ا لجمهورية الجزائرية الديمقراطية الشعبية
RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE

وزارة التعليم العالي و البحث العلمي
MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE SCIENTIFIQUE

Université Frères Mentouri - Constantine 1
Faculté des Sciences de la Nature et de la Vie

جامعة الإخوة منتوري - قسنطينة 1
كلية عاوم الطبيعة والحياة

Département : Biologie Appliquée. **قسم : بيولوجيا تطبيقية**

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

Domaine : Sciences de la Nature et de la Vie

Filière : Sciences Biologiques

Spécialité : *BIOINFORMATIQUE*

Intitulé :

Prédiction des interactions ARN-Protéines basée sur le Deep-Learning

---

Présenté et soutenu par : HAMLAOUI Maria Ouissal **Le : 23/09/2021**

KERRICHE Yousra

Jury d'évaluation :

Président :	**Dr. TEMAGHOULT Mahmoud**	(Université Frères Mentouri - Constantine 1)
Rapporteur :	**Dr. CHEHILI Hamza**	(Université Frères Mentouri - Constantine 1)
Examinateur :	**Dr. KELLOU Kamel**	(Université Frères Mentouri - Constantine 1)

Année universitaire
2020- 2021

Remerciement

Tout d'abord nous remercions **Dieu** le tout puisant qui nous a gardé en bonne santé et nous a donné la force et la patience afin de réaliser ce travail.

Nous tenons à remercier chaleureusement notre encadreur Dr **Chehili Hamza** pour le support qu'il a investi dans la supervision de notre mémoire.

Il a été pour nous un excellent guide au cours de ce périple académique.

Nous sommes très reconnaissants pour la confidence que vous nous avez accordée et pour avoir cru en nos capacités tout au long de ce travail.

Nous vous remercions pour votre patience, ainsi que pour la gentillesse que vous avez manifestée a notre égard.

Nous gardons toujours beaucoup de plaisir à discuter avec vous et bénéficier de votre expérience.

Nos sincères remerciements à nos enseignants de la faculté des sciences de la nature et de la vie pendant les cinq années précédentes.

Nous voudrions également accorder une place spéciale pour « **Mr.Boucheloukh islam** » qui a été un support moral indispensable pendant tout ce temps et qui nous a aidé pour réaliser ce mémoire.

Nous exprimons nos vifs remerciements aux membres de jury :

Mr.Temaghoult Mahmoud (Faculté des sciences de la nature et de la vie UFM Constantine)

Mr.Kellou Kamel (Faculté des sciences de la nature et de la vie UFM Constantine)

A Toutes et à tous qui de loin ou de prés ont contribué à la réalisation et a la finalisation de ce travail sur tout «**Mr.Hamlaoui Salah-Eddine**» . Qu'ils trouvent ici l'expression de nos sincères remerciements et notre profonde reconnaissance.

Merci .....

Dédicace

Je tiens à dédier ce mémoire à mes très chers parents
Rachid et Yasmina, et aussi ma deuxième maman Nadia en témoignage et en gratitude de leurs dévouements, leurs sacrifices illimités, leurs
réconfort moral, leur encouragement non seulement au
cours de mon cycle d'étude; mais aussi tout au long de mon parcours. Eux qui ont consenti beaucoup d'effort pour mon éducation, mon instruction. Et sans lesquels je n'en serais pas la aujourd'hui, vous m'avez donné toutes les chances pour réussir. C'est grâce à vous que je suis ce que je suis maintenant. Pour tout cela et pour ce qui ne peut être dit, mes affections sans limite.

A ma sœur Rayene et mes frères Seif Eddine, Salah Eddine, et Jad; pour leurs aides précieuses et d'être toujours à mes côtés durant ces années, pour tous les moments de joies que nous avons vécues ensemble.

A mon fiancé Moatez Billeh Je te demande de bien vouloir trouver ici ma reconnaissance pour ta patience et ton encouragement qui n'a pas d'équivalent, et ton grand soutien pour réaliser ce travail.

A mes belles sœurs Amina, Warda pour ses profonds sentiments de fraternité et de soutien.

A mon neveu petit ange Abderrahmane et mes nièces Dina, Ayli, Celia d'être mes petits rayons de soleil que j'adore.

A mon binôme Yousra, qui m'a fait confiance.

Cette dédicace ne peut s'achever sans une pensée à mon encadreur Dr Chehili, je suis ravie d'avoir travaillé en votre compagnie tout on long de ce mémoire, j'ai apprécié vos qualités

Maria Ouissal

Dédicace

Je profite à cette honorable occasion pour dédier ce mémoire,

A mon père Abdelghani,
qui m'a toujours poussé et motivé
durant mes études.
A ma chère maman Ferdi Samia, pour ses sacrifices,
son amour, sa tendresse et ses prières .
A mes chers frères que j'aime beaucoup abd elhakim, Midou et ma chère sœur
Maoua pour leurs encouragements, aides et supports
pendant les moments difficiles.
A mon fiancé Souilah Choukri pour sa compréhension qui n'a pas d'équivalent
A toute ma famille, mes cousins,
et mes amis.
surtout ma camarade Maria , pour sa gentillesse.
Sans oublier mon encadreur Mr.Chehili pour son soutien moral, sa patience et sa compréhension tout au long de ce projet.

Que ce travail soit l'accomplissement de vos vœux, et le fruit de votre soutien infaillible.

Merci d'être toujours là pour moi.....

Yousra

Sommaire :

Chapitre 3 : Processus de la prédiction des interactions ARN-Protéines basé sur DL

Chapitre 4 : Implémentation et discussion

Liste des abréviations :

ADN : Acide Désoxyribonucléique.

ARN : Acide ribonucléique.

BM – Biomol : Biologie moléculaire.

CNN : Convolutional Neural Network (Réseau neuronal convolutif).

CSV: Comma-Separated Values.

DL: Deep Learning (Apprentissage profond).

FC : Fully-Connected.

IA : Intelligence artificielle, (Artificial Intelligence).

ML : Machine Learning (Apprentissage Automatique).

MLP : Multi-Layer Perceptron.

NLP : Natural Language Processing (Traitement automatique du langage naturel).

RBP : La rétinol-binding protéine.

ReLU : Unités Rectifié Linéaire.

Tf-IdF: Term Frequency-Inverse Document Frequency

UCLA : Université de Californie à Los Angeles.

UICPA : Union Internationale de Chimie Pure et Appliquée.

Liste des figures :

Introduction générale

Pendant de nombreuses années, les ARN ont été considérés comme de simples intermédiaires entre l'ADN qui contient les gènes (l'information génétique) et les protéines. On sait maintenant que les ARN possèdent de nombreuses autres fonctions et interviennent à différents niveaux de la régulation de l'expression génique. Ces fonctions dépendent de l'établissement d'interactions ARN-composés.

Les interactions ARN-composés interviennent dans de nombreux processus cellulaires fondamentaux et sont essentielles dans tous les règnes du vivant. Elles font intervenir une variété de domaines de liaison à l'ARN dont la plupart reconnaît et lie spécifiquement une courte séquence nucléotidique accessible sous forme simple-brin.

En parallèle et depuis quelques années, l'Intelligence Artificielle (IA) connaît un regain d'intérêt sans précédent grâce à d'importantes avancées technologiques. Notamment dans le domaine de l'apprentissage machine (machine Learning), qui étendent les capacités des ordinateurs et accroissent leurs performances dans un grand nombre de domaines, grâce à de nouveaux algorithmes ainsi qu'à la multiplication des jeux de données et le décuplement des puissances de calcul, ce domaine donne lieu à de nombreux espoirs dans différents domaines.

Les méthodes de l'intelligence artificielles telles que l'apprentissage automatique ou l'apprentissage profond (Deep Learning) ont apporté de nouvelles solutions à différents problèmes de la biologie.

Le Deep Learning est une méthode développée dans le domaine de l'apprentissage automatique, il a été utilisé dans une variété de domaines ces dernières années car il a atteint une précision de prédiction élevée dans les domaines comme la reconnaissance d'image, la reconnaissance vocale et la prédiction d'activité composée.

Les méthodes basées sur l'apprentissage automatique pour prédire les interactions ARN-composé sont largement divisé entre ceux basés sur des données de structure moléculaire et ceux basés sur des données de réseau.

Des méthodes de prédiction d'interaction ARN-composé basées sur l'apprentissage profond ont été sur les données de structure moléculaire. Cependant, ces méthodes n'utilisent que des informations basées sur les séquences nucléotidiques et les structures chimiques, de sorte que les propriétés fonctionnelles des ARN et des composés n'ont pas encore été incorporées dans la prédiction.

Dans les chapitres qui suivent, après une introduction, nous allons prendre en compte l'étude moléculaire des Acides ribonucléique, ARN viral, et les molécules en précisant l'arrangement structural de ces molécules et cela dans le premier chapitre.

Dans le deuxième chapitre les concepts généraux de l'apprentissage automatique, suivie par l'apprentissage profond et le traitement automatique du langage naturel sera souligné.

Afin de représenter la partie pratique de ce travail dans le troisième chapitre et de consiste à implémenter l'approche utilisée. Dans le dernier chapitre nous discuterons l'approche proposé et comment atteindre l'objectif de prédire la protéine inhibitrice d'ARN viral.

Chapitre 1 :

Biologie moléculaire

1. Introduction

La biologie moléculaire « abrégée bio-mol ou BM » a été utilisée la première fois en 1938 par Warren Weaver, désigne également l'ensemble des techniques de manipulation d'acides nucléiques (ADN-ARN) [1].

Depuis la fin des années 1950 et le début des années 1960, les biologistes moléculaires ont appris à caractériser, isoler et manipuler les composants moléculaires des cellules et des organismes. Ces composants incluent l'ADN, support de l'information génétique, l'ARN, et les protéines, molécules structurelles et enzymatiques les plus importantes des cellules [2].

L'information contenue dans les gènes va servir à la fabrication de milliers de protéines qui interviennent dans le fonctionnement de la cellule. La première étape de l'expression d'un gène consiste à recopier son information sous la forme d'une molécule très proche de l'ADN, l'acide ribonucléique ou ARN [3].

La bio-mol a connu d'importants développements pour devenir un outil incontournable de la biologie moderne à partir des années 1970 [2].

2. Biologie moléculaire

La bio-mol est une discipline scientifique au croisement de la génétique, de la biochimie et de la physique, dont l'objet est la compréhension des mécanismes de fonctionnement de la cellule au niveau moléculaire [4].

2.1. ADN

Acide Désoxyribonucléique « abrégée ADN » est un acide nucléique support de l'information génétique et de sa transmission au cours des générations « hérédité ».

L'ADN est le principal constituant des chromosomes, il est la plus grosse molécule du monde vivant et elle est présente dans tous les organismes vivants.

2.1.1. Structure de l'ADN

Une molécule d'ADN est une double hélice composée de deux brins enroulés l'un autour de l'autre, on dit que l'ADN est bi-caténaire chacun de ces brins est constitué d'un enchaînement de bases dites puriques « Guanine G, Adénine A » et pyrimidiques « Cytosine C, thymine T » [5].

Les bases sont reliées entre elles à l'intérieur d'un brin d'ADN par des sucres des oses, appelés désoxyriboses, et par des acides phosphoriques, Une base plus un sucre et un phosphate constituent un nucléotide. L'enchaînement des nucléotides forme un brin d'ADN [5].

L'appariement des deux brins qui composent l'hélice d'ADN est réalisé par les bases : l'adénine peut en effet, se lier par des liaisons faibles à la thymine « AT » et la guanine fait de même avec la cytosine « GC ». En aucun cas, thymine et guanine, ou cytosine et adénine ne peuvent s'apparier [5].

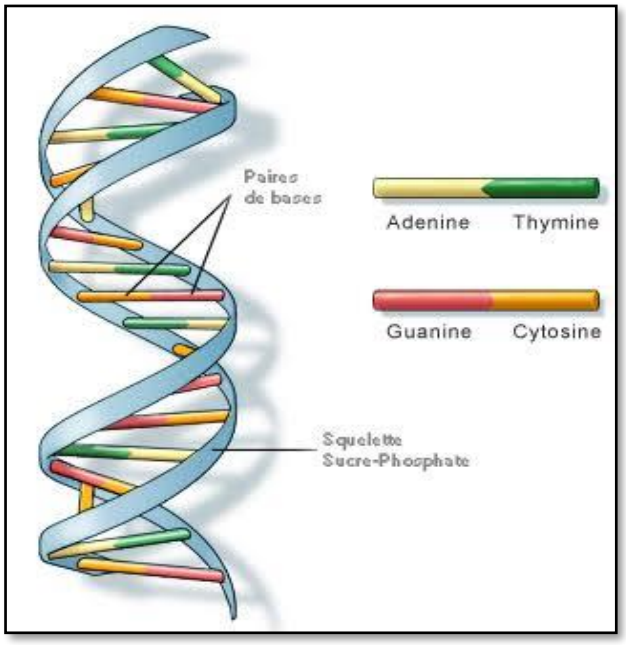

Figure 1 : Structure d'ADN

2.1.2. Fonction de l'ADN

L'ADN a deux rôles fondamentaux : il est le support de l'information génétique, il permet la transmission des informations génétiques de cellule en cellule et de génération en génération.

2.2. ARN

Acide ribonucléique « abrégée ARN » est un acide nucléique formé par une chaîne de ribonucléotides, Composé de ribose, de phosphate, d'adénine, de cytosine, de guanine et d'uracile. Il est présent dans les cellules procaryotes et les eucaryotes, cet acide nucléique résulte de la transcription de l'ADN.

L'ARN est constitué d'une chaîne de monomères appelés nucléotides. Les nucléotides sont joints les uns après les autres par des liaisons phosphodiester chargées négativement.

Les ARN messagers ont copié, au niveau de chaque gène, une séquence d'ADN. Chaque molécule d'ARN messager se fixera sur un ribosome. Le ribosome va glisser le long du messager en donnant naissance à la chaîne protéique [6].

