. Elle scanne les photos de radiographie avec une plus grande précision que l'œil humain et permet donc d'anticiper la prise en charge pour allonger les chances de guérison du malade.

De plus, le Deep Learning aide aussi au diagnostic, la chirurgie assistée par ordinateur, les robots médicaux, la médecine prédictive, l'anticipation d'une épidémie, le triage des patients, le développement de nouveaux traitements [34].

4.2. Dans le secteur de l'agriculture

L'agriculture biologique s'appuie sur des drones intelligents capables d'identifier les mauvaises herbes en scannant au survol plusieurs hectares de plantation. Cela permet aux agriculteurs de concentrer uniquement leur énergie sur les zones qui nécessitent un désherbage [34].

4.3. Dans le domaine biologique

En biologie computationnelle, l'apprentissage profond est utilisé en génomique régulatrice pour l'identification de variantes régulatrices, l'effet de la mutation en utilisant la séquence d'ADN, l'analyse de cellules entières, de populations de cellules et de tissus [34].

5. Traitement automatique du Langage Naturel (TALN)

Le traitement automatique du langage naturel (TALN) est un domaine de recherche en intelligence artificielle qui vise à permettre aux machines de comprendre et de produire du langage naturel tel qu'il est utilisé par les humains. Le TALN utilise des techniques d'apprentissage automatique, de traitement du signal, de statistiques et de linguistique pour analyser et comprendre le langage naturel.

Le TALN peut être utilisé pour analyser et comprendre le langage naturel dans les textes scientifiques, les rapports de laboratoire, les dossiers médicaux et autres documents. Par exemple, il peut être utilisé pour extraire des informations importantes sur les gènes, les protéines et les voies métaboliques à partir de vastes collections de documents scientifiques. Le TALN (Traitement Automatique du Langage Naturel) joue un rôle crucial dans la compréhension des questions liées à la biologie évolutive, à la phylogénie et à la génomique comparative grâce à l'analyse des séquences d'ADN et des annotations fonctionnelles associées. En utilisant des techniques de TALN, il est possible d'extraire des informations pertinentes à partir des vastes ensembles de données génomiques, de faciliter la comparaison et l'alignement des séquences d'ADN, ainsi que d'identifier des motifs et des régions fonctionnelles spécifiques. De plus, le TALN peut contribuer à l'annotation automatique des génomes en associant des fonctions biologiques aux séquences d'ADN. En somme, le TALN est un domaine de recherche en plein essor qui ouvre de nombreuses perspectives d'analyse et de compréhension du langage naturel, en particulier dans les domaines de la biologie et de la médecine. Sa relation avec la génétique réside dans sa capacité à exploiter et à interpréter les données génomiques pour générer de nouvelles connaissances et éclairer notre compréhension des processus biologiques [35].

6. Comment fonctionne le traitement du langage naturel

Les approches actuelles du TALN sont basées sur le deep learning. Les modèles de deep Learning nécessitent d'énormes quantités de données étiquetées pour s'entraîner et identifier les corrélations pertinentes, et l'assemblage de ce type d'ensemble de données volumineuses est actuellement l'un des principaux obstacles du TALN.

Les approches antérieures du TALN impliquaient une approche plus basée sur des règles, dans laquelle des algorithmes d'apprentissage automatique plus simples étaient informés des mots et des phrases à rechercher dans le texte et recevaient des réponses spécifiques lorsque ces phrases apparaissaient. Mais l'apprentissage en profondeur est une approche plus flexible et intuitive dans laquelle les algorithmes apprennent à identifier l'intention des locuteurs à partir de nombreux exemples, presque comme la façon dont un enfant apprendrait le langage humain. Trois outils couramment utilisés pour le TALN sont NLTK, Gensim et Intel TALN Architect. NTLK (Natural Language Tool kit), NTLK, est un module python open source avec des ensembles de données et des didacticiels. Gensim est une bibliothèque Python pour la modélisation de sujets et l'indexation de documents. Intel TALN Architect est également une autre bibliothèque Python pour les topologies et techniques d'apprentissage en profondeur [36].

7. Applications du traitement automatique du langage naturel

Le traitement automatique du langage naturel (TALN) est un domaine de recherche en informatique qui se concentre sur la compréhension et la manipulation du langage naturel par des machines. Il offre de nombreuses applications dans divers domaines, exploitant les avancées technologiques pour analyser et interpréter le texte de manière automatisée. Deux applications courantes du TALN sont la classification des textes et l'analyse du sens [37].

7.1. Classification des textes

La classification de texte est un processus de classification de morceaux de texte en différentes catégories. C'est l'une des tâches TALN les plus simples mais les plus largement utilisées [39].

7.2. Opinion mining (analyse de sentiments)

L’opinion mining, est un type particulier de classification de textes, sa particularité réside dans sa capacité à identifier automatiquement des informations subjectives, telles que des opinions, des émotions ou des sentiments dans le texte. L'une des tâches les plus élémentaires de l’opinion mining est la classification de la polarité, c'est-à-dire de classer si l'opinion exprimée est positive, négative ou neutre [40].

8. Relation entre le traitement automatique du langage naturel et les autres approches d'apprentissage automatique

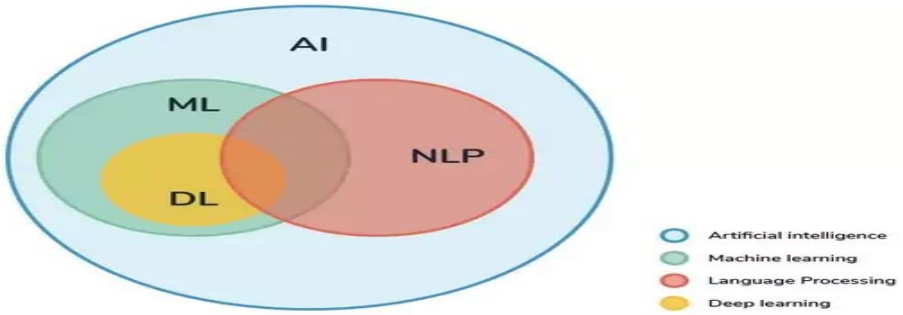

Figure 13: Relation entre DL, ML et NLP [41].

Le TALN accompagnés par les modèles adéquats du ML et du DL, en imitant la réflexion humaine, permet de créer des systèmes intelligents, capables de comprendre l'intention derrière le texte d'un utilisateur et de leur fournir une réponse appropriée. Les séquences génomiques sont représentées par des chaînes de caractères ou chaque caractère représente une base nucléique. Les séquences d'ADN ne sont donc qu'un simple texte d'un point de vue purement informatique, ce qui offre la possibilité de faire usage des techniques du NLP afin d'étudier les séquences d'ADN. Par ailleurs ce travail est basé sur l'une des techniques du NLP pour représenter les séquences d'ADN (GloVe) [1].

9. Les intégrations de mots (Word embidding, Plongement lexical)

Les intégrations de mots (Word embedding, Plongement lexical) sont une méthode de représentation de mots sous forme de vecteurs numériques denses, qui capturent les relations sémantiques et syntaxiques entre les mots. Voici quelques-unes des techniques d'intégration de mots les plus populaires :

9.1. Word2Vec

Une technique d'intégration de mots développée par Google, qui utilise des réseaux de neurones pour apprendre des représentations de mots denses en utilisant des modèles de langage. Les deux architectures principales de Word2Vec sont CBOW (Continuous Bag of Words) et Skip-gram [42].

9.1.1. CBOW

Le modèle CBOW (Continuous Bag-of-Words) est juste l'opposé de Skip-Gram. Pour le modèle CBOW, la tâche du réseau de neurones simple est la suivante : Étant donné un contexte de mots (entourant un mot) dans une phrase, le réseau prédira la probabilité que chaque mot du vocabulaire soit le mot [43].

9.1.2. Skip-Gram

Pour le modèle Skip-Gram, la tâche du réseau de neurones simple est la suivante : Étant donné un mot d'entrée dans une phrase, le réseau prédira la probabilité que chaque mot du vocabulaire soit le mot voisin de ce mot d'entrée [43].

9.2. Doc2Vec

Une extension de Word2Vec qui permet également de représenter des documents sous forme de vecteurs numériques denses. Elle utilise une variante de l'architecture Skip-gram qui prend également en compte le contexte global du document [44].

9.3. GloVe (Global Vectors for Word Representation)

Est une méthode d'intégration de mots développée par l'Université de Stanford, qui utilise des statistiques de co-occurrence de mots dans un corpus de texte pour apprendre des vecteurs de mots. Elle utilise une approche de factorisation matricielle pour apprendre des représentations de mots qui capturent la similarité sémantique et la syntaxe entre les mots [40]. Contrairement aux autres méthodes qui permettent d'étudier le contexte local GloVe permet de combiner les techniques qui étudient le contexte local et le contexte global [45].

10. Techniques basées sur NLP pour le traitement des séquences génomiques

Plusieurs méthodes d'encodage de séquence d'ADN existent, ces méthodes sont principalement basées sur trois modèles de traitement automatique du langage naturel populaires qui sont word2vec, fasttext et GloVe.

10.1. Méthodes basées surWord2Vec

Contient **Gen2vec, Biovec, Protec** et **Dna2vec** qui sont des algorithmes captés pour travailler avec des données spécifiques telles que des séquences génomiques, des protéines ou des textes biologiques. Ces extensions permettent de créer des embeddings spécialisés pour ces types de données, ce qui facilite leur analyse et leur utilisation dans des tâches de classification ou de prédiction [42] [43] [44] [45].

10.2. Méthodes basées sur Doc2Vec

Contient **Seq2Vec** comme extension qui est un algorithme qui permet de représenter des séquences, telles que des phrases ou des paragraphes, sous forme de vecteurs numériques.

Cette extension est particulièrement utile pour l'analyse de texte et la classification de documents [44].

