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 الاخوة منتوري قسنطينة
UNIVERSITÉ DES FRÈRES MENTOURI CONSTANTINE

Faculté des Sciences de la Nature et de la vie — كلية علوم الطبيعة والحياة

Département de Biologie Appliquée — قسم البيولوجيا التطبيقي

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

Exploitation des techniques de Machine Learning (apprentissage automatique) pour la prédiction de la maladie de l'anémie falciforme à partir des séquences d'ADN

Le 18/06/2022

Présenté par :

Boutelala Abir

Merouani Hadil Khadidja

Jury d'évaluation :

Président : Dr GHERBOUDJ AMIRA (Université Frères Mentouri, Constantine1)

Encadreur : Dr OUAHIBA DJAMA (Université Frères Mentouri, Constantine1)

Examinateur : Dr TAMAGOULT MAHMOUD (Université Frères Mentouri, Constantine1)

Année universitaire 2022 – 2023

REMERCIEMENTS

Tout d'abord, je remercie Dieu tout-Puissant de nous avoir accordé la grâce de la raison, la santé et la poursuite de la connaissance et du courage, pour nous avoir donnés la force, la volonté et la patience.

Je remercie chaleureusement mon encadreur Dr OUAHIBA DJAMA pour nous avoir permis de travailler sur un sujet passionnant, pour sa patience, et surtout pour sa confiance, ses commentaires, ses conseils, sa présence et sa gentillesse.

Nous adressons nos sincères remerciements àDr GHERBOUDJ AMIRA

Pour nous avoir honorées en acceptant de présider le jury de ce mémoire.
Nous remercions également Dr TAMAGOULT MAHMOUD d'avoir fait cela pour nous honneur d'avoir accepté de revoir ce travail.

Dédicaces

A nos chers parents pour tous leurs sacrifices, amour et leur tendresse, leur soutien et leurs prières tout au long de mes études.

A mes sœurs pour leurs encouragements constants, leur soutien moral.

A mes frères pour leur soutien et leurs encouragements.

A toutes nos familles pour leur soutien tout au long de ma vie formation universitaire.

A tous nos collègues de promotion de bio-informatique, pour leurs aides et disponibilités.
A nos amis qui nous ont permis d’oublier les moments de stress et de découragement.

Tous ceux qui nous sont chers.

RÉSUMÉ

La drépanocytose ou L’anémie falciforme (AF) est une maladie autosomique récessive causée par une mutation ponctuelle du gène de la globine situé sur le chromosome 11. La mutation du codon 6 entraîne le remplacement de l'acide glutamique 6 par la valine. Les globules rouges changent leur forme d'une boule ronde à une forme de croissant lorsqu'un trouble survient dans les gènes responsables de la formation de l'hémoglobine. Afin de trouver les mutations qui se produisent dans les polymorphismes sur ce gène spécifique, notre travail a mis en évidence l'importance d'utiliser FASTA_format pour étudier et analyser l'AF, et a développé une approche bio-informatique d'apprentissage automatique pour classer la maladie et les stades sains. Le but de cette approche est de développer un modèle explicatif afin de déterminer l’existence de maladie chez les individus notamment avec l’absence des symptômes ou d'estimer son stade, à l'aide d'un classifieur automatique, de ce fait, nous avons classé les données génétiques fonctionnelles avec une précision de test de 70%.

SUMMARY

Sickle cell disease or Sickle cell anemia (FA) is an autosomal recessive disease caused by a point mutation of the globin gene located on chromosome 11. The mutation of codon 6 results in the replacement of glutamic acid 6 by valine. The red blood cells change their shape from a round ball to a crescent shape when a disorder occurs in the genes responsible for the formation of haemoglobin. To find the mutations that occur in the polymorphisms in that specific gene, our work highlighted the importance of using FASTA_format to study and analyze FA, and developed a machine learning bioinformatics approach to classify disease and healthy stages. The aim of this approach is to develop an explanatory model in order to determine the type of disease in individuals or to estimate its stage, using an automatic classifier. Therefore, we have classified the genetic data functional with a test accuracy of 70%.