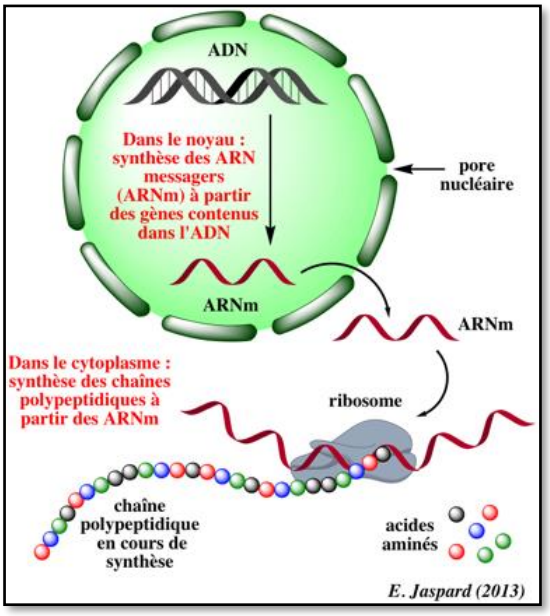

Figure 2 : Transcription et traduction « La synthèse des protéines, université Angers »

2.2.1. Structure de l'ARN

Les ribonucléotides sont différents des désoxynucléotides par la présence d'un groupement OH en 2' du ribose, mais aussi par le fait que la thymine (T) est remplacée par l'uracile (U) [4].

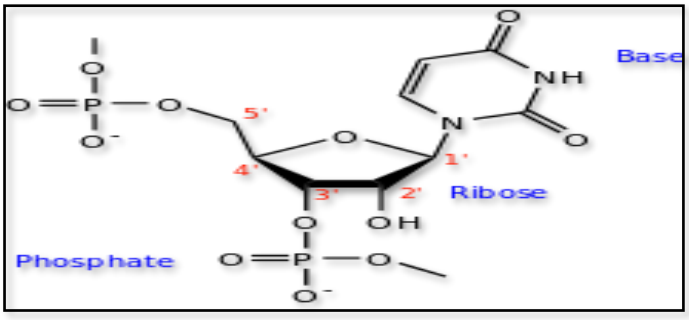

Figure 3 : Nucléotide

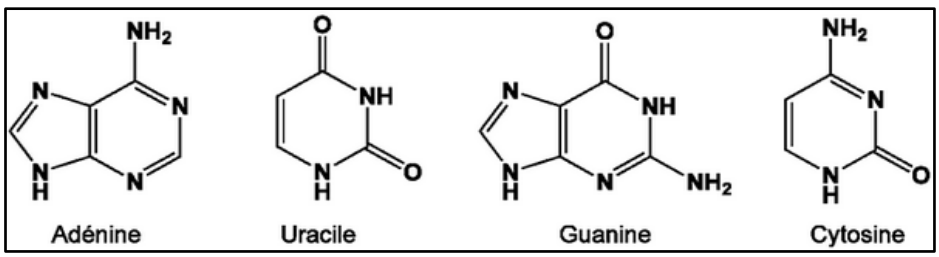

Figure 4 : Bases azotées d'ARN

2.2.2. Classification des ARN

La classification des ARN se fait en deux (2) classes :

a) ARN Codant

Dans le cas des gènes codants, l'ARN messager issu de la transcription, puis de la maturation éventuelle, contient toute l'information nécessaire à la production d'une protéine :

- **ARN messagers (ARNm)**

Est un support temporaire de l'information génétique. Il est utilisé par la cellule pour transmettre l'information correspondant à un gène donné, a séquence nucléotidique de l'ARNm est une séquence linéaire complémentaire et antiparallèle à la séquence matrice de l'ADN dont elle est issue « ARNm représente 5% des ARN totaux » [2].

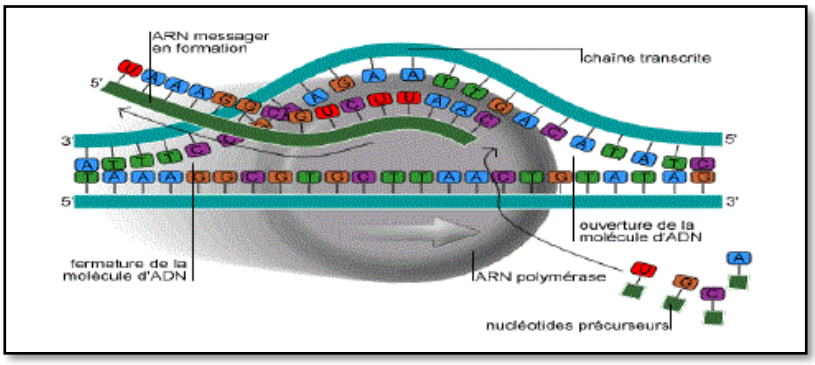

Figure 5 : Formation d'ARN à partir de la double hélice d'ADN « Thème 1 - Transmission, variation et expression du patrimoine génétique, QCM SVT »

La maturation des ARNm

Cette étape est spécifique du monde des eucaryotes, son rôle principal est de protégé les ARNm pendant leur transport du noyau jusqu'au cytoplasme ou ils vont être traduit en fonction du code génétique, elle fait intervenir trois évènements :

Formation de la coiffe en 5' des ARNm dans le noyau dès le début de la transcription, son rôle est de protéger la région 5' des ARN pré messager vis-à-vis des nucléases et des phosphatases.

La polyadénylation est un signal permettant de moduler la stabilité des ARN, elle consiste en l'addition d'une queue poly (A) « une succession de nombreux ribonucléotides de type Adénosine (A) » à l'extrémité 3' des ARNm, Chez les eucaryotes cette étape s'effectue dans le noyau.

L'étape finale de la maturation assure l'élimination des introns lors de l'excision suivie de l'épissage des exons. Ainsi cette maturation aboutit à des ARNm plus courts puisque débarrassés des séquences non codantes. Ces étapes sont catalysées par diverses enzymes nucléaires.

b) ARN Non-Codant

A l'inverse des ARN messagers issus de la transcription de gènes codants, les ARN non codants issus de la transcription de gènes à ARN n'ont pas vocation à coder pour des protéines. Les ARN non-codants assurent diverses fonctions pour la plupart déterminées par la structure spatiale qu'adopte la molécule d'ARN en se repliant sur elle-même. Les ARN non-codants qui ne se replient pas de manière spécifique s'associent le plus souvent à d'autres molécules telles que des protéines pour former des complexes. Les fonctions de ces ARN « non-structures » sont alors en lien étroit avec leur séquence [5], Dans la suite nous nous intéressons essentiellement aux ARN non-codants qui adoptent une structure caractéristique :

- **ARN ribosomiques (ARNr)**

Les ARN ribosomiaux représentent plus de 80% des ARN cellulaires totaux s'associent à des protéines pour former le ribosome qui est le support de la synthèse des protéines. Les ribosomes sont une association de 2 sous unités : 50S et 30S chez les procaryotes et 60S et 40S chez les eucaryotes [9].

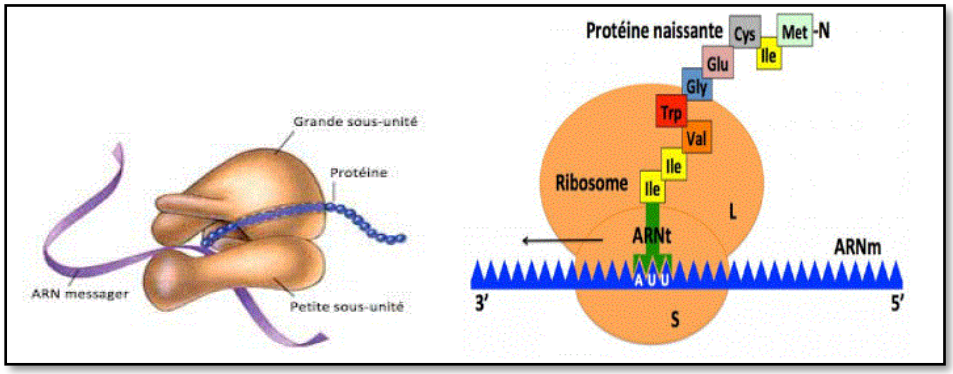

Figure 6 : ARN ribosomiques

- **ARN de transfert (ARNt)**

Un ARNt est une courte molécule d'ARN, de structure complexe qui joue un rôle fondamental dans la synthèse des protéines. Si on aplatit artificiellement la molécule d'ARNt, elle apparaît en forme de croix avec trois boucles et quatre régions double brin. (ARNt représente 15% des ARN totaux) [8].

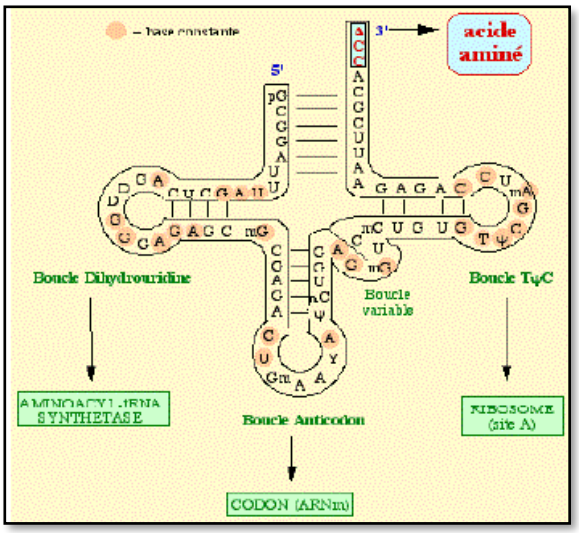

Figure 7 : ARN de transfert « La synthèse des protéines, université angers »

- **ARN de régulation**

Un certain nombre de types d'ARN sont impliqués dans la régulation de l'expression des gènes, y compris l'ARN micro (miARN), le petit ARN d'intervention (siARN) et l'ARN antisens (aARN) [9].

- **D'autres familles**

Ils existent des nouvelles classes (ARNsno, ARNnc...) sont régulièrement découvertes.

Chez certains virus (le virus de la mosaïque du tabac, le VIH...), l'ARN constitue le génome (alors que chez la grande majorité des organismes, c'est l'ADN qui remplit cette fonction).

2.3. Gène

Le Gène est un élément d'information héréditaire situé sur un chromosome ou un locus donné, Chaque gène correspond à un caractère héréditaire particulier et constitue donc une unité d'information génétique.

Plus précisément le gène est un fragment d'ADN contenant toutes les informations nécessaires pour produire un ARN ou, le plus souvent, une protéine. Un gène correspond à une instruction à effectuer par la cellule.

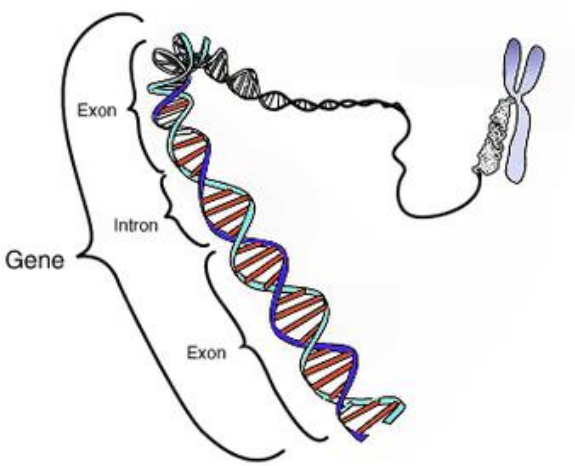

Figure 8 : Gène

2.4. Génome

Le mot « génome » est la combinaison des mots « gène » et « chromosome ».

Le Génome est l'ensemble de l'information génétique d'un organisme contenu dans chacune de ses cellules sous la forme de chromosomes, Le support matériel du génome est l'ADN, sauf chez certains virus où il s'agit d'ARN [10].

2.5. Protéine

Les protéines constituent le principal matériel de construction des êtres vivants, elles jouent un rôle actif et vital dans le fonctionnement des cellules « enzymes, anticorps, antigènes, toxines ... » [11].

2.5.1. Structure des protéines

Les protéines peuvent être décrites selon quatre niveaux d'organisation structurale.

a) Structure primaire

Une séquence linéaire d'acides aminés, formant une chaîne polypeptidique, constitue la structure primaire de la protéine. Cette structure qui ressemble à un chapelet de « perles » d'acides aminés, est le squelette de la molécule de protéine [11].

b) Structure secondaire

Les protéines n'existent pas sous forme de chaînes linéaires d'acides aminés, elles se tordent et se replient sur elles-mêmes « C'est leur structure secondaire ». La structure secondaire la plus courante est celle de l'hélice alpha (α). Dans l'hélice alpha la chaîne primaire s'enroule sur elle-même puis est stabilisée par des liaisons hydrogène entre les groupes NH et CO, à tous les quatre acides aminés environ [11].

Le feuillet plissé bêta (β) est une autre structure secondaire, où les chaînes polypeptidiques primaires ne s'enroulent pas mais se lient côte à côte au moyen de liaisons hydrogène et forment une sorte d'échelle pliante [11].

c) Structures tertiaire et quaternaire

Un grand nombre de protéines se complexifient jusqu'à la structure tertiaire, une structure très spécifique formée à partir de la structure secondaire, dans une structure tertiaire des régions hélicoïdales ou plissées de la chaîne polypeptidique se replient les unes sur les autres et forment une molécule en forme de boule, ou molécule globulaire. La structure tertiaire est maintenue par des liaisons (covalentes, hydrogène ...) entre des acides aminés souvent très éloignés sur la chaîne primaire [12].

La structure quaternaire correspond à l'association spécifique de plusieurs chaînes peptidiques en une unité d'ordre supérieur seule capable d'assurer complètement les fonctions biologiques, l'hémoglobine possède ce niveau d'organisation structurale dans lequel deux chaînes α a sont associées à deux chaînes β [12].

2.6. Interaction ARN-Protéines

Il existe une grande quantité d'ARN qui peuvent former des associations avec des protéines. Ces interactions sont essentielles à plusieurs processus biologiques comme la biosynthèse des protéines, l'épissage de l'ARN et la réplication des virus ; de nouvelles structures tridimensionnelles de complexe ARN-protéine sont élucidées et nous permettent ainsi de faire des analyses comparatives approfondies des patrons de reconnaissance récurrents entre celles-ci, Ces patrons, que l'on peut nommer motifs structuraux, sont une construction régulière habituellement associée à des fonctions précises. Dans notre cas, ce sont des unités de reconnaissance entre les structures de l'ARN et de la protéine. Il existe plusieurs motifs découverts à ce jour qui semblent posséder des propriétés pour lier l'ARN de façon spécifique. Mais jusqu'à maintenant la communauté scientifique n'a pu apporter de confirmation sur ce sujet [13].