10.3. Méthodes basées sur FastText

Sont des extensions de l'algorithme **FastText** adaptées pour travailler avec des données spécifiques telles que des séquences génomiques ou des données textuelles volumineuses. FastDNA permet de représenter des séquences d'ADN sous forme de vecteurs numériques, tandis que LSHvec utilise des hachages localement sensibles pour accélérer la recherche de similitude entre les vecteurs denses [46].

10.4. Méthodes basées sur GloVe

Sont des algorithmes qui permettent de représenter des séquences d'ADN sous forme de vecteurs numériques denses. **GlvDNA** permet l'analyse de séquences d'ADN et la classification de données génomiques. Quant à **glvProtéine** permet l'analyse des séquences protéiques et la prédiction de leur structure et de leur fonction. Les embeddings générés par **glvDNA** et **glvProtéine** peuvent être utilisés pour diverses tâches d'analyse de séquences biologiques [45].

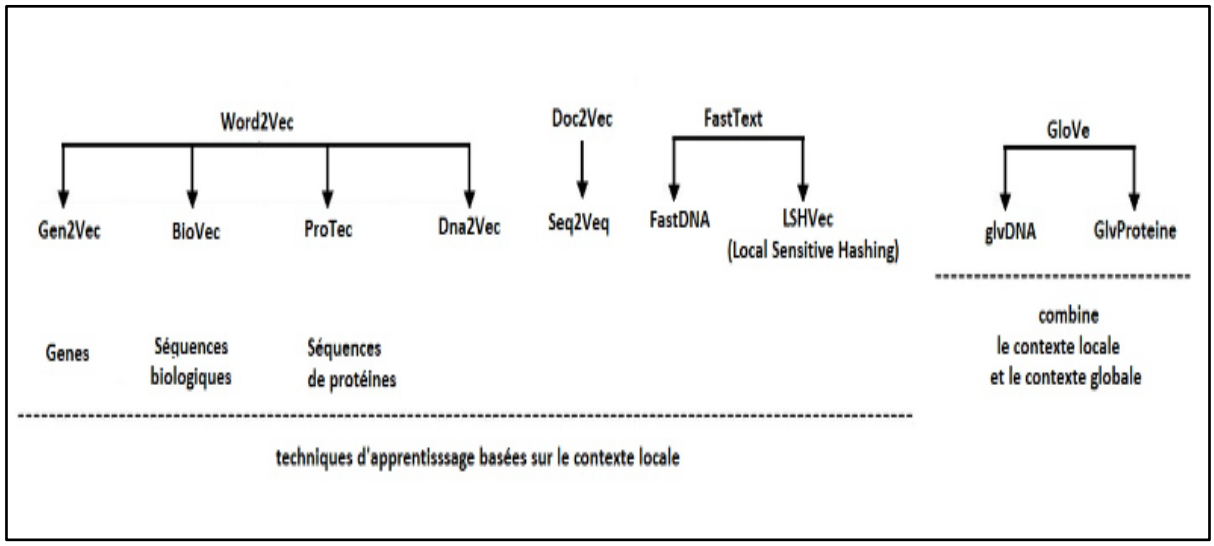

Figure 14: les différentes techniques d'apprentissage basées sur le contexte local et celle qui combine le contexte locale et globale.

CHAPITRE 3 : CONTRIBITION

1. Introduction

La classification taxonomique est une méthode utilisée pour analyser l'ADN environnemental dans le but d'identifier les organismes présents dans un échantillon provenant de l'environnement. Elle se base sur des techniques de séquençage haut débit et permet d'étudier la diversité et la fonction des communautés microbiennes dans des écosystèmes complexes. Les méthodes d'apprentissage automatique, notamment les réseaux de neurones profonds, sont de plus en plus employés dans la classification taxonomique en raison de leur aptitude à apprendre à partir de vastes ensembles de données et à modéliser des schémas complexes [47]. Dans ce travail nous avons tiré profit de la ressemblance entre le langage naturel et les séquences de l'ADN qui sont tous les deux basés sur les caractères, (a, b, cd,...xyz) et (A, T, G, C) pour proposer une représentation des lectures métagénomiques basée sur les plongements lexicaux. Ces plongements lexicaux sont ensuite utilisés pour entraîner un modèle d'apprentissage profond de type LSTM afin d'obtenir un modèle capable de classifier des séquences métagénomiques. Pour entrainer ce modèle nous avons utilisé des séquences ARN 16s de la base de données SILVA. Ces séquences correspondent seulement à neuf espèces à cause de nos limites en termes de ressources matérielles. Nous subdiviserons cet ensemble de données en un ensemble d'apprentissage et un ensemble de test, puis entraînerons notre modèle sur l'ensemble d'apprentissage et l'évaluerons sur l'ensemble de test afin de mesurer sa précision de prédiction. Enfin, nous utiliserons notre modèle pour prédire les étiquettes de classe d'un Ensemble de données de validation. Ces séquences correspondent seulement à neuf espèces à cause de nos limites en termes de ressources matérielles.

2. Matériel et méthodes

2.1. Architecture de la solution

Notre système est partagé principalement en deux parties :

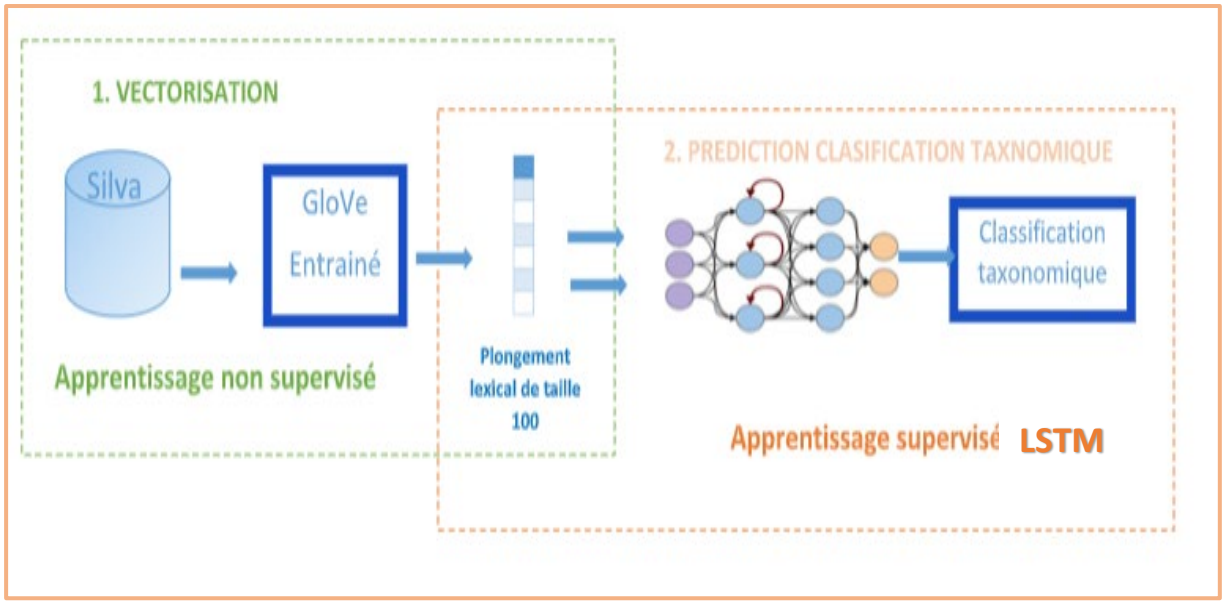

Figure 15: Architecture de la solution.

La première partie basée sur l'apprentissage non supervisé veut dire que les données d'entraînement ne sont pas étiquetées, cette étape consiste à implémenter une approche qui permet de représenter les séquences d'ARN 16S rRNA en représentation vectorielle de taille d en basant sur une méthode TALN.

Une deuxième partie basée sur l'apprentissage supervisé (les données sont étiquetées) tels que les vecteurs en résultats de la première partie seront utilisés pour entraîner un modèle d'apprentissage profond de type LSTM. Afin d'avoir une classification taxonomique.

2.2. Explication détaillée

Dans cette section on va détailler tous les éléments qui constituent notre architecture.

2.2.1. Base de données SILVA

La base de données SILVA est une ressource essentielle pour l'identification taxonomique des organismes grâce à ses séquences d'ARN ribosomal (Figure 16). Elle comprend des Séquences d'ARN ribosomal provenant de diverses espèces, accompagnées d'informations taxonomiques. SILVA est régulièrement mise à jour et largement utilisée dans les études de métagénomique. Elle fournit des ensembles de données complets et de qualité contrôlée, avec des séquences alignées des petites (16S/18S, SSU) et grandes sous-unités (23S/28S, LSU) d'ARN ribosomal pour les bactéries, les archées et les eucaryotes. La sélection de la version SSU r138.1 de SILVA pour l'identification des séquences, spécifiquement adaptée aux régions ARN 16S

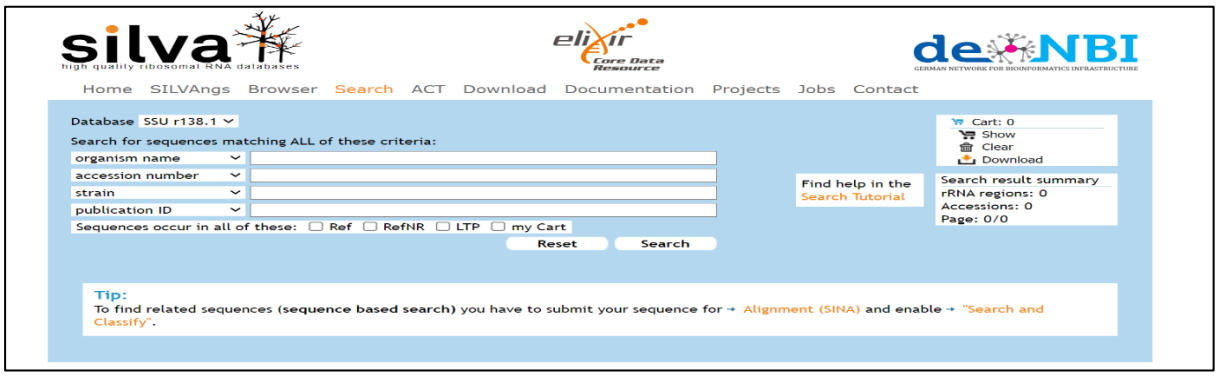

Figure 16: La base des données SILVA.

Pour accéder à ces séquences, il suffit de fournir le nom de la bactérie, son numéro d'identification taxonomique (ID) ou une séquence d'ARNr 16S connue. Spécifier une qualité de séquence supérieure à 70, puis téléchargez les séquences sans lacunes au format "zip" et les enregistrer sous format FASTA.

2.2.2. Prétraitement des données

Cette étape a pour but de s'adapter et de se transformer les fragments d'ARN métagénomiques en représentation vectorielle via une méthode de traitement du langage naturel basée sur les plongements et l'apprentissage profond, afin de préparer les données de la classification. Notre méthode proposée pour le traitement des données métagénomiques est effectuée comme suit :

I. Construire le corpus

Dans cette étape, nous construisons le corpus qui contient toutes les possibilités de concaténation de nucléotides dans un ensemble de données donné de la manière suivante

$$\boldsymbol{Corpus(dataset) = W1^K + W2^K + W3^K, \ldots\ldots Wi^K}$$

- **W est un mot**
- **K est la longueur du mot (k-mer)**
- **i est l'index du mot**

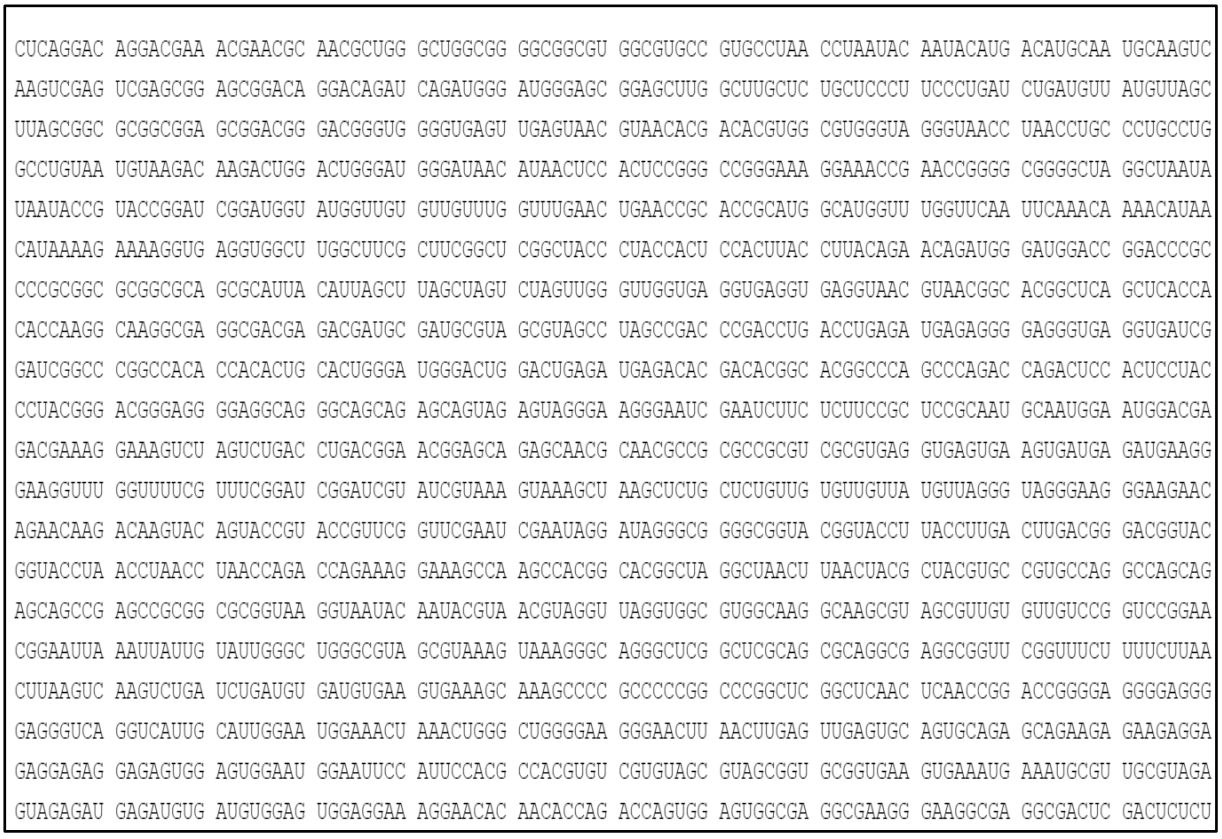

Figure 17: Fractionnement des fragments d'ARN 16 s.

Pour construire un corpus, nous prenons chaque séquence d'un fichier fasta, nous la nettoyons des caractères manquants (par exemple X, -, etc.). Ensuite, nous faisons glisser une fenêtre de longueur k (Figure 17) sur la séquence pour générer des k-mers non superposés séparés par des tabulations qui ont la même longueur.

II. **Entraînement du modèle Glove**

Nous avons entraîné un modèle GloVe pour qu'il puisse transformer des fragments d'ARN en représentation vectorielle. Ce modèle d'embedding de mots fait partie des algorithmes les plus connus pour le traitement du langage naturel. La méthode utilisée est basée sur l'utilisation de la cooccurrence mot à mot pour construire un modèle qui vise à apprendre les représentations vectorielles des mots composant une séquence. Les mots qui partagent des contextes similaires sont représentés par des vecteurs numériques proches. Nous avons utilisé le corpus généré à l'étape précédente pour entraîner le modèle embedding et obtenir un transformateur entraîné noté Vec. Ce transformateur fait correspondre un mot x ∈ Corpus à un espace vectoriel continu de taille d. Le vecteur résultant a une dimension de d= 100.

a) **Description des paramètres choisis**

Pour construire le modèle GloVe, de nombreux essais de configuration du modèle ont été testés. La meilleure configuration a été retenue en termes de complexité et de performance.

Tableau 3: Paramètres de réglage du modèle d'intégration.

Paramètres	Valeur	Description
VOCAB_MIN_COUNT	0	Les mots qui ont moins de occurrences que cette valeur sont ignorés
VECTOR_SIZE	15	Taille du vecteur d'embedding
MAX_ITER	15 100	Nombre d'itérations d'entraînement
WINDOW_SIZE	15	Nombre de mots de contexte à gauche

b) **Résultats de l'entraînement**

Durant l'apprentissage et à chaque itération une valeur de coût est affichée. Les valeurs de coût diminuent au fur et à mesure que l'entraînement progresse, ce qui indique que le modèle s'améliore progressivement. Initialement, le coût peut être relativement élevé (Figure 18), mais à chaque itération, le coût diminue jusqu'à atteindre une valeur minimale qui ne peut pas diminuer davantage.

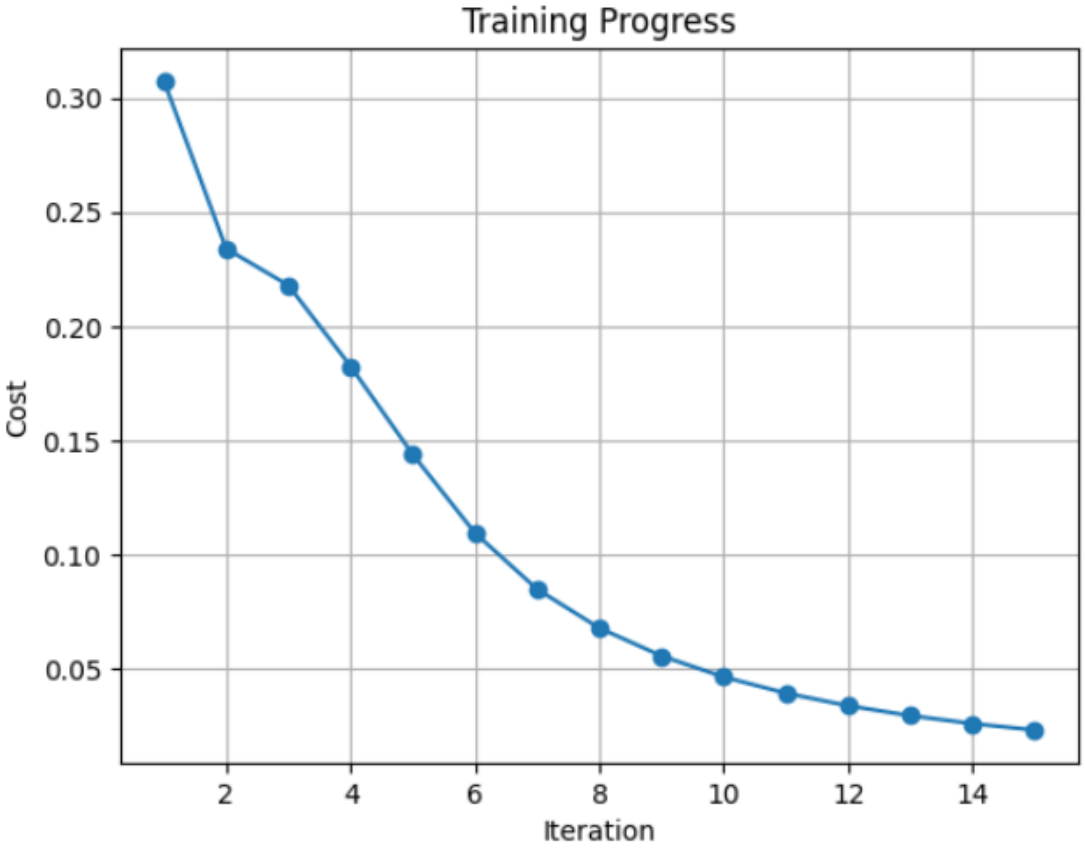

Figure 18: Resultants de l'entraînement.

c) **Transformer les fragments d'ARN en vecteur numérique à l'aide d'un modèle embedding entraîné**

Le modèle GloVe entraîné nous permet de représenter les fragments d'ARN 16S en une forme numérique (figure 19) que les algorithmes d'apprentissage profond peuvent traiter. Tout d'abord, nous avons divisé les fragments d'ARN 16S (X) en k-mers notés (Xi). Ensuite le vecteur numérique de chaque fragment (X) est généré en calculant la moyenne des vecteurs des plongements lexicaux comme suit :

$$Vec(X) = \sum_{i=1}^{N} \frac{\mathbf{Vec(Xi)}}{\mathbf{N}}$$

Où {X1, X2, X3.....XN} est un ensemble de mots de longueur k =8, Le vecteur résultant a une dimension de taille =100.