Les protéines peuvent reconnaître et interagir avec certaines localisations spécifiques sur des éléments de la structure d'un ARN. Premièrement, le repliement de la molécule est très important, la forme de l'ARN doit permettre un bon positionnement de la protéine pour lui permettre d'accomplir ses fonctions. L'encombrement stérique de l'ARN ne doit également pas encombrer ces sites fonctionnels. Les éléments qui permettent un bon positionnement spatial de la protéine face à l'ARN et sa fixation sur la molécule d'ARN se retrouvent au niveau de la structure secondaire de cette protéine [13].

Deuxièmement, la protéine interagit avec des éléments spécifiques dans la séquence du nucléotide. Ces contacts sont formés grâce à des ponts hydrogène entre les résidus nucléotidiques et les chaînes latérales ou des portions du squelette des protéines [13].

Troisièmement, l'ARN est une molécule flexible et peut adopter une structure locale non-canonique. Cet aspect important de la reconnaissance des acides nucléiques par la protéine est appelé la déformation séquence dépendante. Cette caractéristique qu'adopte l'ARN favorise l'énergie libre du complexe, En se basant sur les structures de complexe ARN-protéine existantes, il existe différentes stratégies de reconnaissance spécifique de sites d'ARN par des protéines [13].

Malgré le peu de travaux qui a été réalisé jusqu'à maintenant dans le domaine des complexes ARN-Protéine. Ce domaine deviendra de plus en plus important étant donné que

ces complexes deviendront de nouvelles cibles thérapeutiques à certaines maladies pour l'instant incurables. Il devient essentiel de développer de nouvelles méthodes d'analyse des interactions moléculaires. De ce fait, la compréhension des mécanismes d'interaction occasionnera la prédiction précise des complexes potentiels entre les ARN et les protéines. Pour l'instant, le problème demeure entier et toutes les avenues de recherche qui sont exploitées offriront peut-être quelques réponses [13].

2.6.1. Virus à ARN

Le virus une particule microscopique infectieuse qui ne peut se répliquer qu'en pénétrant dans une cellule et en utilisant sa machinerie cellulaire, il est un pathogènes responsables de maladies transmissibles, défini par une structure se résumant à deux ou trois éléments donc il est totalement différent des bactéries ou des parasites.

La découverte des virus trouve son origine au XIX^{e} siècle, avec les travaux d'A. Mayer sur la mosaïque du tabac. Le scientifique allemand a découvert que cette maladie qui touchait les feuilles de tabac pouvait se propager d'une plante à l'autre. Mais il n'a pas pu trouver de bactérie responsable de cette maladie. Plus tard, le microbiologiste Beijerinck comprit que la particule infectieuse devait être bien plus petite qu'une bactérie. Le virus de la mosaïque du tabac (VMT) n'a été identifié qu'en 1935 par Wendell Stanley [14].

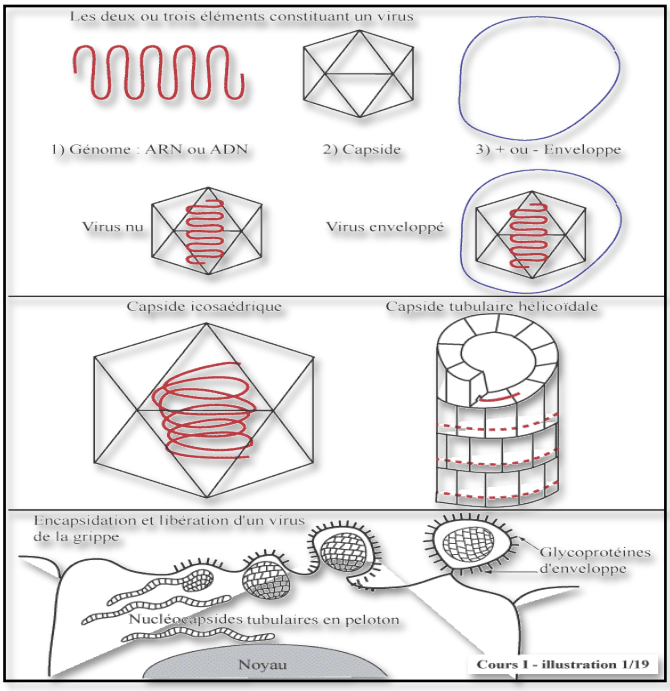

Figure 9 : Eléments constituent le virus

a) Génome viral

Un virus comporte toujours un génome qui est de l'ADN ou de l'ARN, de sorte que dans la classification des virus on distingue en premier lieu virus à ADN et virus à ARN, ce génome peut être monocaténaire (à simple brin) ou bicaténaire (à double brin).

D'une façon générale, la réplication du génome des virus à ARN est beaucoup moins fidèle que celle du génome des virus à ADN**.** Les virus à ARN sont particulièrement sujets aux variations génétiques (par exemple : HIV, virus de l'hépatite C), contrairement aux virus à ADN.

b) Recherche antivirale

La recherche antivirale est une discipline encore récente et le nombre de molécules disponibles pour lutter contre les infections virales demeure insuffisant. Pourtant, tant les pathologies causées par des virus endémiques ou émergents que l'existence de résistance de certains virus aux antiviraux rendent indispensables une recherche constante de nouvelles molécules antivirales.

L'industrie pharmaceutique se tourne aujourd'hui vers de nouvelles solutions antivirales telles que les peptides, qui constituent un nouveau champ d'exploration pour la thérapie. Ça prend créer de façon rationnelle de nouveaux médicaments plus sélectifs et plus efficaces, l'identification et la mise au point de ces molécules nécessitent l'utilisation de nouveaux outils bio-informatiques et des techniques récentes en biologie moléculaire.

2.6.2. Les molécules

La molécule est la structure de toute matière**,** elle est définie par l'UICPA **:** Une molécule est une entité électriquement neutre comprenant plus d'un atome.

Autrement dit la molécule est un ensemble d'atomes (au moins deux) identiques ou non, ces derniers sont liés par différentes forces physiques que l'on appelle liaisons. Là aussi plusieurs types existent : on retrouvera par exemple les liaisons covalentes, les liaisons hydrogènes ou encore les forces de van der Waals [15].

a) Composition et structure des molécules

Les compositions des molécules dépendent de chacune d'entre elles. On peut déduire la composition de ces dernières en fonction de leur formule chimique**,** par exemple CO_2 est une molécule de dioxyde de carbone. Cette dernière est donc composée d'un atome de carbone et de deux atomes d'oxygènes [15].

b) Etat de la molécule

La molécule existe dans plusieurs états, qui déterminent également sa structure :

- **Etat solide :** est l'état le plus "serré" de la matière. Dans ce cas, toutes les molécules sont collées les unes aux autres, c'est ce qui donne à l'élément sa solidité.
- **Etat liquide :** est un intermédiaire entre l'état solide et l'état gazeux, La matière y est **malléable** et coule.
- **Etat gazeux :** est l'état dans lequel la matière est la plus dissipée.

A chaque état correspond une structure particulière, à l'état solide les molécules s'empilent le plus souvent de manière régulière, a l'état liquide l'espacement entre les molécules est petit, ce qui fait qu'elles sont peu agitées. En revanche, à l'état gazeux, les molécules sont espacées et leur agitation est maximale, c'est pour cela que l'on dit qu'un gaz prend toute la place qu'on lui offre.

c) Types de liaisons

On distingue deux types de liaisons :

- **Liaisons chimiques**

  3 types limites de liaisons chimiques:

  Liaison covalente: Se forme entre atomes d'électronégativités voisines « La liaison covalente = un non-métal + un non-métal » [15].

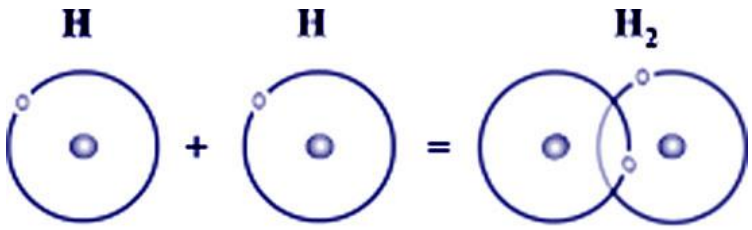

Figure 10 : Liaison covalente « Campus Odontologie Dr. A. RASKIN : Rappels atomistiques, structure des métaux, des alliages et des céramiques « Société Francophone des Biomatériaux Dentaires (SFBD) » »

Liaison ionique: Se forme entre atomes d'électronégativités très différentes « La liaison ionique = métal fort + non-métal fort » [15].

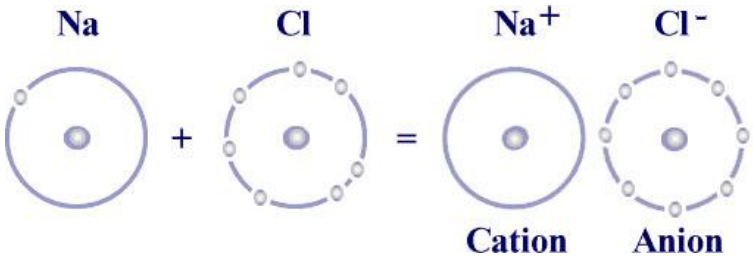

Figure 11 : Liaison ionique « Campus Odontologie Dr. A. RASKIN : Rappels atomistiques, structure des métaux, des alliages et des céramiques « Société Francophone des Biomatériaux Dentaires (SFBD) » »

Liaison métallique: Se forme entre atomes d'électronégativités voisines, cette liaison est donc un ensemble d'ions +. Chaque charge + est entourée par une charge -., elle est beaucoup plus faible que les liaisons ioniques et les liaisons covalente [15].

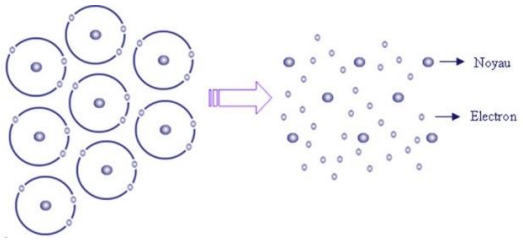

Figure 12 : Liaison métallique « Campus Odontologie Dr. A. RASKIN : Rappels atomistiques, structure des métaux, des alliages et des céramiques « Société Francophone des Biomatériaux Dentaires (SFBD) » »

Liaisons physiques

Deux types limites de liaisons physiques

Liaison hydrogène : Se produit lorsqu'un atome électronégatif (avec un ou plusieurs doublets libres) se trouve à proximité d'un atome d'hydrogène lié de façon covalente à un autre atome électronégatif [15].

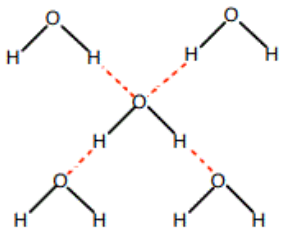

Figure 13 : Liaison hydrogène

Liaisons de Van der Waals : Ce sont des interactions de faibles intensités entre atomes ou molécules, Sont des liaisons électriques comme toutes les liaisons mais ce sont des liens physiques entre molécules, elles sont faibles mais suffisantes pour créer un état liquide [15].

3. Conclusion

Dans le cadre de ce chapitre, nous avons voulu répondre au problème biologique, à savoir s'il existe des facteurs de reconnaissance qui permettent de prédire l'agencement des ARN et des autres molécules entre eux. La formation de complexes ARN-Protéines semblent se réaliser grâce à des parties spatiales compatibles qui forment des ensembles d'interactions spécifiques. Il n'existe pas de moyen automatique de recherche de ces ensembles d'interactions « que nous nommerons motifs » dans les structures résolues jusqu'à présent. Nous avons tenté de solutionner ce problème en élaborant un programme qui détermine les régions structurales répétitives au sein d'un groupe de complexes ARN-Protéines.

Pour la réalisation de cette étude qui vise à améliorer les performances de prédiction des interactions ARN-Protéines en intégrant plusieurs données d'interactome hétérogènes dans les prédictions des interactions ARN-Protéines. La complexité de ces études nous pousse à utiliser des méthodes avancées en informatique que les méthodes traditionnelles, c'est ce que nous allons aborder dans le deuxième chapitre.

Chapitre 2 :

Deep learning

1. Introduction

L'Intelligence Artificielle (abrégée IA) apparue en 1956 est la science dont le but est de faire par une machine des tâches que l'homme accomplit en utilisant son intelligence [17].

Le chercheur français en informatique Jean-Louis Laurière (J.L.Laurière) a défini que l'IA est l'Etude des activités intellectuelles de l'homme pour lesquelles aucune méthode n'est a priori connue [16].

L'Informatique est la science du traitement de l'Information, l'IA s'intéresse à tous les cas où ce traitement ne peut être ramené à une méthode simple, précise, algorithmique. Un algorithme est une suite d'opérations ordonnées, bien définies, exécutables sur un ordinateur actuel, et qui permet d'arriver à la solution en un temps raisonnable (minutes, heures, ou plus, ... mais pas des siècles) [17].

La recherche en IA a donné lieu à des vrais succès et a nourri largement l'histoire des mathématiques et de l'informatique, beaucoup de disciplines ont d'ailleurs profité de cette avancée.

Le domaine de l'IA connait de nos jours une très grande évolution et il risque de changer catégoriquement le monde. La détection de la fraude, les systèmes de recommandation, la reconnaissance faciale et l'automatisation de la prise de décision, ne sont que quelques exemples parmi une variété où l'intelligence artificielle a déjà pris place. Le domaine financier n'a pas été épargné. D'ailleurs, on estime qu'environ plus de 80% des transactions boursières effectuées sur les marchés américains d'actions sont faites par des robots : négociation algorithmique [18].

2. Apprentissage automatique

L'apprentissage automatique (appelé aussi apprentissage machine ou bien Machine Learning en anglais) est une branche de l'IA qui concerne la conception et le développement d'algorithmes permettant à un ordinateur (une machine au sens large) d'apprendre à exécuter des tâches très complexes sans avoir été explicitement programmé (Koza, Bennett, Andre, & Keane, 1996) [19].

Le ML est une branche consiste à programmer des algorithmes permettant d'apprendre automatiquement à partir des données et d'expériences passées ou par interaction avec l'environnement. Ce qui rend l'apprentissage machine vraiment utile est le fait que l'algorithme peut "apprendre" et adapter ses résultats en fonction de nouvelles données sans aucune programmation à priori [19].