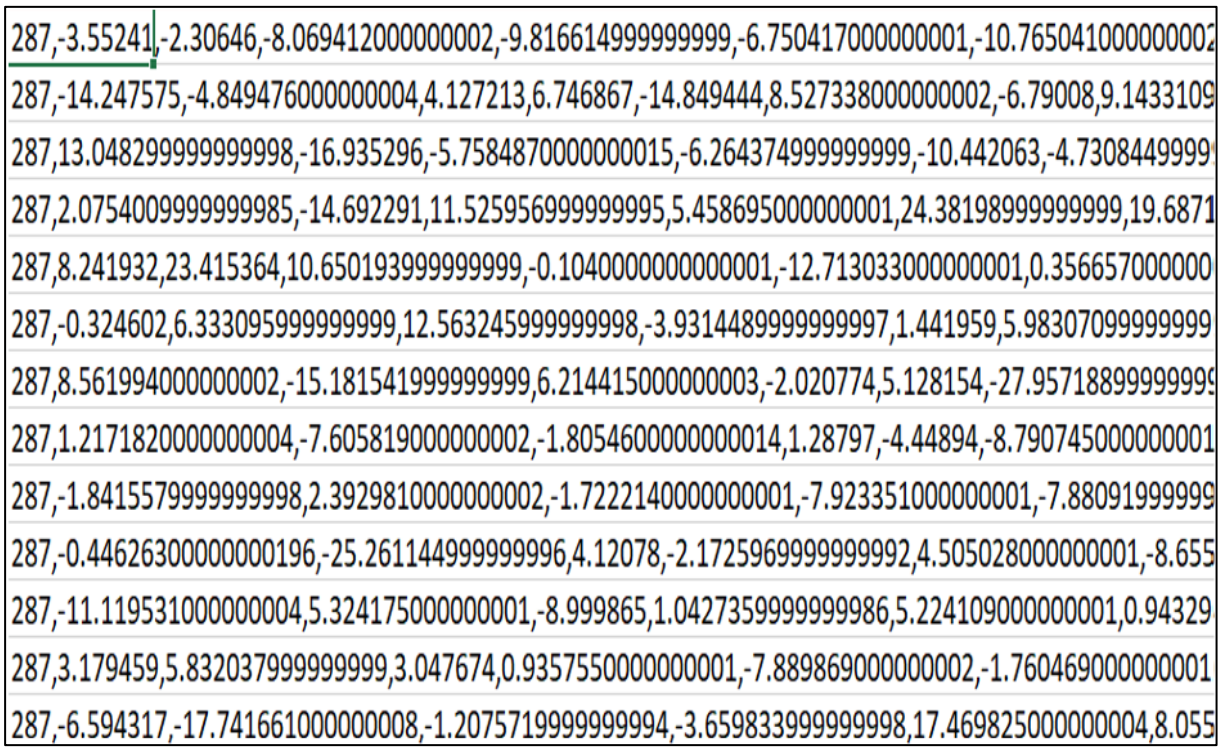

Figure 19: Représentation numérique des fragments (vecteurs).

2.2.3. Entraînement du modèle LSTM

Les LSTM (Long Short-Term Memory) sont des réseaux de neurones récurrents spéciaux qui excellent dans la capture de dépendances à long terme dans les séquences de données. Grâce à leur structure spécifique de cellule LSTM et à l'utilisation de portes pour réguler le flux d'informations, les LSTM sont devenus une méthode très populaire et efficace en apprentissage automatique. Ils sont particulièrement adaptés pour modéliser les relations temporelles complexes, ce qui en fait un outil puissant dans des domaines tels que la traduction automatique, la reconnaissance vocale et le traitement du langage naturel.

I. Composants du modèle LSTM

a) Répartition des données pour l'apprentissage et validation

Le fichier de données contient 27017 entrés qui sont les représentations numériques de 27017 fragments d'ARN, ces données sont divisées en deux ensembles (figure 20). Le premier représente 75% (soit 20263 fragments) des données. Elles ont servi à entraîner le modèle. La deuxième fraction du dataset représente les 25% (soit 6754 fragments) qui serviront à tester le modèle après chaque époque, ce qui permettra au modèle d'optimiser son apprentissage. La division a été effectuée aléatoirement en utilisant la fonction **train_test_split** de sklearn.

```
# Split the data into training and testing sets
features_train, features_test, labels_train, labels_test = train_test_split(features, labels, test_size=0.25, random_state=42)
```

Figure 20: Répartition des données via la fonction train_test_split.

b) Création du modèle LSTM

Notre modèle est composé des éléments suivants (Figure 21):

- **La couche d'entrée**

C'est la première couche du notre système, composé de 100 variables d'entrée qui correspondent à la dimension du plongement lexical (d= 100). Cette couche collecte l'entrée brute et l'utilisent dans le processus de calcul.

- **Les couche LSTM**

La première couche : Contient 64 unités LSTM. Les unités LSTM sont des composants qui permettent de capturer et de modéliser les motifs séquentiels dans les données d'entrée. Plus le nombre d'unités LSTM est élevé, plus le modèle est capable de capturer des motifs complexes et d'apprendre des représentations séquentielles plus riches (Figure 22).

Deuxième couche : Contient 32 unités LSTM. Cette couche LSTM est ajoutée après la première couche LSTM et joue un rôle supplémentaire dans la modélisation des motifs séquentiels. Le choix du nombre d'unités LSTM pour cette couche dépend de la complexité des données et du niveau d'abstraction souhaité dans la modélisation des séquences (Figure 22).

➢ **La Couche de sortie**

Composée de neuf neurones qui correspondent au nombre actuel d'identifiants taxonomiques que notre système est capable de prédire.

✧ **La fonction d'activation**

L'activation **softmax** est une fonction d'activation couramment utilisée dans les tâches de classification multi-classes. Elle est souvent utilisée comme dernière couche d'un modèle de classification pour obtenir des probabilités normalisées pour chaque classe de sortie.

```
num_features = features.shape[1]
features_train = np.reshape(features_train, (features_train.shape[0], 1, num_features))
features_test = np.reshape(features_test, (features_test.shape[0], 1, num_features))


model = Sequential()
model.add(LSTM(64, return_sequences = True, input_shape=(1, num_features)) )
model.add(LSTM(32, return_sequences=False))
model.add(Dense(labels.shape[1], activation='softmax'))
```

Figure 21: Architecture du modèle LSTM proposé.

Lors de la compilation, le modèle vérifiera que les options choisies sont compatibles les unes avec les autres. (Voir figure 22)

```
print(model.summary())
```

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 lstm (LSTM)                 (None, 1, 64)             42240