2.1. Les grandes classes d'apprentissage automatique

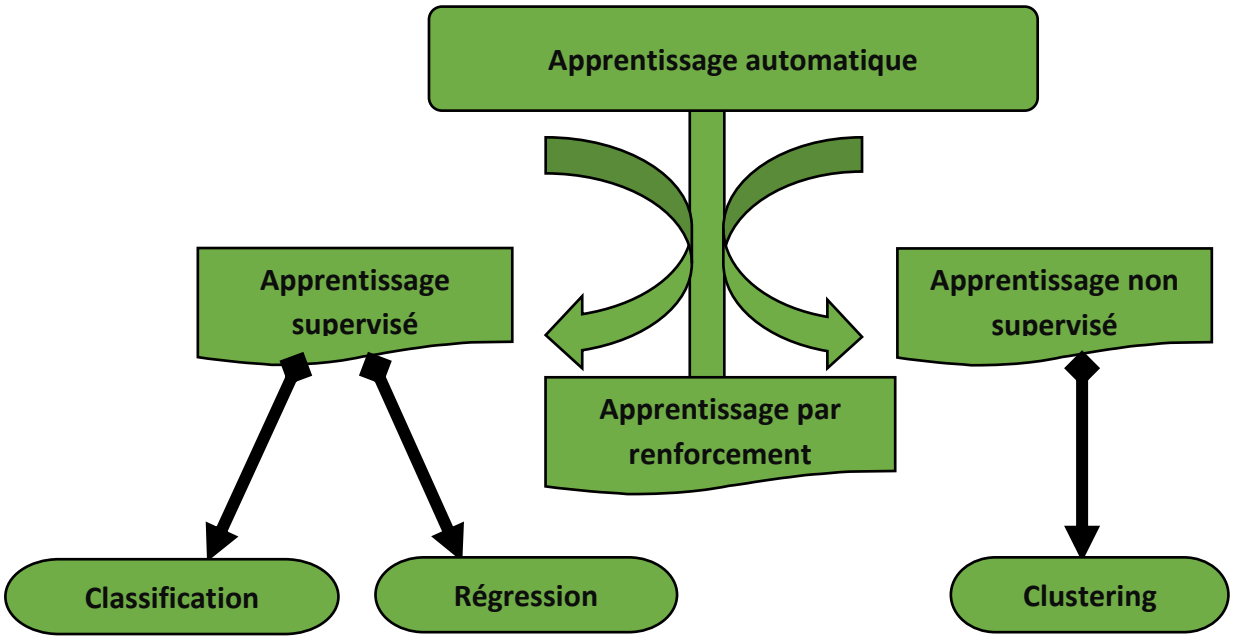

Figure 14 : Grandes classes d'apprentissage automatique

2.1.1. L'apprentissage supervisé

L'algorithme est entrainé en utilisant une base de données d'apprentissage contenant des exemples de cas réels traités et validés. L'objectif est de trouver des corrélations entre les

données d'entrée (variables explicatives) et les données de sorties (variables à prédire), pour ensuite inférer la connaissance extraite sur des entrées avec des sorties inconnues [19].

En apprentissage supervisé, on distingue entre deux types de tâches :

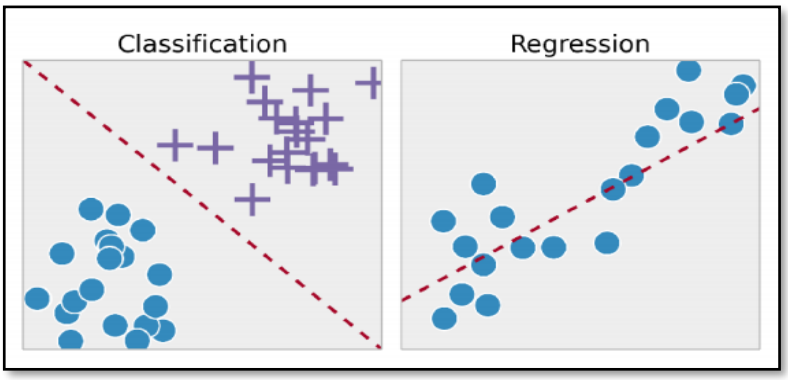

Figure 15: Illustration de la différence entre classification linéaire et régression linéaire « initiez-vous au machine learning ; Openclassroom »

2.1.2. L'apprentissage non-supervisé

Contrairement à l'apprentissage supervisé, l'apprentissage non supervisé est utilisé pour tirer des conclusions et trouver des tendances à partir de données d'entrée sans étiquettes (ou labels). Cela retourne des résultats étiquetés et fait apparaître des « catégories ». Les deux principales méthodes utilisées dans l'apprentissage non supervisé comprennent le groupement (cluster) et la réduction de la dimensionnalité [20].

Le clustering est une technique consiste à regrouper, ou à mettre en clusters, des points de données. Elle est fréquemment utilisée pour la segmentation de clients, la détection des fraudes et la classification des documents [20].

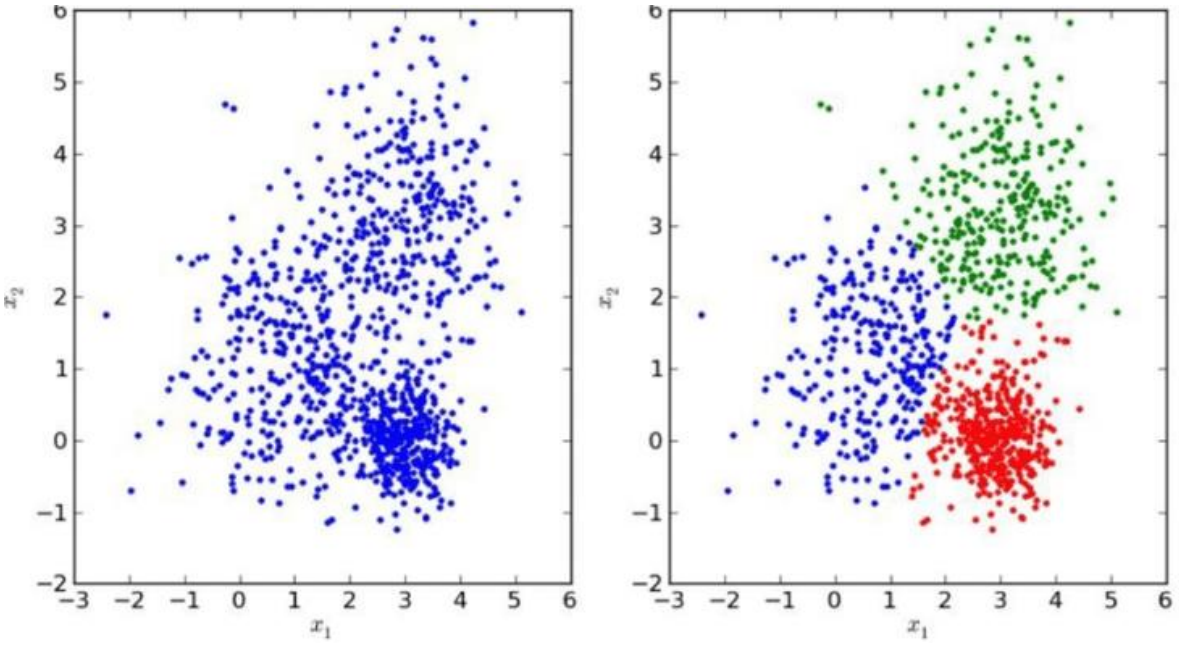

Figure16 : Exemple d'apprentissage non supervisé « MONCOACHDATA »

2.1.3. L'apprentissage par renforcement

L'apprentissage se fait sans supervision, par interaction avec l'environnement (principe d'essai / erreur), en observant le résultat des actions prises. Chaque action de la séquence est associée à une récompense. Le but est de déterminer la stratégie comportementale optimale afin de maximiser la récompense totale. Pour cela, un simple retour des résultats est nécessaire pour apprendre comment la machine doit agir. Ceci est appelé le signal de renforcement [19].

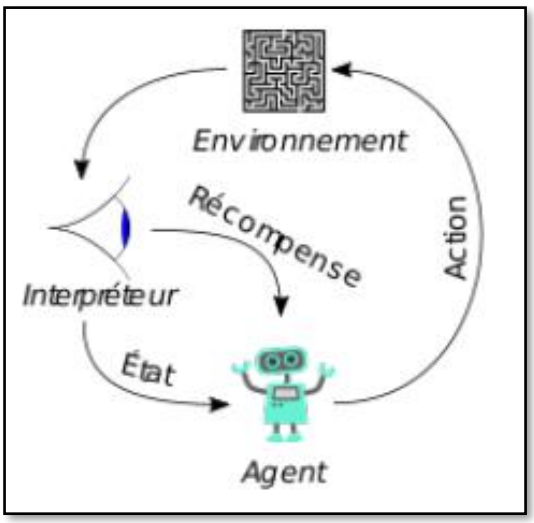

Figure 17: Schéma descriptive de l'apprentissage par renforcement

3. Apprentissage profond

Le Deep Learning « abrégé (DL) ou apprentissage profond » est un sous-domaine d'intelligence artificielle considéré comme une évolution du Maching Learning (apprentissage automatique) où la machine est capable d'apprendre par elle-même, contrairement à la programmation où elle se contente d'exécuter à la lettre des règles prédéterminées [21].

Autrement dit le DL est une forme d'apprentissage fondée sur des approches mathématiques, utilisées pour modéliser des données. Pour mieux comprendre ces techniques, il faut remonter aux origines de l'intelligence artificielle en 1950, année pendant laquelle Alan Turning s'intéresse aux machines capables de penser [21].

Le DL est un système avancé basé sur le cerveau humain, qui comporte un vaste réseau de neurones artificiels. Ces neurones sont interconnectés pour traiter et mémoriser des informations, comparer des problèmes ou situations quelconques avec des situations similaires passées, analyser les solutions et résoudre le problème de la meilleure façon possible [21].

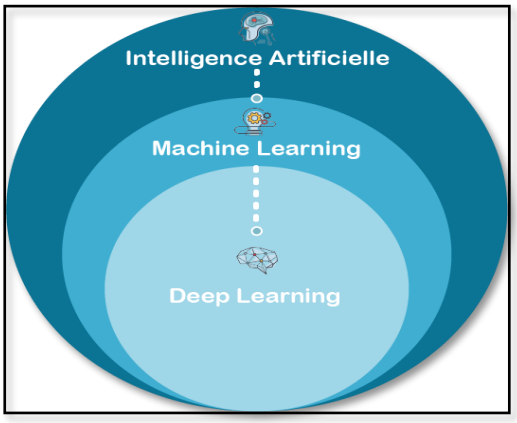

Figure 18 : Relation entre le DL ; ML ; IA « Comprendre le machine Learning et le deep Learning ; Bial-R »

Par exemple, le modèle de DL connu sous le nom de réseau neuronal convolutif peut être entraîné à l'aide d'un grand nombre (des millions) d'images, des images représentant des chats par exemple. Ce type de réseau neuronal tire son apprentissage des pixels contenus dans les images reçues. Il peut classer des groupes de pixels en fonction des caractéristiques du chat telles que les griffes, les oreilles, les yeux indiquant la présence de l'animal dans l'image

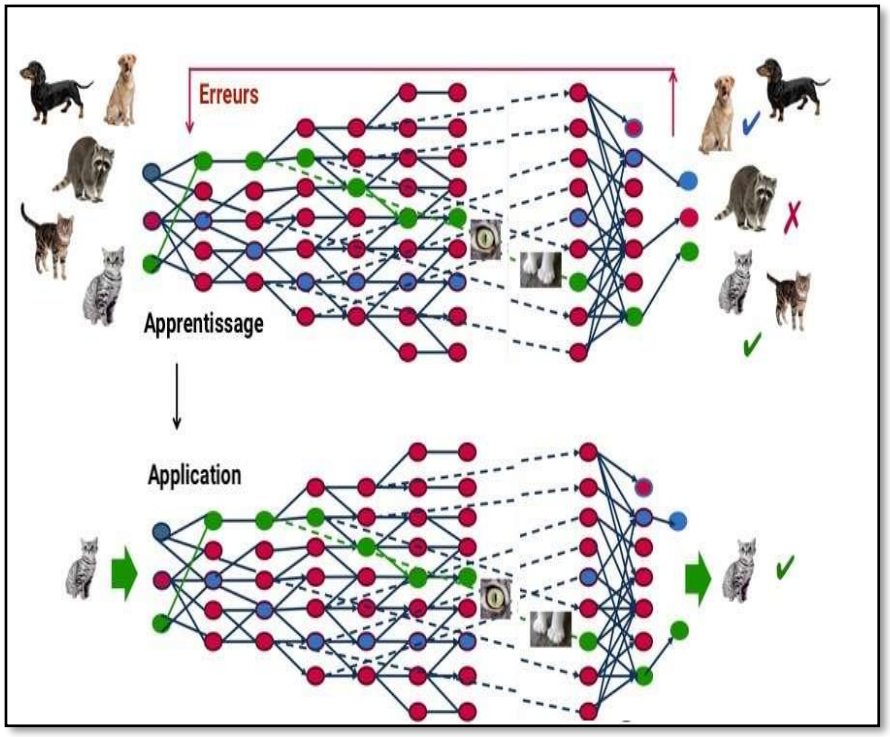

Figure 19 : Identification un chat sur une photo par le deep Learning « MapR, C.D, Futura »

A travers un processus d'autoapprentissage, le Deep Learning est capable d'identifier un chat sur une photo. À chaque couche du réseau neuronal correspond un aspect particulier de l'image.

3.1. Fonctionnement de DL

Le DL s'appuie sur un réseau de neurones artificiel s'inspirant du cerveau humain, ce réseau est composé de dizaines voire de centaines de « couches » de neurones, chacune recevant et interprétant les informations de la couche précédente. Le système apprendra par exemple à reconnaître les lettres avant de s'attaquer aux mots dans un texte, ou détermine s'il y a un visage sur une photo avant de découvrir de quelle personne il s'agit [22].

3.1.1. Réseaux neurones artificiels

Les réseaux de neurones artificiels (artificial neural networks en anglais) constituent une branche spécifique de la recherche en informatique et en neuro-informatique, la présence de différents types de réseaux neurones artificiels donne des diverses possibilités pour le traitement de l'information [23].

a) Modèles de neurone artificiel

Quatre éléments d'un modèle de neurone artificiel sont identifiés comme suit :

- **Structure du réseau**

Un réseau de neurones est en général composé d'une succession de couches dont chacune prend ses entrées sur les sorties de la précédente, Chaque couche (i) est composée de N_i neurones, prenant leurs entrées sur les N_i-1 neurones de la couche précédente, a chaque synapse est associé un poids synaptique, l'effet d'une synapse est incarné par le poids entre deux neurones [23].