 lstm_1 (LSTM)               (None, 32)                12416

 dense (Dense)               (None, 9)                 297

=================================================================
Total params: 54,953
Trainable params: 54,953
Non-trainable params: 0
_________________________________________________________________
None
```

Figure 22: Récapitulation du modèle Séquentiel LSTM.

✧ **Algorithme d'optimisation et la fonction d'objectif**

Le modèle est compilé en utilisant la fonction d'activation **'Adam'** pour l'optimiseur. L'optimiseur est un algorithme qui ajuste les poids du modèle afin de minimiser la fonction de perte, également connue sous le nom de fonction objectif. La fonction de perte utilisée est la

'Categorical_crossentropy', qui est couramment utilisée pour les problèmes de classification avec plusieurs classes.

II. Les résultats de l'apprentissage du modèle LSTM

Deux méthodes sont utilisées pour évaluer les résultats de l'apprentissage.

La matrice de confusion : est représentée dans (la figure 23), sous forme d'un tableau 9×9. Le nombre de lignes et de colonnes est en fonction du nombre de classes (neuf classes). Les lignes correspondent aux valeurs réelles d'une classe tandis que les colonnes indiquent les valeurs prédites. La matrice de confusion nous aide à visualiser si le modèle est confus ou bien performant dans la discrimination entre les 9 classes. La matrice de confusion de notre modèle montre qu'il y a une compatibilité presque complète entre les identifiants taxonomiques réelles et les identifiants taxonomiques prédits pas le modèle :

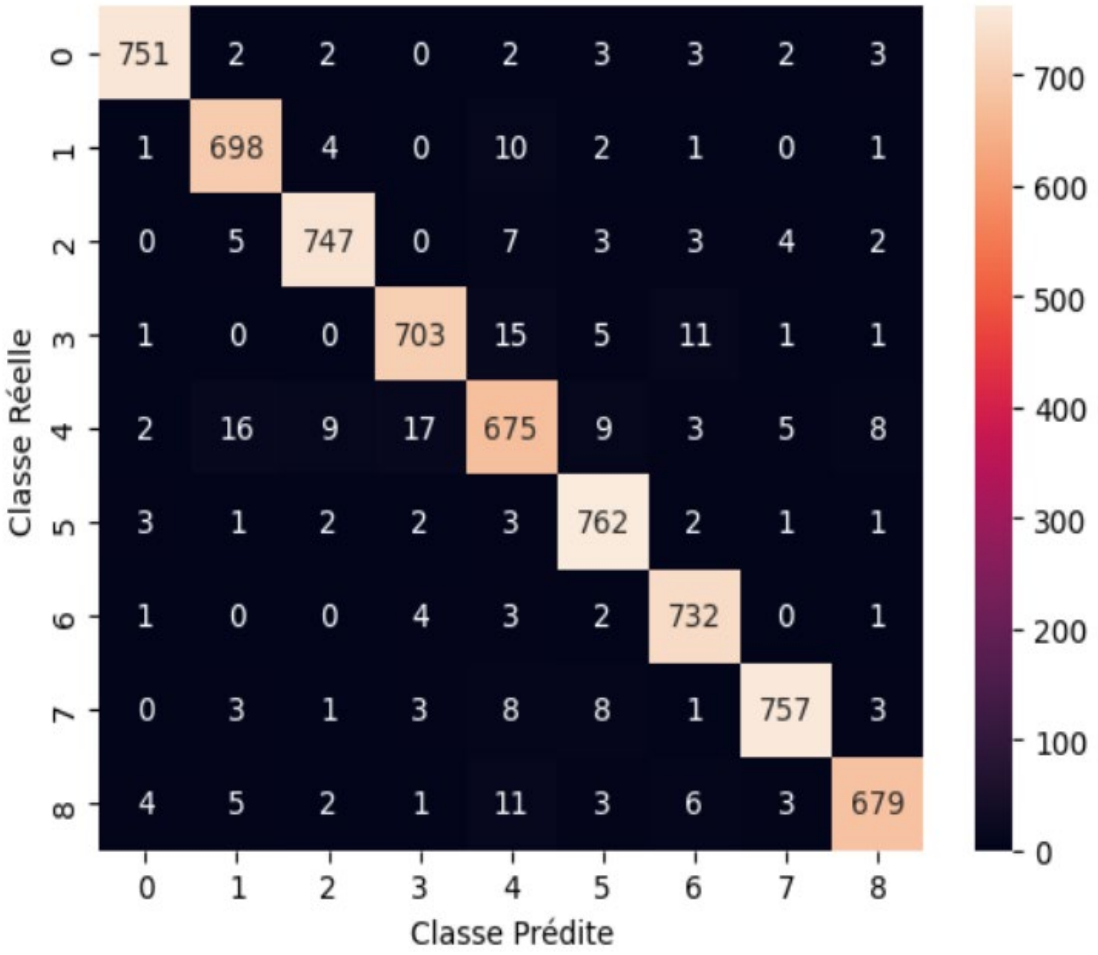

Figure 23: Matrice de confusion du modèle proposé.

Enfin, on trace l'historique de l'apprentissage pour voir son évolution, et pour comparer les valeurs d'apprentissage et de validation. À la fin des 50 époques, nous avons une précision (accuracy) pour l'ensemble d'apprentissage 0.98 et 0.96 pour l'ensemble de validation. La valeur de fonction de perte diminue pour les deux ensembles de données d'apprentissage et de test (figure 24, figure 25).

$$\boldsymbol{Accuracy} = \frac{\mathbf{Numberofcorrectpredictions}}{\mathbf{Totalnumberofpredictions}}$$

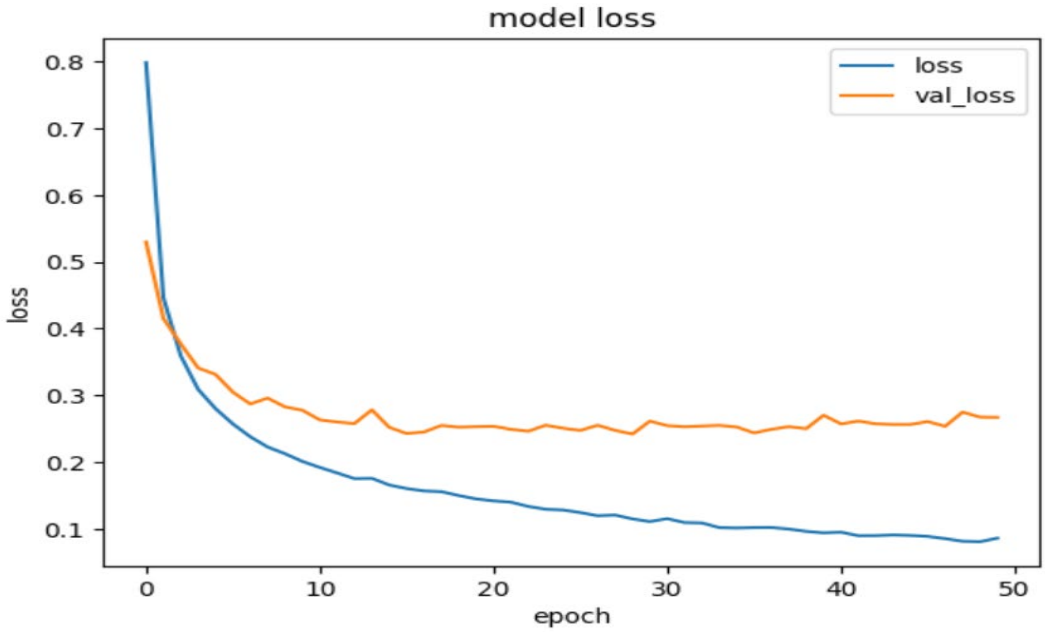

Figure 24: Évolution de la fonction perte pour l'ensemble de données de test et d'apprentissage pour 50 itérations.

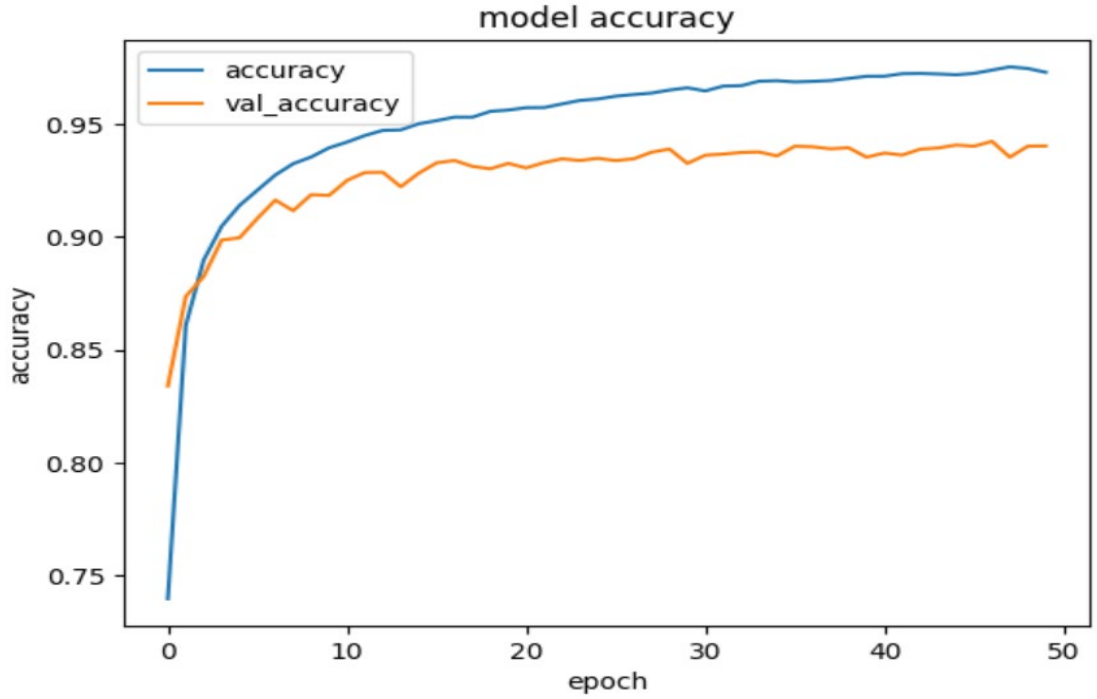

Figure 25: Évolution de la précision pour l'ensemble de données de test et d'apprentissage pour 50 itérations.

2.2.4. Visualisations des données

I. L'arbre phylogénétique :

L'ensemble des identifiants taxonomiques prédits par notre modèle LSTM sont visualisés sous forme d'un arbre phylogénétique (Figure 26). Cet arbre consiste à représenter et montrer le plus petit abre qui relie relie tous les identifiants taxonomiques prédits à l'aide d'une méthode interne qui interroge une base de données NCBI pour faire cette représentation.

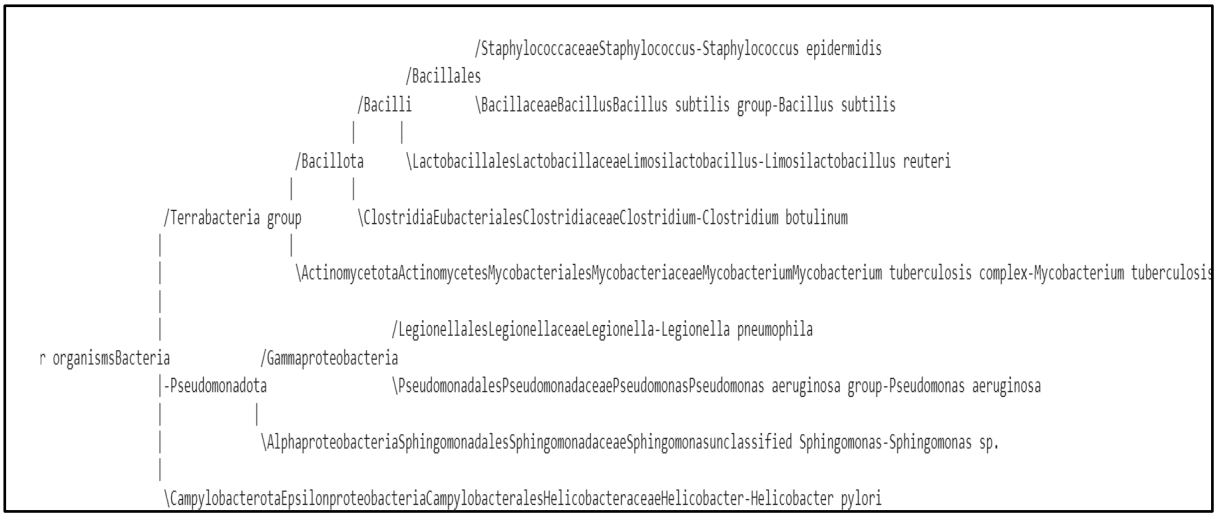

Figure 26: L'arbre phylogénétique.

2.3. Implémentation et expérimentation

2.3.1. Environnement de travail

- **Python**

Python est un langage de programmation open source créé par Guido van Rossum en 1991, inspiré de l'émission Monty Python's Flying Circus. Il s'agit d'un langage interprété, ce qui signifie qu'il n'a pas besoin d'être compilé pour être exécuté. Il offre la possibilité de voir rapidement les résultats des changements apportés au code. Cependant, cela le rend généralement plus lent que les langages compilés tels que le C. Python est apprécié pour sa simplicité et sa lisibilité, ce qui permet aux programmeurs de se concentrer sur la résolution des problèmes plutôt que sur les détails techniques. Cela se traduit souvent par une réduction du

temps nécessaire pour développer des applications par rapport à d'autres langages de programmation [1].

- **Jupyter Notbook**

Jupyter Notbook est un environnement de développement interactif largement utilisé dans le domaine de la science des données et de la programmation. Il permet de créer et d'exécuter des documents appelé « Notebook » qui contient du code, du texte explicatif, des images et des visualisations [48].

2.4. Bibliothèques python

- **Pandas**

Pandas est une bibliothèque conçue pour le langage de programmation Python qui facilite la manipulation et l'analyse des données. Elle offre des structures de données puissantes ainsi que des opérations avancées pour travailler avec des tableaux numériques et des séries temporelles [49].

- **NumPy :**

NumPy est un projet open source qui vise à faciliter les calculs numériques avec Python. Il a été créé en 2005 en s'appuyant sur les travaux initiaux des bibliothèques Numerical et Numarray. NumPy reste entièrement open source et gratuit pour tous les utilisateurs. Son développement est collaboratif et transparent, avec la participation active de la communauté scientifique Python, et il est hébergé sur GitHub pour favoriser la contribution et le consensus au sein de cette communauté [50].

- **biopython**

Est considéré comme le package de bioinformatique le plus vaste et le plus populaire pour Python. Il offre une gamme étendue de modules qui couvrent diverses tâches bioinformatiques courantes. Ces modules permettent notamment d'effectuer des alignements d'ADN, d'ARN.... Biopython est une ressource précieuse pour les chercheurs et les professionnels de la bioinformatique, offrant des fonctionnalités avancées pour l'analyse et la manipulation de données biologiques [51].

- **Matplotib**