- **Fonction de combinaison**
- **Fonction d'activation**

La fonction d'activation (ou fonction de seuillage, ou encore fonction de transfert) sert à introduire une non-linéarité dans le fonctionnement du neurone [23].

- **Propagation de l'information**

Ce calcul effectué, le neurone propage son nouvel état interne sur son axone.

b) Types de neurones artificiels

Quelques types de neurones artificiels sont identifiés comme suit :

- **Perceptron**

La forme la plus simple de réseau neuronal artificiel était composé d'un seul neurone, son rôle est de classifie les données à partir d'une combinaison linéaire de ses entrées [23].

- **Perceptron multicouche**

Appelé aussi Multi-layer Perceptron (MLP) il est un réseau de neurones plus généraux que le perceptron, ils sont composés d'une multitude de neurones interconnectés et organisé en couche successives [23].

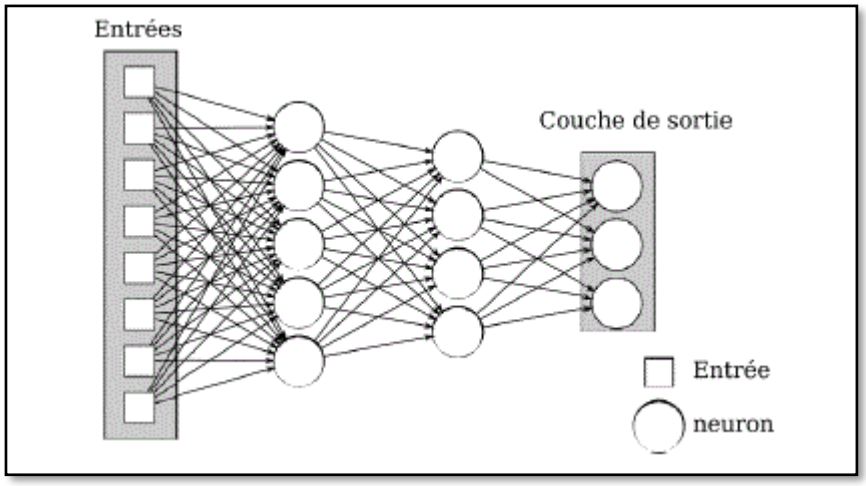

Figure 20 : Exemple d'une représentation d'un MLP « THÈSE DE DOCTORAT - UNIVERSITE DE RENNES 1 COMUE UNIVERSITE BRETAGNE LOIRE – par Corentin HARDY 8 avril 2019 »

- **Rétro-propagation**

C'est un algorithme d'apprentissage adapté aux réseaux de neurones multicouches de nos jours, les réseaux multicouches et la rétro-propagation de gradient reste le modèle le plus productif au niveau des applications [23].

3.1.2. Réseau neurone convolutif

Un réseau de neurone convolutionel est un type particulier de réseau de neurone qui se base sur l'opération de convolution, les réseaux à convolution sont dérivés des architectures de type perceptron multicouches (Multi Layer Perceptron : MLP), cependant ils utilisent des poids partagés, liés à la fenêtre de convolution qui leur permettent une extraction implicite de caractéristiques locales, les CNN sont particulièrement adaptés à la reconnaissance d'images [23].

Un réseau est dit convolutif quand chaque neurone reçoit ses informations non pas de toute la couche précédente, mais seulement des neurones situés dans son champ réceptif [23].

a) La couche de convolution

La couche de convolution est la composante clé des réseaux de neurones convolutifs, et constitue toujours au moins leur première couche, son but est de repérer la présence d'un ensemble des caractéristiques des données reçues en entrée [24].

b) La couche de pooling

Un autre outil très puissant utilisé par les CNNs s'appelle le Pooling. Qui est une méthode permettant de prendre une large image et d'en réduire la taille tout en préservant les informations les plus importantes qu'elle contient [25].

c) Unités Rectifié Linéaire (ReLU)

L'unité rectifié linéaire est la fonction d'activation la plus couramment utilisée dans les modèles d'apprentissage profond [25], sa fonction est définie par : f(x)= max (0, x)

d) Couche entièrement connectée(FC):

Après plusieurs couches de convolution et de max-pooling, le raisonnement de haut niveau dans le réseau neuronal se fait via des couches entièrement connectées (FC=fully-connected), Les neurones dans une couche (FC) ont des connexions vers toutes les sorties de la couche précédente. Les calculs de sa fonction d'activation se fais par une multiplication matricielle suivie d'un décalage de polarisation [25].

e) Couche de perte (LOSS) :

La couche de perte (LOSS) constitue normalement la dernière couche d'un réseau de neurones, Diverses fonctions de perte adaptées à différentes tâches peuvent y être utilisées comme : La perte « Soft max », La perte euclidienne [23]…

3.2. Application du Deep learning

Prenons l'exemple d'application de « DL » dans la recherche médicale, À l'aide du « DL » les chercheurs en cancérologie peuvent dépister automatiquement les cellules cancéreuses. Des équipes de l'Université de Californie à Los Angeles (UCLA) ont conçu un microscope qui génère un ensemble de données de grande dimension afin d'entraîner une application de « DL » à identifier avec précision des cellules cancéreuses.

Les applications de « DL » sont utilisées dans divers secteurs:

1) Reconnaissance d'image
2) Traduction automatique

3) Voiture autonome
4) Recommandations personnalisées
5) Modération automatique des réseaux sociaux
6) Identification de pièces défectueuses
7) Détection de malwares ou de fraudes
8) Robots intelligents.

4. Traitement Automatique du Langage Naturel en français (TAL / NLP)

Le traitement du Langage Naturel « Natural Language Processing (NLP) en anglais ; Traitement Automatique du Langage naturel (TAL) en francais » est l'un des domaines de recherche les plus actifs en science des données actuellement. C'est un domaine à l'intersection du Machine Learning et de la linguistique. Il a pour but d'extraire des informations et une signification d'un contenu textuel [26].

- **Objectif :** Dialoguer naturellement avec une machine comme avec une personne.
- Le TAL est généralement composé de deux à trois grandes étapes :

1) **Prétraitement :** une étape qui cherche à standardiser du texte afin de rendre son usage plus facile
2) **Représentation du texte comme un vecteur** : Cette étape peut être effectuée via des techniques de sac de mots (Bag of Words) ou Term Frequency-Inverse Document Frequency (Tf-IdF). On peut également apprendre des représentations vectorielles (embedding) par apprentissage profond.
3) **Classification :** trouver la phrase la plus similaire… (optionnel).

- **Niveaux de traitement automatique**

Pour traiter le langage naturel, on a besoin d'informations coordonnées et pertinentes sur la langue à des niveaux divers. Le plus souvent on a recours à cinq niveaux de connaissances sur une langue : phonologique, morpho-lexical, syntaxique, sémantique et pragmatique. Ces niveaux se superposent, chacun apportant des problèmes spécifiques à résoudre relatif à un niveau donné. Cela nous donne la hiérarchie suivante [27].

- **Le niveau phonologique :** La machine doit reconnaître les signaux acoustiques (domaine de la phonétique) et les identifier en tant que mots. Plus précisément, il s'agit de reconnaître dans le flot sonore les unités acoustiques élémentaires (phonèmes). La difficulté est que la forme acoustique d'un phonème varie selon plusieurs facteurs : le sexe, l'âge, la région, la fatigue, la peur, l'intensité de la parole, etc…
- **Le niveau morpho-lexical :** il concerne l'étude de la formation des mots et de leur variation de formes.

- **Le niveau syntaxique :** il s'intéresse à l'agencement des mots et à leurs relations structurelles.
- **Le niveau sémantique :** se consacre au sens des énoncés.
- **Le niveau pragmatique :** prend en compte le contexte d'énonciation.

5. Deep Learning dans la prédiction des interactions ARN-Protéines

De nombreuses méthodes de calcul ont été proposées pour prédire les interactions ARN-protéine, parmi eux :

5.1. Identification des couples d'interaction ARN-Protéines

Beaucoup de méthodes basées sur l'apprentissage profond ont été développées pour prédire les paires d'interactions ARN-protéine où les modèles sont entraînés par des paires ARN-protéine construites à partir de complexes ARN-protéine.

Par exemple, IPMiner (Pan, Fan, Yan et Shen,2016) entraîne un auto-encodeur empilé pour apprendre des caractéristiques abstraites à partir de la fréquence k-mer, puis les caractéristiques abstraites apprises sont introduites dans un RF pour classer les paires ARN-protéine.

Enfin, un classificateur d'ensemble empilé combine différents prédicteurs pour améliorer les performances de prédiction. Basé sur le cadre de l'IPMiner, RPI-SAN introduit des informations de conservation dans les profondeurs cadre d'apprentissage et obtient en outre de meilleures performances (Yi et al., 2018).

En plus de l'auto encodeur empilé, les CNN sont utilisés pour prédire les paires d'interaction ARN-protéine. RPIFSE combine des CNN profonds et des méthodes de sélection de fonctionnalités dans un ensemble moyen de classer les paires ARN-protéine (L. Wang et al., 2019).

ELM utilise d'abord des CNN profonds pour extraire des fonctionnalités de haut niveau de séquences, qui sont ensuite introduites dans une machine d'apprentissage extrême (ELM) pour la classification (L. Wang et al., 2018).

Actuellement les méthodes basées sur l'apprentissage profond pour prédire les paires d'interaction protéine-ARN ne sont toujours pas de bout en bout, elles utilisent toutes d'abord encodeur automatique ou CNN empilés pour extraire des fonctionnalités de haut niveau, qui sont ensuite intégrées à un modèle d'apprentissage automatique conventionnel.

L'une des raisons est que la méthode de formation de bout en bout pour l'apprentissage profond nécessite un plus grand nombre d'échantillons de formation.

Cependant, le nombre d'échantillons d'entraînement dérivés de complexes ARN-protéine est encore peu nombreux [28].

5.2. Prédiction des sites de liaison ARN-Protéines

Cette méthode est différente de la prédiction des paires ARN-protéines, La prédiction des sites de liaison ARN-protéine se concentre sur la prédiction des sites de liaison RBP « La rétinol-binding protéine » sur les ARNs.

L'entrée n'a besoin que des représentations des ARNs et chaque modèle est entraîné par RBP, étant donné que différents RBP ont différentes spécificités de liaison.

Les sites de liaison sont générés à l'aide d'une technologie de séquençage à haut débit, et un grand nombre de sites de liaison pour des protéines individuelles sont collectés.

Par rapport à la prédiction des paires ARN-protéine, les méthodes de protéine spécifique ne peuvent prédire que les ARN de liaison des protéines spécifiques qu'avec suffisamment de données d'apprentissage [28].

DeepBind est la première méthode basée sur CNN pour déduire la préférence ARN/ADN de liaison des RBP en utilisant un nucléotide codé à chaud (Alipanahi et al., 2015), et il détecte les motifs de liaison à partir des paramètres appris de la convolution couche.

DeepBind surpasse de loin les méthodes de pointe. DeeperBind (Hassanzadeh & Wang, 2016) ajoute une Couche LSTM sur DeepBind pour apprendre la longue dépendance au sein des séquences afin d'améliorer encore les performances de prédiction. DanQ utilise également une architecture similaire composée de CNN et de LSTM pour estimer l'impact des mutations (Quang & Xie, 2016) [28].

6. Conclusion

Ce chapitre a pour objectif de définir les concepts qui seront discutés dans ce projet. Il peut être réparti en trois parties. Dans la première, nous allons présenter les concepts des techniques d'apprentissage automatique « Machine Learning », suivie par l'apprentissage profond « Deep Learning » dans la seconde partie puis dans la dernière partie nous aborderons le traitement automatique du lagunage naturel « NLP ».

Dans le chapitre suivant, nous allons présenter notre contribution …

Chapitre 3 :

Processus de la prédiction des interactions ARN-protéines basée sur DL

1. Description globale de l'approche proposée

Cette partie décrit l'approche développée pour atteindre l'objectif cible. Ce dernier vise à améliorer les performances de prédiction des interactions ARN-Protéines tant que les méthodes manuelles sont difficiles et donnent des résultats incorrects. Il est nécessaire de développer des outils bio-informatiques dédiés et adaptés.

Pour une description globale du déroulement de l'approche proposée, On distingue deux phases, la 1ère phase prétraitement de donnée et une classification binaire. La 2ème phase est la prédiction d'une nouvelle protéine inhibitrice ciblée pour les ARN-viraux.