Est une bibliothèque Python utilisée pour tracer et visualiser des données sous forme de graphiques. Elle est souvent combinée avec les bibliothèques NumPy et SciPy, qui sont dédiées au calcul scientifique en Python. Matplotlib est distribuée gratuitement en tant que logiciel open source, ce qui signifie qu'elle est libre d'utilisation et accessible à tous les utilisateurs. Cette bibliothèque offre une grande flexibilité et une large gamme de fonctionnalités pour la création de graphiques de qualité professionnelle dans divers formats et styles [52].

- **Sklearn**

Est un module Python qui intègre des algorithmes classiques de machine learning dans l'écosystème des packages scientifiques Python tels que NumPy, SciPy et Matplotlib. Son objectif principal est de fournir des solutions simples et efficaces aux problèmes d'apprentissage automatique, rendant ainsi ces techniques accessibles à tous et réutilisables dans divers contextes. Scikit-learn est largement utilisé comme un outil polyvalent pour la science et l'ingénierie, offrant une grande variété d'algorithmes de classification, de régression, de clustering et de prétraitement des données. Il est également apprécié pour sa documentation complète, ses exemples d'utilisation et sa communauté active qui facilite l'apprentissage et l'application de la machine learning en Python [53].

- **TensorFlow**

Est une plateforme open source qui se concentrent sur la machine learning (ML) et le deep learning (DL). Elle offre un éventail complet d'outils, de bibliothèques et de ressources communautaires flexibles, permettant aux chercheurs d'avancer dans le domaine de l'intelligence artificielle (IA). De plus, elle facilite la création et le déploiement d'applications exploitant cette technologie pour les développeurs [54].

Tableau 4: Caractéristiques des différents outils/bibliothèques informatiques utilisées.

Outils / bibliothèques	**Versions**
Pandas	Pandas1.3.3
Numpy	numpy1.21.2
Biopython	biopython1.79
Matplotlib	matplotlib3.5.1
Sklearn	scikit-learn 0.24.2
Python	Python 3.9.7
Tk	tk 3 .7

2 .5. Présentation de l'interface graphique

L'interface graphique que nous avons développée est conçue pour simplifier l'utilisation du modèle. Nous avons utilisé **Tkinter**, une bibliothèque Python populaire pour le développement d'interfaces graphiques.

Tout d'abord, l'utilisateur doit télécharger des fragments de texte sous forme de fichier fasta au niveau de « Apload file » (Figure 27).

Figure 27 : Téléchargement du fichier.

Après le téléchargement (Upload) du fichier on pourrait également afficher ce fichier (voir figure 28) et verctoriser ce fichier (transformation des séquences en des vecteurs numériques en utilisant le modèle GloVe entraîné) Figure 29.

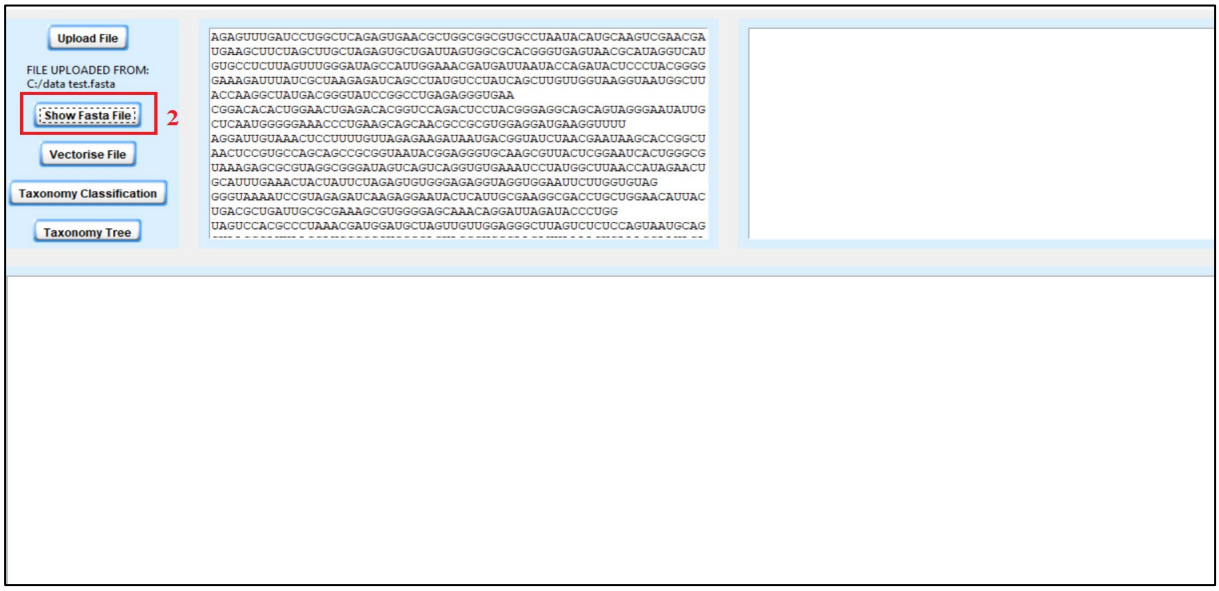

Figure 28: Affichage du fichier Fasta.

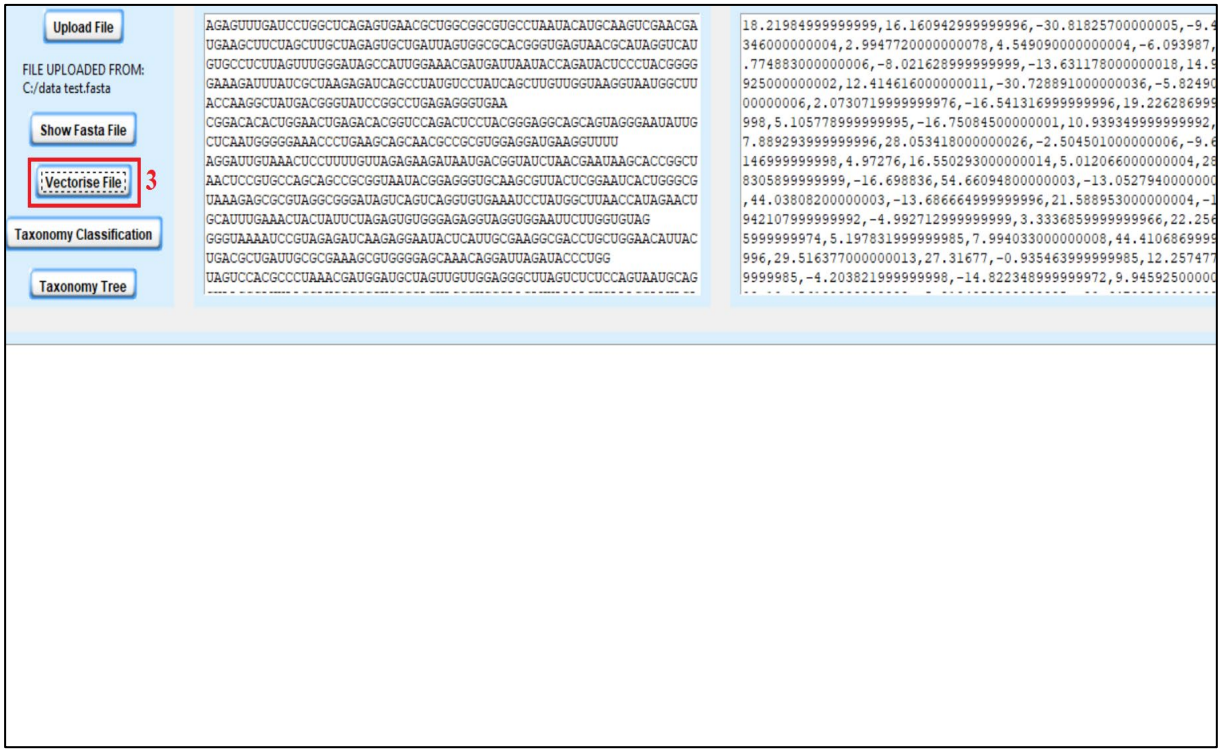

Figure 29: Fichier vectorisé.

L'étape qui suit la vectorisation est la classification taxonomique où le modèle prédit des valeurs et les classe à l'aide de la base de données NCBI. Les résultats obtenus de cette étape sont affichés sous forme de lignes (Figure 30). L'utilisateur peut aussi afficher les résultats de classification sous forme d'un arbre ETE (Figure 31).

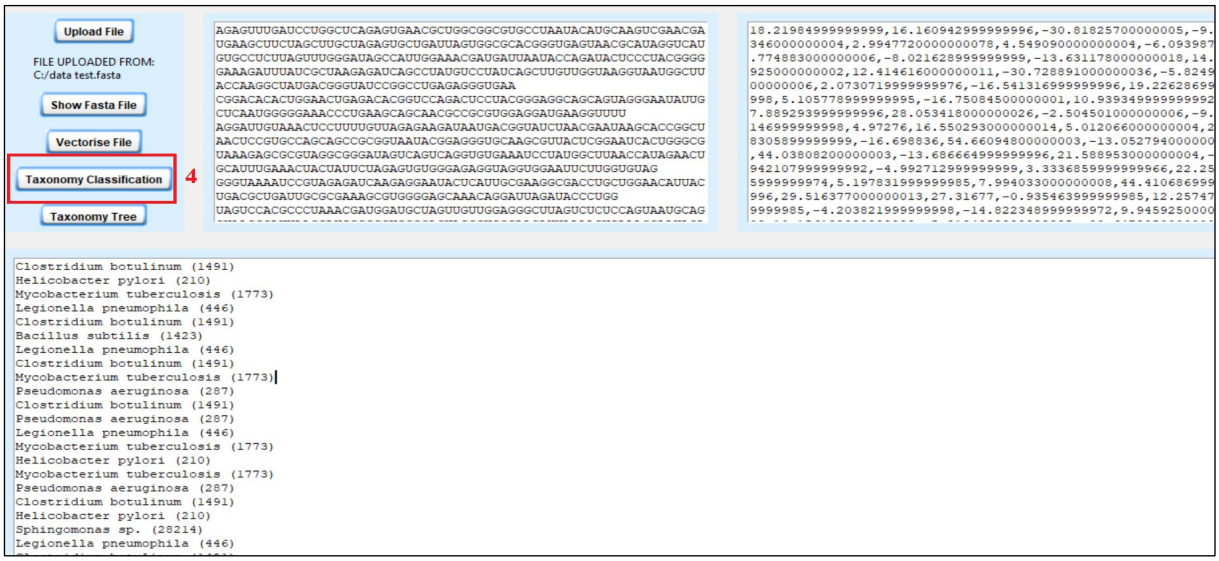

Figure 30: Classification taxonomique.

Figure 31: Affichage de l'arbre phylogénétique.

3. Résultat et discussion

Pour mieux tester le modèle proposé, nous avons utilisé un dataset dont on connaît son appartenance taxonomique mais qui n'a pas été utilisé auparavant dans le processus d'apprentissage. Les résultats obtenus sont affichés dans le tableau

Tableau 5: Résultats de classification taxonomique utilisant le modèle LSTM avec plongements lexicaux GloVe.

Modèle	Précision
LSTM avec plongements lexicaux GloVe	98%

- **Discuter les résultats**

  Les résultats obtenus ont révélé une précision de 98% pour notre modèle, ce qui suggère que notre approche est prometteuse pour prédire l'identifiant taxonomique d'un fragment d'ARNr donné. Cette performance démontre la capacité de notre modèle à extraire et exploiter les caractéristiques distinctives des séquences d'ARNr, permettant ainsi une classification précise.

  Bien que nous n'ayons pas pu comparer notre modèle avec les autres classifieurs taxonomiques et ceci parce que la plupart d'entre eux utilisent des bases de données de plusieurs Giga (≃ Téra) et qui nécessite des ressources matérielles importantes, les résultats obtenus sont encourageants. Ils soulignent le potentiel de notre méthode basée sur les plongements lexicaux et les réseaux LSTM dans le domaine de la classification taxonomique.

  Les résultats suggèrent que notre modèle peut être considéré comme un outil fiable pour révéler la composition taxonomique d'un échantillon taxonomique, en particulier dans des situations où les ressources matérielles sont limitées. Cependant, il serait bénéfique de mener des études futures en utilisant des bases de données plus vastes et en effectuant une comparaison approfondie avec d'autres classifié pour une évaluation plus complète de l'efficacité de notre modèle dans un contexte plus large.

4. Conclusion

Dans ce travail, nous avons présenté une nouvelle méthode basée sur la composition des séquences pour la classification taxonomique fondée sur les principes des plongements lexicaux GloVe et une architecture d'apprentissage profond de type LSTM. Notre approche se compose de deux étapes. La première étape vise à obtenir une représentation vectorielle numérique (plongements lexicaux) de fragments d'ARN à l'aide d'un modèle TALN. Ensuite, ces plongements sont utilisés pour créer un classifier. Ce dernier pourra prédire un identifiant taxonomique pour un fragment d'ARN donné.

Nous avons évalué le modèle proposé avec un dataset externe qui n'a pas été utilisé lors de l'apprentissage. [98%] de précision est obtenu ce qui nous ramène à penser que cette méthode peut être utilisée dans une approche globale afin de présenter un outil capable de révéler la composition taxonomique d'un échantillon taxonomique.