Le traitement des données se fait en une succession d'étapes :

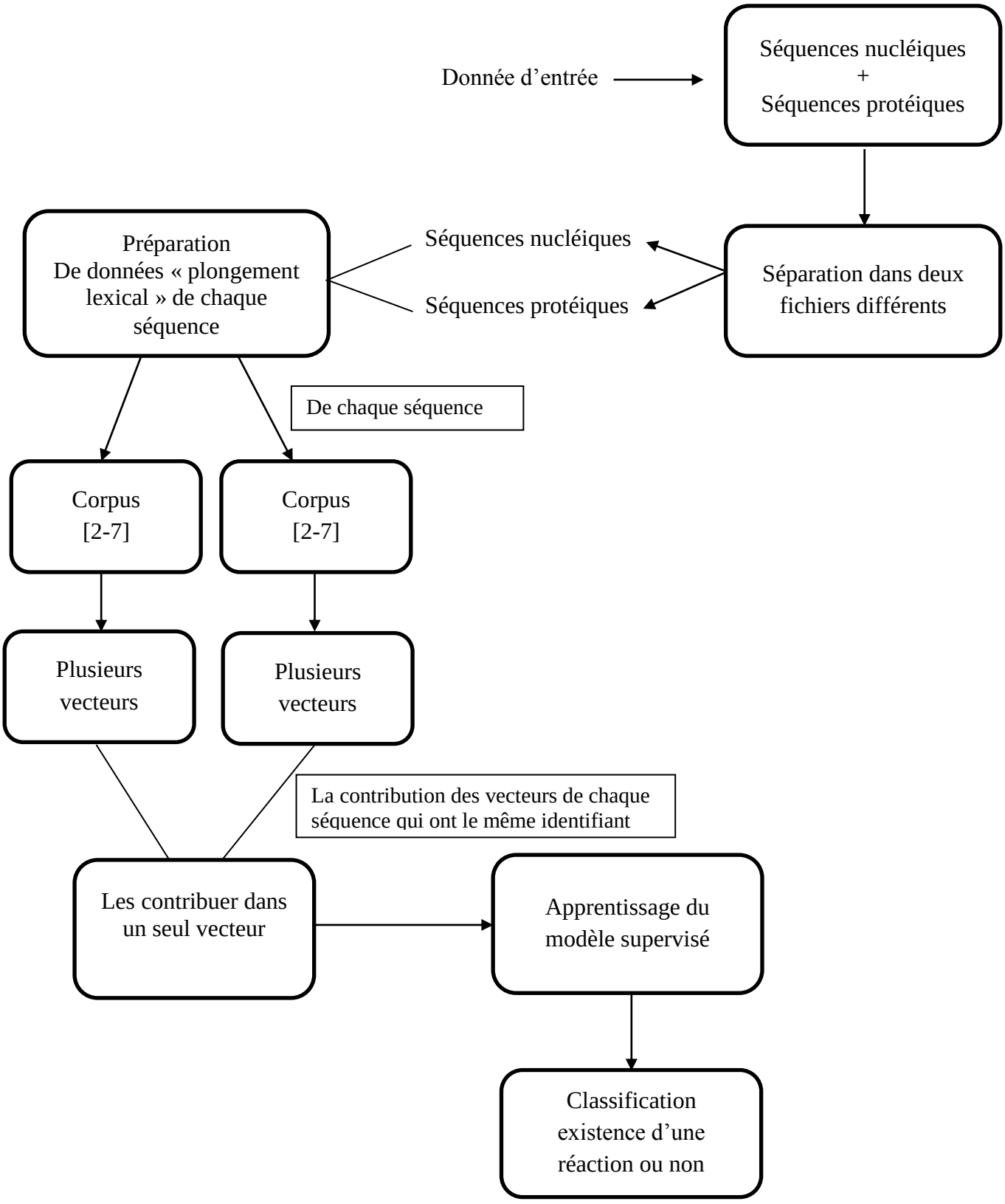

Figure 21 : Phase -1- prétraitement de donnée et une classification binaire sur un modèle supervisé

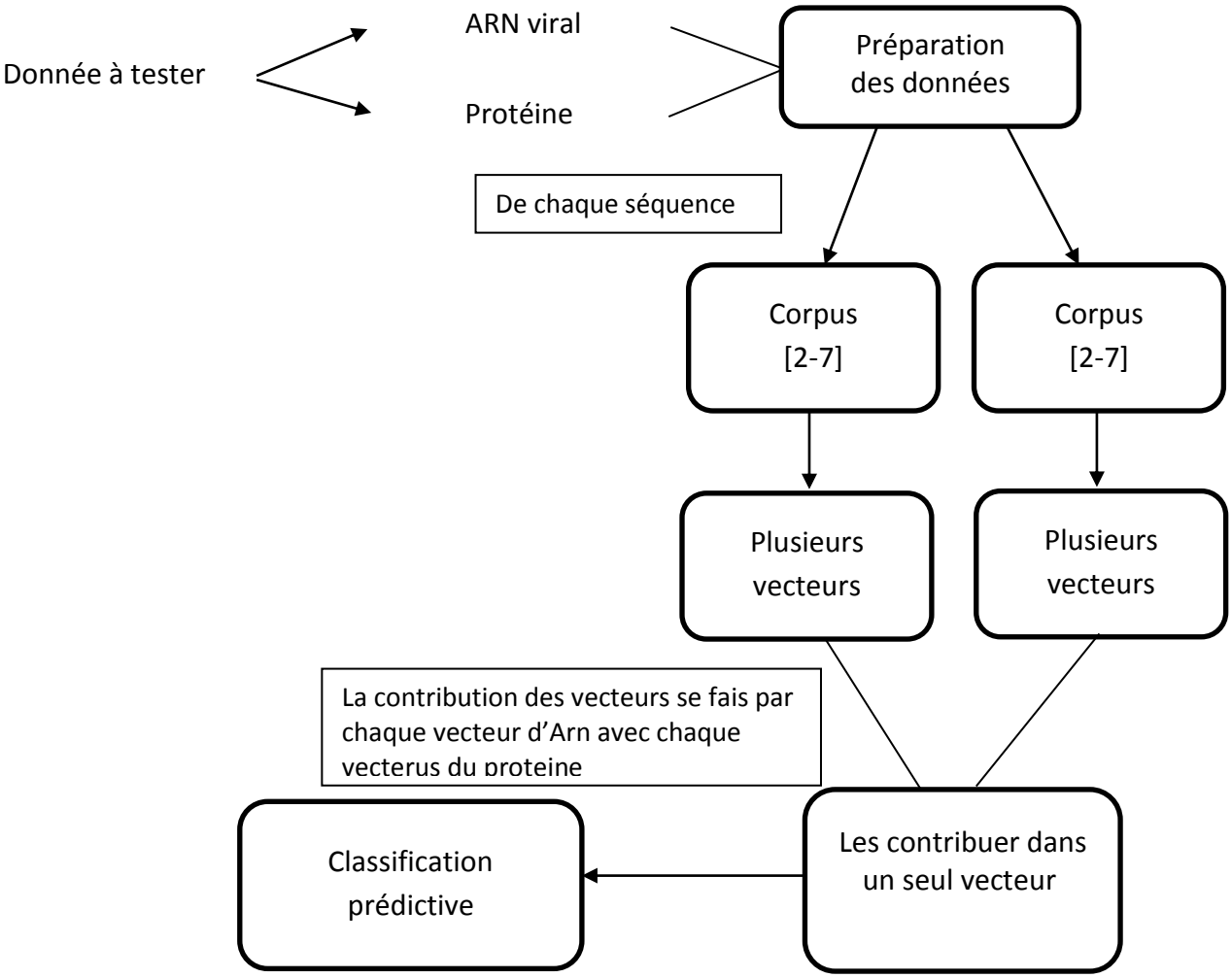

Figure22 : Phase -2- prédiction d'une protéine inhibitrice ciblé pour les ARN viraux

2. Description détaillée de l'approche proposée

2.1. Prétraitement de données et l'apprentissage d'un modèle supervisé

Les données extraites du fichier CSV « récupéré sur » ont été placées dans deux fichiers différents un fichier pour les séquences nucléiques et un autre pour les séquences protéiques en conservant le classement des séquences et l'ID Number.

Par un modèle Biovec pour le plongement lexical (word embedding), ce modèle est un réseau de neurones artificiels à deux couches entraînés pour reconstruire le contexte linguistique des mots. Autrement dit, divisé en ordre les séquences des deux fichiers sous forme des corpus et les mettre dans un seul vecteur.

Afin de pouvoir procéder à l'entrainement, les données doivent être transformé, c'est-à-dire convertir les séquences protéiques et les séquences nucléiques en données numériques traitables par le réseau de neurones convolutif. Ensuite ces données sont divisées en deux échantillons, le 1er pour le résultat qui contient une interaction, et le second pour le résultat qui ne contient pas une interaction. Ensuite, le modèle est créé, entrainé et évaluer, et on modifie à chaque fois les paramètres du modèle jusqu'à l'obtention d'un résultat satisfaisant. Une fois les paramètres optimaux du modèle trouvé, le modèle est enregistré.

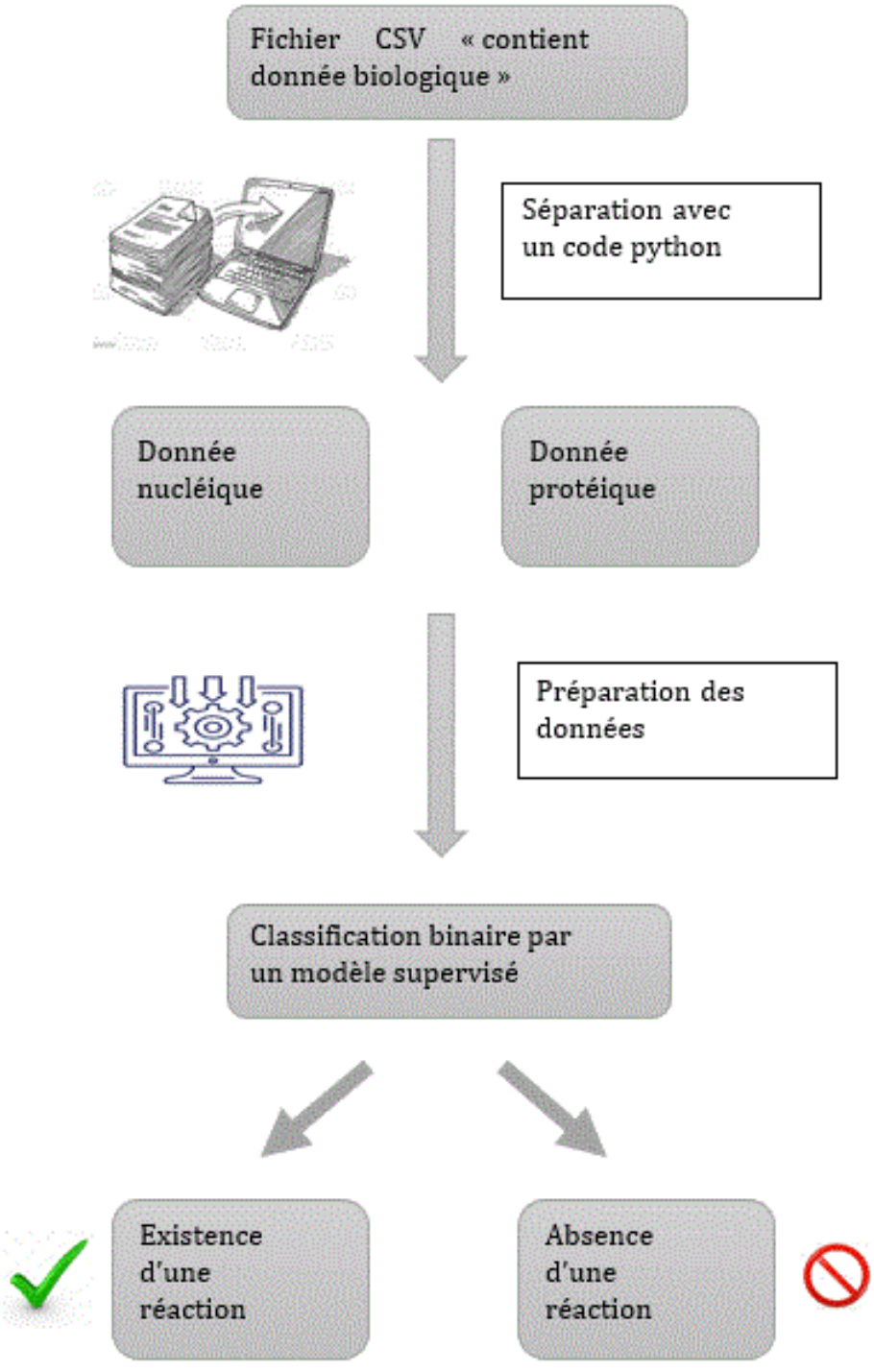

Figure23 : Processus de la première étape

2.2. Prédiction

Une fois le modèle enregistré, il suffit de récupérer des autres petites séquences protéiques qui ont une activité inhibitrice et des autres séquences nucléiques des ARN viraux et travailler de la même manière qu'avant pour extraire des données numériques mais que chaque vecteur des séquences du ARN viral sera contribuer avec tous les vecteurs du séquences protéiques et les comparer par les résultats optimaux de la première étape, cela peut prédire la protéine inhibitrice correspondant à l'ARN viral proposé.

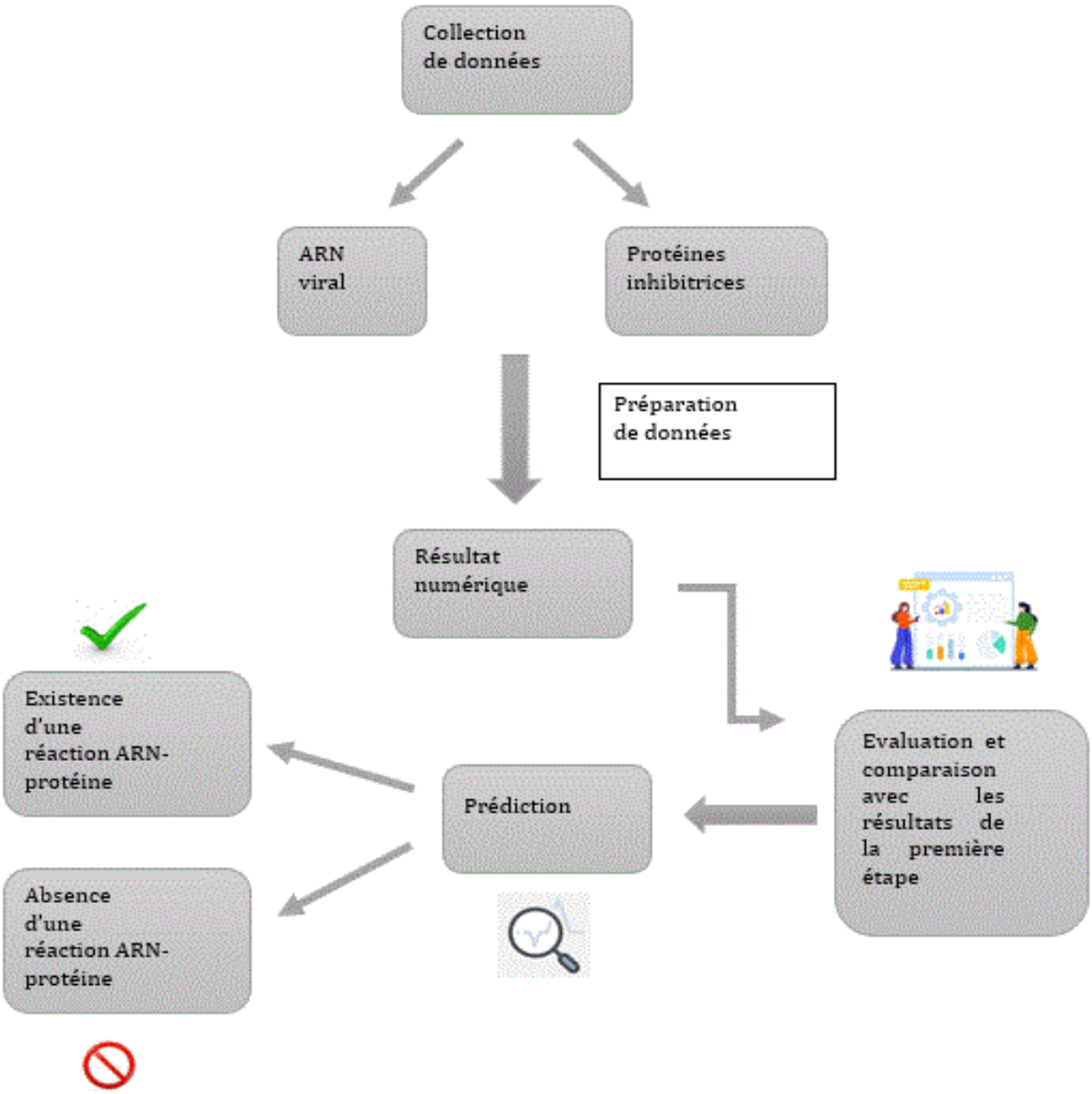

Figure24 : Prédiction d'une protéine inhibitrice

Chapitre 4 :