BIBLIOGRAPHIE

[1] Aliouane, S. E., Bendahmane, A. (2020). "Nouvelle approche de prédiction des classes protéiques issues d'un séquençage NGS par deep learning." Mémoire de Master en bioinformatique, Université Frères Mentouri Constantine 1, 17 septembre 2020.

[2] Diene, S. M., Bertelli, C., et al. (2014). Génomique et métagénomique bactérienne : Applications cliniques et importance médicale. Revue médicale suisse, 10, 2155-2161.

[5] Caboche, S. (05 et 06 décembre 2018). Cycle de formation NGS Module 5 : Métagénomique Partie 1 : Métagénomique ciblée [Diapositive].

[6] Turnbaugh, P. J., Ley et al. (2007). The Human Microbiome Project. Nature, 449(7164), 804-810.

[8] Gilbert, J. A., Jansson, J. K., et al. (2014). The Earth Microbiome Project: Successes and aspirations. BMC Biology, 12: 69. Doi: 10.1186/s12915-014-0069-1

[9] Greve, C. (2015). Développement et applications d'outils d'analyse métagénomique des communautés microbiennes associées aux insectes (Doctoral dissertation, Université Claude Bernard Lyon 1).

[10] Lamoril, J., Ameziane, N., et al. (2008). Les techniques de séquençage de l'ADN : Une révolution en marche. Première partie [DNA sequencing technologies: A revolution in motion. Part one]. Immuno-analyse & Biologie Spécialisée, 23(5), 260-279.

[11] Bichat, A. (2020). Prise en compte de l'organisation hiérarchique des espèces pour la découverte de signatures métagénomiques multi-échelles. Thèse de doctorat de l'Université Paris-Saclay, École Doctorale de Mathématique Hadamard (EDMH) n°574, Spécialité de doctorat : Mathématiques appliquées, Unité de recherche : Laboratoire de mathématiques et modélisation d'Évry (UEVE), UMR 8071 CNRS-INRA. Référent : Université d'Évry. Thèse présentée et soutenue à Paris.

[14] Siegwald, L. (2017). Solutions d'amélioration des études de métagénomique ciblée (Université de Lille - École doctorale Biologie et Santé).

[15] Xiong, W., Giannone, R. J., et al. (2011). Metagenomic functional and taxonomic profiling for complex microbial community analysis. Nature Methods, 8(10), 956-959. Doi: 10.1038/nmeth.1702.

[16] Wikström, M. (2020). Taxonomic classification of metagenomic short reads. Department of Mathematical Sciences, Chalmers University of Technology, Gothenburg, Sweden, p. 5.

[17] Chentli, A. (2021). Taxonomie bactérienne [Diapositive 6].

[19] Alauzet, C. (2009). Taxonomie des bactéries anaérobies : de la reclassification à la découverte de nouveaux pathogènes. Thèse de doctorat, Université Henri Poincaré, Mention : Génomique.

[20] Anwar, A., Ebersold, S., et al. (2009). Vers une approche à base de règles pour la composition de modèles. Application au profil VUML. Laboratoire IRIT, Université de Toulouse II le Mirail, Toulouse, France. Laboratoire SI2M, ENSIAS, Rabat, Maroc. Laboratoire LRI-MIARF, Université Mohammed V-Agdal, Faculté des sciences, Rabat, Maroc.

[21] Hassabis, D., Kumaran, D., et al. (2017). Neuroscience-inspired artificial intelligence. Neuron, 95(2), 245-258.

[22] Annina, S., Mhimah, S., et al. (2016). "An overview of machine learning and its applications," International Journal of Electrical Sciences and Engineering (IJESE).

[24] Sutton, R. S., Barto, A. G. (2018). Reinforcement Learning: An Introduction. The MIT Press.

[27] Patterson, J., Gibson, A. (2017). Deep Learning: A Practitioner's Approach, 1st ed. O'Reilly Media, Inc.

[29] LeCun, Y., Bengio, Y.,et Hinton, G. (2015). Deep Learning. Nature, 521(7553), 436-444.

[30] Bouzelifa, R. ET Rouimel, S. (2019). Mémoire de fin d'études pour l'obtention du diplôme de Master en Électronique, Option Électronique des Systèmes Embarqués : Reconnaissance des images avec les réseaux de neurones artificiels. Faculté des Sciences et de la Technologie, Département d'Électronique, Université Mohammed Seddik Ben Yahia Jijel.

[31] D'Acremont, A. (2020). Réseaux de neurones profonds pour la classification d'objets en imagerie infrarouge : apports de l'apprentissage à partir de données synthétiques et de la détection d'anomalies (Thèse de doctorat). École doctorale no 601, Mathématiques et Sciences et

Technologies de l'Information et de la Communication, Spécialité : Signal, Image, Vision, ENSTA Bretagne.

[32] Bouaziz, M. (s. d.).(2017). Réseaux de neurones récurrents pour la classification de séquences dans des flux audiovisuels parallèles (Thèse de doctorat). Université d'Avignon et

des Pays de Vaucluse, École doctorale 536 "Sciences et Agrosciences", Laboratoire d'Informatique (EA 4128), Laboratoire d'Informatique d'Avignon.

[33] Young, T., Hazarika, et al. (2018). Recent Trends in Deep Learning Based Natural Language Processing. IEEE Computational Intelligence Magazine, 13(3), 55-75.

[34] Deng, L., Yu, D. (2014). Deep Learning: Methods and Applications. Foundations and Trends in Signal Processing, 7(3-4), 197-387.

[35] Jurafsky, D., Martin, J. H. (2020). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (3rd ed.). Pearson.

[36] Badaoui, M. Y. Z. (2020). Génération in-silico des molécules à visées thérapeutiques basée sur les méthodes d'intelligence computationnelle. Mémoire de master, Université Frère Mentouri – Constantine 1, Faculté des Sciences de la Nature et de la Vie, Département de Biochimie et de Biologie Moléculaire et Cellulaire.

[37] Young, T., Hazarika, D., et al. (2018). Recent Trends in Deep Learning Based Natural Language Processing. IEEE Computational Intelligence Magazine, 13(3), 55-75.

[38] Belainine, B. (2017). Classification supervisée de textes courts et bruités : application au domaine des médias sociaux. Mémoire présenté comme exigence partielle de la maîtrise en informatique, Université du Québec à Montréal.

[39] Boullier, D., Lohard, A. (2012). Opinion Mining et Sentiment Analysis: Méthodes et Outils. Paris: Open Editions Press.