Implémentation et discussion

1. Données utilisés

1.1. Données biologiques

Nous allons travailler sur un complexe des séquences « ARN-Protéines », le premier ensemble de données provient d'un ensemble de séquences nucléiques des ARN, la deuxième donnée c'est un ensemble de séquences protéiques. Voici un exemple du complexe ARN-Protéines.

```
>1FFK_A_0|interactive
GRRIQGQRRGRGTSTFRAPSHRYKADLEHRKVEDGDVIAGTVVDIEHDPARSAPVAAVEFEDGDRRLILAPEGVGVGDELQVGVDAEIAPGNTLPLAEIPEGVPVCNVESSPGDGGKFARASG
VNAQLLTHDRNVAVVKLPSGEMKRLDPQCRATIGVVGGGGRTDKPFVKAGNKHHKMKARGTKWPNVRGVAMNAVDHPFGGGGRQHPGKPKSISRNAPPGRKVGDIASKRTGRGGNE
UUGGCUACUAUGCCAGCUGGUGGAUUGCUCGGCUCAGGCGCUGAUGAAGGACGUGCCAAGCUGCGAUAAGCCAUGGGGAGCCGCACGGAGGCGAAGAACCAUGGAUUUCCGAAUGAGAAUCUC
UCUAACAAUUGCUUCGCGCAAUGAGGAACCCCGAGAACUGAAACAUCUCAGUAUCGGGAGGAACAGAAAACGCAAUGUGAUGUCGUUAGUAACCGCGAGUGAACGCGAUACAGCCCAAACCGA
AGCCCUCACGGGCAAUGUGGUGUCAGGGCUACCUCUCAUCAGCCGACCGUCUCGACGAAGUCUCUUGGAACAGAGCGUGAUACAGGGUGACAACCCCGUACUCGAGACCAGUACGACGUGCGG
UAGUGCCAGAGUAGCGGGGGUUGGAUAUCCCUCGCGAAUAACGCAGGCAUCGACUGCGAAGGCUAAACACAACCUGAGACCGAUAGUGAACAAGUAGUGUGAACGAACGCUGCAAAGUACCCU
CAGAAGGGAGGCGAAAUAGAGCAUGAAAUCAGUUGGCGAUCGAGCGACAGGGCAUACAAGGUCCCUCGACGAAUGACCGACGCGCGAGCGUCCAGUAAGACUCACGGGAAGCCGAUGUUCUGU
CGUACGUUUUGAAAAACGAGCCAGGGAGUGUGUCUGCAUGGCAAGUCUAACCGGAGUAUCCGGGGAGGCACAGGGAAACCGACAUGGCCGCAGGGCUUUGCCCGAGGGCCGCCGUCUUCAAGG
GCGGGGAGCCAUGUGGACACGACCCGAAUCCGGACGAUCUACGCAUGGACAAGAUGAAGCGUGCCGAAAGGCACGUGGAAGUCUGUUAGAGUUGGUGUCCUACAAUACCCUCUCGUGAUCUAU
GUGUAGGGGUGAAAGGCCCAUCGAGUCCGGCAACAGCUGGUUCCAAUCGAAACAUGUCGAAGCAUGACCUCCGCCGAGGUAGUCUGUGAGGUAGAGCGACCGAUUGGUGUGUCCGCCUCCGAG
AGGAGUCGGCACACCUGUCAAACUCCAAACUUACAGACGCCGUUUGACGCGGGGAUUCCGGUGCGCGGGGUAAGCCUGUGUACCAGGAGGGGAACAACCCAGAGAUAGGUUAAGGUCCCCAAG
UGUGGAUUAAGUGUAAUCCUCUGAAGGUGGUCUCGAGCCCUAGACAGCCGGGAGGUGAGCUUAGAAGCAGCUACCCUCUAAGAAAAGCGUAACAGCUUACCGGCCGAGGUUUGAGGCGCCCAA
AAUGAUCGGGACUCAAAUCCACCACCGAGACCUGUCCGUACCACUCAUACUGGUAAUCGAGUAGAUUGGCGCUCUAAUUGGAUGGAAGUAGGGGUGAAAACUCCUAUGGACCGAUUAGUGACG
AAAAUCCUGGCCAUAGUAGCAGCGAUAGUCGGGUGAGAACCCCGACGGCCUAAUGGAUAAGGGUUCCUCAGCACUGCUGAUCAGCUGAGGGUUAGCCGGUCCUAAGUCAUACCGCAACUCGAC
UAUGACGAAAUGGGAAACGGGUUAAUAUUCCCGUGCCACUAUGCAGUGAAAGUUGACGCCCUGGGGUCGAUCACGCUGGGCAUUCGCCCAGUCGAACCGUCCAACUCCGUGGAAGCCGUAAUG
GCAGGAAGCGGACGAACGGCGGCAUAGGGAAACGUGAUUCAACCUGGGGCCCAUGAAAAGACGAGCAUAGUGUCCGUACCGAGAACCGACACAGGUGUCCAUGGCGGCGAAAGCCAAGGCCUG
UCGGGAGCAACCAACGUUAGGGAAUUCGGCAAGUUAGUCCCGUACCUUCGGAAGAAGGGAUGCCUGCUCCGGAACGGAGCAGGUCGCAGUGACUCGGAAGCUCGGACUGUCUAGUAACAACAU
AGGUGACCGCAAAUCCGCAAGGACUCGUACGGUCACUGAAUCCUGCCCAGUGCAGGUAUCUGAACACCUCGUACAAGAGGACGAAGGACCUGUCAACGGCGGGGGUAACUAUGACCCUCUUAA
GGUAGCGUAGUACCUUGCCGCAUCAGUAGCGGCUUGCAUGAAUGGAUUAACCAGAGCUUCACUGUCCCAACGUUGGGCCCGGUGAACUGUACAUUCCAGUGCGGAGUCUGGAGACACCCAGGG
GGAAGCGAAGACCCUAUGGAGCUUUACUGCAGGCUGUCGCUGAGACGUGGUCGCCGAUGUGCAGCAUAGGUAGGAGACACUACACAGGUACCCGCGCUAGCGGGCCACCGAGUCAACAGUGAA
AUACUACCCGUCGGUGACUGCGACUCUCACUCCGGGAGGAGGACACCGAUAGCCGGGCAGUUUGACUGGGGCGGUACGCGCUCGAAAAGAUAUCGAGCGCGCCCUAUGGCUAUCUCAGCCGGG
ACAGAGACCCGGCGAAGAGUGCAAGAGCAAAAGAUAGCUUGACAGUGUUCUUCCCAACGAGGAACGCUGACGCGAAAGCGUGGUCUAGCGAACCAAUUAGCCUGCUUGAUGCGGGCAAUUGAU
GACAGAAAAGCUACCCUAGGGAUAACAGAGUCGUCACUCGCAAGAGCACAUAUCGACCGAGUGGCUUGCUACCUCGAUGUCGGUUCCCUCCAUCCUGCCCGUGCAGAAGCGGGCAAGGUGAG
GUUGUUCGCCUAUUAAAGGAGGUCGUGAGCUGGGUUUAGACCGUCGUGAGACAGGUCGGCUGCUAUCUACUGGGUGUGUAAUGGUGUCUGACAAGAACGACCGUAUAGUACGAGAGGAACUAC
GGUUGGUGGCCACUGGUGUACCGGUUGUUCGAGAGAGCACGUGCCGGGUAGCCACGCCACACGGGGUAAGAGCUGAACGCAUCUAAGCUCGAAACCCACUUGGAAAAGAGACACCGCCGAGGU
CCCGCGUACAAGACGCGGUCGAUAGACUCGGGGUGUGCGCGUCGAGGUAACGAGACGUUAAGCCCACGAGCACUAACAGACCAAAGCCAUCAU
```

Figure25 : Fichier contient des interactions ARN-protéines

La 1ère ligne est l'en-tête des deux séquences et commence par un > : elle représente l'identifiant du complexe.

La 2ème ligne représente la séquence protéique jusqu'à la fin de son enchainement. Cette séquence est classiquement représentée par une chaîne de caractères qui utilise un alphabet de vingt lettres des acides aminés.

Après avoir terminé la séquence d'acides aminés, nous passons à la ligne nous trouvons la séquence nucléotidique représentée par une chaîne de caractères.

Ces données se présentent en fichier CSV (Comma-separated values), et contiennent beaucoup d'attributs. Ceux qui ont été utilisées sont : l'entrée, la séquence protéique et la séquence nucléique.

1.2. Outils informatiques

Notre travail consiste à utiliser des moyens plus développées loin des moyens manuels

1.2.1. Environnement de travail

- **Google Colab**

Google Colaboratory ou Colab, un outil Google simple et gratuit pour vous initier au Deep Learning ou collaborer avec vos collègues sur des projets en science des données, Colab permet :

- D'améliorer vos compétences de codage en langage de programmation Python.
- De développer des applications en Deep Learning en utilisant des bibliothèques Python populaires telles que Keras, TensorFlow, PyTorch et OpenCV.
- D'utiliser un environnement de développement (Jupyter Notebook) qui ne nécessite aucune configuration [29].

- **Notebook Jupyter**

Jupyter Notebook est une application Web open source qui vous permet de créer et de partager des documents contenant du code en direct, des équations, des visualisations et du texte narratif. Les utilisations incluent : le nettoyage et la transformation des données, la simulation numérique, la modélisation statistique, la visualisation des données, l'apprentissage automatique et bien plus encore [30].

1.2.2. Bibliothèques Python utilisées

- **NumPy**

NumPy est une bibliothèque pour le langage de programmation Python qui permet plus de stockage de données avec moins de mémoire. Avec un tableau multidimensionnel et d'autres ressources, Il dispose d'un grand nombre de fonctions mathématiques qui peuvent

être appliquées directement à un tableau. Dans ce cas, la fonction est appliquée à chacun des éléments du tableau [31].

- **Pandas**

Pandas est un outil d'analyse et de manipulation de données open source rapide, puissant, flexible et facile à utiliser, construit sur le langage de programmation Python [33].

- **Matplotlib**

Matplotlib est une bibliothèque complète permettant de créer des visualisations statiques, animées et interactives en Python, Il rend les choses faciles et les choses difficiles possibles. Matplotlib est hébergé sur GitHub, elle est distribuée librement et gratuitement [34].

- **Biovec**

Biovec est une nouvelle approche pour représenter les séquences biologiques.9 Il est léger permet des paramètres et une texture définis par l'utilisateur et s'exécute sur les plates-formes Windows ou Linux [36].

- **Keras**

Keras est une bibliothèque open source de composants de réseaux neuronaux écrits en Python, Composée d'une bibliothèque de composants d'apprentissage automatique couramment utilisés, notamment des objectifs, des fonctions d'activation et des optimiseurs, la plate-forme open source de Keras prend également en charge les réseaux de neurones récurrents et convolutifs. De plus, Keras propose le développement de plates-formes mobiles pour les utilisateurs souhaitant mettre en œuvre des modèles d'apprentissage en profondeur sur les smartphones, à la fois iOS et Android. En 2018, la bibliothèque avait une utilisation de 22% par rapport à ses plus de 200 000 utilisateurs [32].

- **Sklearn**

Sklearn est une Open source utilisable commercialement, Accessible à tous et réutilisable dans divers contextes et construit sur NumPy, SciPy et matplotlib, Il vise à

apporter des solutions simples et efficaces aux problèmes d'apprentissage. Les fonctionnalités fournies par scikit-learn incluent :

- Classification : Identifier à quelle catégorie appartient un objet.
- Régression : Prédiction d'un attribut à valeur continue associé à un objet.
- Clustering : Regroupement automatique d'objets similaires en ensembles.
- Réduction de la dimensionnalité : Réduire le nombre de variables aléatoires à considérer.
- Sélection du modèle : Comparer, valider et choisir des paramètres et des modèles.
- Prétraitement : Extraction et normalisation de caractéristiques [35].

- **TensorFlow**

La principale bibliothèque Open Source développé par des chercheurs de Google pour exécuter l'apprentissage automatique, l'apprentissage profond et d'autres charges de travail d'analyse statistique et prédictive.