[42] Ghannay, S. (2018). Étude sur les représentations continues de mots appliquées à la détection automatique des erreurs de reconnaissance de la parole.

[44] Lau, J. H., Baldwin, T. (2016), "An Empirical Evaluation of Doc2Vec with Practical Insights into Document Embedding Generation," Proceedings of the 1st Workshop on Representation Learning for NLP, Berlin, Germany, pp. 78-86, arXiv:1607.05368 [cs.CL].

[45] Pennington, J., Socher, R.,et Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. Computer Science Department, Stanford University, Stanford, CA 94305.

[46] Athiwaratkun, B., Wilson, A. G.,et Anandkumar, A. (2018). Probabilistic FastText for Multi-Sense Word Embeddings. ArXiv: 1806.02901 [cs.CL].

[47] Alauzet, C. (2009). Taxonomie des bactéries anaérobies: de la reclassification à la découverte de nouveaux pathogènes. Thèse de doctorat, Université Henri Poincaré, Mention: Génomique

Webographie :

[3] URL **:** "Genomique et metagenomique bacteriennes : applications cliniques et importance medicale." Consulté le 11 avril 2023. Disponible sur : https://www.revmed.ch/revue-medicale-suisse/2014/revue-medicale-suisse-450/genomique-et-metagenomique-bacteriennes-applications-cliniques-et-importance-medicale.

[4] URL: Metagenomics. Consulter le 13 avril 2023. https://youtu.be/6QYF4lCAIRm.

[7] URL: Étude d'un exemple. Consulter le 15 avril 2023 : L'expédition Tara. https://svtlyceedevienne.com/term-ens-scientifique/theme-3-une-histoire-du-vivant/le-recensement-de-la-biodiversite-dun-ecosysteme/etude-dun-exemple-lexpedition-tara/.

[12] URL: MinION. Consulter le 29 avril 2023. https://site.lookatsciences.com/minion-un-sequenceur-portatif-de-la-taille-dune-clef-usb/.

[13] URL: Introduction à la métagénomique. Consulter le 29 avril 2023 ; https://dridk.me/metagenomique.html.

[18] URL: Tree of Life. Consulter le 29 avril 2023. https://fr.wikipedia.org/wiki/Fichier:Rangs_taxonomiques.svg.

[23] URL: Apprentissage supervisé et non supervisé. Consulter le 1 mai 2023. https://www.lemagit.fr/conseil/Machine-Learning-les-9-types-dalgorithmes-les-plus-pertinents-en-entreprise.

[25] URL : Introduction à l'apprentissage profond (deep learning) de l'intelligence artificielle. Consulter le 1 mai 2023. https://culturesciencesphysique.ens-lyon.fr/ressource/IA-apprentissage-Rousseau.xml.

[26] URL: Une représentation de l'organisation des différentes disciplines présentées. Consulter le 3 mai 2023. https://openclassrooms.com/fr/courses/6417031-objectif-ia-initiez-vous-a-lintelligence-artificielle/6822141-reperez-vous-dans-le-champ-de-lintelligence-artificielle.

[28] URL: Le perceptron multicouches. Consulter le 3 mai 2023. https://www.researchgate.net/figure/Le-perceptron-multicouches_fig6_30517821.

[40] URL : Opinion mining et Sentiment analysis consulter le 3 mai 2023 https://books.openedition.org/oep/198?lang=fr.

[41] URL: Relation entre DL, ML et NLP. Consulter le 3 mai 2023. https://www.researchgate.net/figure/Relationship-between-AI-ML-DL-and-NLP-7_fig8_343079524.

[43] URL : Word Embeddings: CBOW vs Skip-Gram consulté le 3 mai 2023 https://www.baeldung.com/cs/word-embeddings-cbow-vs-skip-gram.

[48] URL: Jupyter Notebook. Consulter le 8 juin 2023. https://jupyter.org/.

[49] URL: Pandas 2023. Consulter le 8 juin 2023. https://pandas.pydata.org/.

[50] URL: NumPy. Consulter le 8 juin 2023. https://numpy.org/.

[51] URL: Biopython 2023. Consulter le 9 juin 2023. https://biopython.org/.

[52] URL: Matplotlib 2023. Consulter le 10 juin 2023. https://matplotlib.org/.

[53] URL: Scikit-learn 2023. Consulter le 10 juin 2023. https://scikit-learn.org/.

[54] URL: TensorFlow 2023. Consulter le 10 juin 2023. https://www.tensorflow.org/.

Année universitaire : 2022-2023	Présenté par : GUECHTAL Loubna et OUELBANI Rania Nour et TALBI Yasmina

Une approche basée sur le traitement automatique du langage naturel (TALN) pour la classification taxonomique des séquences métagénomiques 16S rRNA.

Mémoire pour l'obtention du diplôme de Master en Bioinformatique.

Cette étude vise à développer une approche basée sur les réseaux LSTM (Long Short-Term Memory) pour simuler les différentes étapes du processus de la métagénomique, en se concentrant spécifiquement sur l'analyse des données générées par les technologies de séquençage de nouvelle génération (NGS). Cette approche repose sur deux axes clés de l'intelligence artificielle, à savoir le traitement automatique du langage naturel (NLP) et l'apprentissage profond (DL). En utilisant un ensemble de données d'apprentissage composé de neuf bactéries, des tests ont été effectués et ont abouti à un taux de précision de 98%. Ces résultats démontrent l'efficacité de l'approche, notamment en ce qui concerne la phase de prédiction basée sur le TALAN et le DL. La combinaison de ces deux outils a permis de développer un modèle possédant une grande capacité d'extraction de connaissances à partir des données génomiques, permettant ainsi la prédiction et la classification taxonomique des génomes. Ce modèle a été entraîné de manière approfondie en exploitant les séquences génomiques. Les résultats de cette recherche mettent en évidence l'apport significatif de cette approche pour améliorer la précision de la classification des génomes.

Mot clés : Apprentissage ; prédiction ; Intelligence Artificielle ; TALAN ; NGS ; LSTM ;

Centre de recherche : Centre de recherche en biotechnologie de Constantine

Président : Dr. KELLOU Kamel	(Université Frères Mentouri, Constantine 1).
Encadreur : Dr. MATOUGUI Brahim	(Centre de recherches biotechnologie, Constantine).
Co-Encadreur : Dr. GHERBOUDJ Amira	(Université Frères Mentouri, Constantine 1).
Examinateur 1 : Dr. CHEHILI Hamza	(Université Frères Mentouri, Constantine 1).