2. Approche utilisée

Cette section présente l'implémentation des étapes décrite en chapitre trois « 3 », Dans la suite nous détaillons chaque étape :

2.1.1. Prétraitement de données et apprentissage d'un modèle supervisé

Cette étape inclut deux sous étapes qui sont :

- **Prétraitement de donnée**

Récupération des données

On récupère les séquences nucléiques et les séquences protéiques sur la banque de données et on les sépare dans deux fichiers différents.

```
1- Prétraitement de données

[ ] from google.colab import drive
    drive.mount('/content/drive')

[ ] cd /content/drive/MyDrive/Hamlaoui_Kerriche

file =open("RNA-SmallProtein.txt")
protein_file = open("protein.fasta","w")
rna_file = open("rna.fasta","w")
...
for line in file:
  if(">" in line):
    protein_file.write(line)
    ...
    ...
    ...
  else :
    if(i==0):
      protein_file.write(line)
      ...

      i = 1
    else :
      if(i==1):
        rna_file.write(line)
        ...
        i = 0
```

Figure26 : Récupération de données

Préparation des données

Avec un modèle Word2vec qui est responsable à mettre chaque séquence de chaque fichier en format des corpus et rassembler les séquences de deux fichier qui ont même identifiant dans un seul vecteur « plongement lexical ».

```
[ ] import biovec
    pv_protein = biovec.models.ProtVec("protein.fasta", corpus_fname="protein.corpus", n=6)
    pv_protein.save('myprotein.model')

[ ] pv_rna = biovec.models.ProtVec("rna.fasta", corpus_fname="rna.corpus", n=6)
    pv_rna.save('myrna.model')

    import biovec
    pv_protein = biovec.models.load_protvec("myprotein.model")
    pv_rna = biovec.models.load_protvec("myrna.model")
    vectorized = open("vectorized_data.csv","w")
    file =open("RNA-SmallProtein.txt")
    i = 0
    for line in file:
      if(">" in line):
        ...
        ...
      else :
        if(i==0):
          ...
          ...
          ...
          i = 1
        else :
          if(i==1):
            ...
            ...
            ...
            i = 0
```

Figure27: Préparation de données

- **Apprentissage d'un modèle supervisé**

Cette phase distingue deux étapes principales

Chargement des bibliothèques python nécessaires

Pour initier le travail sur le modèle de deep Learning, il est nécessaire d'importer les bibliothèques nécessaires. Selon les commandes affichées

2- Apprentissage d'un modéle supervisé

```
import numpy as np
import biovec
from keras import backend as K
from sklearn.model_selection import train_test_split
import matplotlib.pyplot as plt
import pandas as pd
...
...
...
...
import sklearn
import os
import numpy
...
...
```

Figure28 : Chargement des bibliothèques

Application du modèle supervisé

Le modèle est désormais prêt à être utilisé pour classer les données qui ont devenu numérique en deux classe soit il existe une interaction ou bien l'absence de cette dernière.

```
[ ] def defCNN():
      model = Sequential()
      ...
      ...
      ...
      model.compile(optimizer='adam', loss='binary_crossentropy')
      return model
```

```
dataset = pd.read_csv("vectorized_data.csv",delimiter="\t")
...
...
...
m = KerasClassifier(build_fn=defCNN, epochs=300, batch_size=2000, verbose=0)
results = cross_validate(m, X, Y, cv=5,scoring=['accuracy'])
print(results)
```

Figure29 : Apprentissage du modèle supervisé

2.1.2. Prédiction d'une protéine inhibitrice pour ARN-viral

Dans la suite c'est la prédiction d'une protéine inhibitrice pour les ARN viraux. Des nouvelles séquences d'ARN viral en été collecter dans la banque de donnée « Virus Host DB », et des séquences des protéines inhibitrices dans « NCBI ». Les séquences sont placées en format FASTA dans deux fichiers différents.

Les mêmes procédures de la première phase sont effectuées sur les deux nouveaux fichiers, Et on applique le plongement lexical par le modèle word2vec.

Par le même modèle supervisé l'apprentissage se fait et Les résultats de la deuxième étape sont comparés à ceux de la première.

3. Résultat

Cette présentation suit les étapes du processus, Elle commence par l'implémentation du modèle supervisé afin de rassembler les Reads et extraire s'il existe une interaction entre ces derniers ou non. La dernière partie consiste à prédire la protéine inhibitrice correspondant à l'ARN viral proposé.

3.1.1. Résultats de la phase 01: Prétraitement de données et l'apprentissage d'un modèle supervisé

Dans cette étape, les données issues de la préparation et la séparation seront enregistrées dans deux fichiers différents sur Google drive, Apres l'enregistrement, les données sera organiser sous forme des corpus après les vectorisé dans un seul vecteur par un modèle Word2vec « WordEmbedding Models »

Figure30: Generate corpus file from Fasta file

Afin de s'assurer de la précision de l'algorithme, c'est le tour de l'apprentissage d'un modèle supervisé pour une classification binaire et c'est présenté dans la figure qui suit

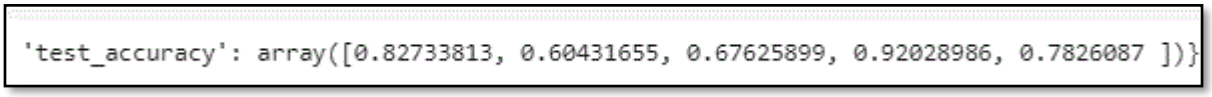

Figure31: Résultat du modèle supervisé

4. Discussion

L'étude des interactions ARN-Protéines est essentielle à plusieurs processus biologiques. Ce domaine deviendra de plus en plus important étant donné que ces complexes deviendront de nouvelles cibles thérapeutiques à certaines maladies mais leurs études par les méthodes classiques « In vitro » sont très couteuses et prends beaucoup de temps, et en plus sont fait sur des petites quantités.

Par contre de nombreuses nouvelle méthodes de calcul « in silico » basées sur le DL ont été proposées pour prédire les interactions ARN-protéine par exemple : IPMiner entraine un auto-encodeur empilé pour apprendre des caractéristiques abstraites à partir de la fréquence k-mer, puis les caractéristiques abstraites apprises sont introduites dans un RF pour classer les paires ARN-protéine. Enfin, un classificateur d'ensemble empilé combine différents prédicteurs pour améliorer les performances de prédiction.

Un autre exemple différent de la prédiction des paires ARN-protéine, la prédiction des sites de liaison ARN-protéine se concentre sur la prédiction des sites de liaison RBP sur les ARN$_S$. Dans un autre cas aussi loin de la prédiction par le DL, la prédiction des sites de liaison RBP ce fait par le docking moléculaire.

L'objectif et l'implémentation de l'approche proposée est différent du reste, cette approche vise les ARN viraux donc elle a une cible thérapeutique à certaines maladies.

Dans la première phase du processus, les données sont vectorisées pour prendre toutes les prédictions du site d'ARN et protéiques disponibles. On considère qu'une précision de 80 % et plus signifie qu'il existe une interaction entre l'ARN est la protéine donnée. Dans la deuxième phase on compare les résultats avec ceux de la première phase pour réaliser la prédiction.

Conclusion

Conclusion

Dans ce mémoire, nous avons étudié de manière approfondie les méthodes basées sur l'apprentissage profond pour prédire les interactions ARN-Protéines. Ainsi ; nous avons décrit comment formuler les prédictions d'interaction ARN-Protéines pour différentes tâches. Ensuite, nous avons fourni les détails sur les modèles de deep-Learning les plus couramment utilisés.

Nous avons proposé une approche pour prédire les interactions ARN-Protéines par un traitement sur les ARN et ses interactions avec les protéines comme une première étape et après de faire un traitement sur les ARN viraux et les protéines inhibitrice pour atteindre notre objectif qui vise les cibles thérapeutiques.

Après la présentation de quelques aspects d'implémentation de l'approche proposée, nous avons discuté les résultats de l'approche de prédiction utilisant l'apprentissage profond et d'autres directions possibles.

Après la simulation du déroulement du processus proposé sur un nombre réduit de données, nous pouvons envisager les perspectives suivantes :

- Utilisation des données volumineuses pour avoir des résultats satisfaisants.
- Utilisation des moyens de calcul puissants comme un HPC.

.

Références

Reference:

1) Dr.nehal fatima université hassiba benbouali chlef « Cour biologie moléculaire et genie genetique »
2) Dr Benaissa.Y Université Oran 1, Faculté de Médecine « Cour génétique moléculaire »
3) Thèse doctorat « Chloé Bessiere » « 27 novembre 2018 Université montpellier » Etude des éléments régulateurs de l'expression des gènes chez l'humain
4) Pr. C. Housset et Pr. A. Raisonnier Université Pierre et Marie Curie « cour biochimie »
5) Dr. YOUSFI « Cours de Biologie Moléculaire » niversité de boira
6) Dr DJEBIEN. S « STRUCTURE DES ACIDES NUCLEIQUES » UNIVERSITE BADJI MOKHTAR-ANNABA
7) Thèse doctorat « Thomas CLEMENT » Recherche de liens entre expression d'ARN non codants et physiopathologies articulaires, utilisation des microARN comme biomarqueurs du phénotype chondrocytaire « 10 septembre 2014 UNIVERSITE DE LORRAINE »
8) Thèse doctorat « clément joret » Etude de la structure et de la fonction d'un complexe constitué de 5 protéines non ribosomiques Npa1p, Npa2p, Dbp6p, Nop8p et Rsa3p essentielles à la formation de la grande sous unité des ribosomes eucaryotes « 19 octobre 2016 université Toulouse »
9) Thèse doctorat « Feifei LIANG » Induction de l'expression génique par des petits ARN dans des cellules de mammifère « 15 Décembre 2011 Université Paris-sud XI Faculté de Médecine »
10) Dr Hammouda.F « organisation des génome » université annaba
11) Dr Tabti M Biochimie et régulation « université chlef »
12) Professeur Michel Seve « les protéines : Définition et Structure » université joseph fourier et Grenoble
13) Article : Deep learning integration of molecular and interactome data for protein-compound interaction prediction « Narumi Watanabe, Yuuto Ohnuki and Yasubumi Sakakibara» Department of Biosciences and Informatics, Keio University, Yokohama, Kanagawa 223-8522, Japan
14) https://www.futura-sciences.com/sante/definitions/medecine-virus-291/

15) Campus Odontologie Dr. A. RASKIN : Rappels atomistiques, structure des métaux, des alliages et des céramiques « Société Francophone des Biomatériaux Dentaires (SFBD) »
16) Thèse de doctorat Benlazaar sid ahmed nadjib « Requête naturelles appliquées à la stratégie managériale en productique (Maintenance Industrielle) : Les Agents autonomes de Conrad ; Université d'Oran Es-Senia »
17) Université Paris 5 - Maîtrise de mathématiques Dominique Pastre, Module INTELLIGENCE ARTIFICIELLE 1999/2000
18) Article : Moustafa Zouinar , Évolutions de l'Intelligence Artificielle : quels enjeux pour l'activité humaine et la relation Humain-Machine au travail ? ; Activités [En ligne], 17-1 | 2020, mis en ligne le 15 avril 2020, consulté le 17 septembre 2021.
19) Rachid MIFDAL : Application des techniques d'apprentissage automatique pour la prédiction de la tendance des titres financiers « ÉCOLE DE TECHNOLOGIE SUPÉRIEURE UNIVERSITÉ DU QUÉBEC 12 novembre 2019 »
20) https://moncoachdata.com/blog/modeles-de-machine-learning-expliques/
21) Thèse de Corentin Hardy. "Contribution au développement de l'apprentissage profond dans les systèmes distribués." 8 avril 2019
22) https://www.futura-sciences.com/tech/definitions/intelligence-artificielle-deep-learning-17262/
23) Claude Touzet. LES RESEAUX DE NEURONES ARTIFICIELS, INTRODUCTION AU CONNEXIONNISME : COURS, EXERCICES ET TRAVAUX PRATIQUES. EC2, 1992, Collection de l'EERIE
24) Analyse fine 2D/3D de véhicules par réseaux de neurones profonds Florian Chabot Université Clermont Auvergne, 2017. Français
25) Khouloud Dahmane. Analyse d'images par méthode de Deep Learning appliquée au contexte routier en conditions météorologiques dégradées. Vision par ordinateur et reconnaissance de formes [cs.CV]. Université Clermont Auvergne, 2020. Français
26) KADI ALLAH Fayçal Magister La Traduction automatique « université d'oran 2011 » : Etat de l'art et les problèmes inhérents.
27) https://www.loukam.net/TALN_Chap1.pdf
28) Recent methodology progress of deep learning for RNA–protein interaction prediction "Xiaoyong Pan* | Yang Yang| Chun-Qiu Xia| Aashiq H. Mirza |Hong-Bin Shen"

29) https://moov.ai/fr/blog/deep-learning-avec-google-colab/
30) https://jupyter.org/
31) https://courspython.com/apprendre-numpy.html
32) https://deepai.org/machine-learning-glossary-and-terms/keras
33) https://pandas.pydata.org/
34) https://matplotlib.org/
35) https://scikit-learn.org/stable/
36) https://pypi.org/project/biovec/

Résumés

RÉSUMÉ

Le but de ce travail est de proposer une approche de prédiction des interactions ARN-molécule basée sur le deep Learning en visant les ARN viraux pour ciblé les maladies. Cette approche est basée sur deux axes de l'intelligence artificielle (IA) : le traitement automatique du langage naturel (NLP) et l'apprentissage profond (DL).

ملخص

الهدف من هذا العمل هو اقتراح منهج للتنبؤ بتفاعلات جزيء الحمض النووي الريبي على أساس التعلم العميق من خلال استهداف الحمض النووي الريبي الفيروسي للحد من الأمراض. يعتمد هذا النهج على محورين للذكاء الاصطناعي :(AI) المعالجة التلقائية للغة الطبيعية. (NLP) والتعلم العميق .(DL)

Abstract

The goal of this work is to propose an approach for predicting RNA-molecule interactions based on deep learning by targeting the RNAviral to target diseases. This approach is based on two axes of artificial intelligence (AI): automatic natural language processing (NLP) and deep learning (DL).

Année universitaire 2020 – 2021	**Présenté et soutenu par :** HAMLAOUI Maria Ouissal KERRICHE Yousra

Prédiction des interactions ARN-Protéines basée sur le Deep-Learning

Mémoire présenté en vue de l'obtention du Diplôme de Master
BIOINFORMATIQUE

Le but de ce travail est de proposer une approche de prédiction des interactions ARN- Protéines basée sur le deep Learning en visant les ARN viraux pour cibler les maladies.

Cette approche est basée sur deux axes de l'intelligence artificielle (IA) : le traitement automatique du langage naturel (NLP) et l'apprentissage profond (DL).

الهدف من هذا العمل هو اقتراح منهج للتنبؤ بتفاعلات جزيء الحمض النووي الريبي على أساس التعلم العميق من خلال استهداف الحمض النووي الريبي الفيروسي للحد من الأمراض.
يعتمد هذا النهج على محورين للذكاء الاصطناعي :(AI) المعالجة التلقائية للغة الطبيعية. (NLP) والتعلم العميق .(DL)

The goal of this work is to propose an approach for predicting RNA-molecule interactions based on deep learning by targeting the RNAviral to target diseases.

This approach is based on two axes of artificial intelligence (AI): automatic natural language processing (NLP) and deep learning (DL).

Président : Dr. TEMAGHOULT Mahmoud (Université Frères Mentouri - Constantine 1)
Rapporteur : **Dr. CHEHILI Hamza** (Université Frères Mentouri - Constantine 1)
Examinateur : **Dr. KELLOU Kamel** (Université Frères Mentouri - Constantine 1)