الملخص

فقر الدم المنجلي (FA) هو مرض وراثي جسمي متنحي ناجم عن طفرة نقطية في جين الغلوبين الموجود على الكروموسوم 11. ينتج عن طفرة الكودون 6 حيث حدث استبدال حمض الجلوتاميك 6 بالفالين. تغير خلايا الدم الحمراء شكلها من كرة مستديرة إلى شكل هلال عندما يحدث اضطراب في الجينات المسؤولة عن تكوين الهيموغلوبين. لإيجاد الطفرات التي تحدث في تعدد الأشكال على هذا الجين المحدد ،سلط عملنا الضوء على أهمية استخدام FASTA_format لدراسة و تحليلFA ،طورنا منهجًا باستخدام التعلم الالي للمعلوماتية الحيوية لتصنيف المرض و المراحل الصحية. الهدف من هذا النهج هو تطوير نموذج توضيحي من أجل تحديد نوع المرض لدى الأفراد أو لتقدير مرحلته باستخدام مصنف أوتوماتيكي، لذلك قمنا بتصنيف البيانات الجينية الوظيفية بدقة اختبار تبلغ 70٪

LISTES DES FIGURES

LISTE DES TABLEAUX

LISTE D’ABBREVIATION

- ADN : acide désoxyribonucléique.
- A :l’adénine.
- G : la guanine.
- C :la cytosine
- T :la thymine.
- GTF :Gene Transfer Format.
- GFF :General Feature Format.
- HGP : Human Genome Project (Projet du génome humain)
- RFLP :Polymorphisme de longueur de fragments de restriction.
- SSRP : polymorphisme de la séquence répété.
- SNP : polymorphisme simple de nucléotide.
- SNPs : signal nucléotide polymorphisme.
- GWAS : Études d'association à l'échelle du génome.
- AF : l’anémie falciforme.
- HB : L’hémoglobine.
- HBS : L'hémoglobine S.
- AI : Artificial Intelligence (Intelligence artificielle).
- ANN : Artificial Neural Network (Réseau de neurones artificiels).
- ML : Machine Learning (Apprentissage Automatique)
- DP : L’apprentissage profond ou Deep Learning.
- NCBI : Le National Center for Biotechnology Information.
- VP : Vrai Positif.
- VN : Vrai Négatif.
- FP : Faux Positif.
- FN : Faux Négatif.

TABLE DES MATIÈRES

TABLE DES MATIÈRES

Introduction

INTRODUCTION

Les maladies génétiques sont des maladies causées par des anomalies dans les gènes ou les chromosomes. Certaines anomalies génétiques peuvent être transmises à la progéniture, Ces maladies génétiques sont donc des maladies familiales. Ce sont des maladies rares qui peuvent se présenter à la naissance (ou même dans l'utérus), ou parfois plus tard dans la vie. Ces anomalies génétiques peuvent être détectées en annotant le génome et en déterminant le risque de développer une maladie. Un défaut dans une ou plusieurs paires de bases d'ADN dans un gène est une variante de ce gène qui peut affecter le fonctionnement du gène. Dans ce cas, la structure chromosomique n'est pas altérée et l'anomalie n'est pas visible par l'analyse du caryotype ou d'autres tests chromosomiques. D'autres analyses génétiques doivent alors être effectuées. Certaines variantes génétiques ne posent aucun problème et d'autres ont peu de conséquences. En revanche, d'autres sont à l'origine de troubles graves comme la drépanocytose.

La drépanocytose, aussi appelée anémie falciforme, est une maladie génétique héréditaire touchant les globules rouges (ou hématies). Elle est caractérisée par une anomalie de l'hémoglobine, principale protéine du globule rouge. Elle est causée par l'hérédité homozygote des gènes de l'hémoglobine (Hb) S. Les globules rouges falciformes provoquent une vaso-occlusion et sont sujets à l'hémolyse, entraînant de graves crises de douleur, une ischémie d'organe et d'autres complications systémiques.

L'intelligence artificielle repose sur de puissants algorithmes capables d'analyser toutes les données de santé liées au patient : âge, sexe, poids, symptômes, antécédents personnels et familiaux, activité physique ou traitements antérieurs. L'utilisation d'un tel système permettrait de détecter plus tôt la maladie, de mieux identifier ses symptômes, de mieux comprendre son comportement et son évolution et d'ajuster plus précisément les traitements.

Le but de notre travail est d'apporter une contribution pratique qui permet à l'utilisateur de faire une prédiction meilleure, plus rapide et plus précise basée sur l'approche de l'intelligence artificielle qu'est l'apprentissage automatique (ML).

PARTIE 1 : RECHERCHE BIBLIOGRAPHIQUE

CHAPITRE 1 : LES MALADIES GÉNÉTIQUES

1. LE GÉNOME HUMAIN

Le matériel génétique est constitué de nucléotides liés l'un à l'autre. Cette succession de nucléotides constitue un brin d'acide désoxyribonucléique (ADN) qui est couplé à un brin complémentaire (Lambert ,2014). Chaque nucléotide est composé d'une base azotée purique, l'adénine (A) ou la guanine (G), ou pyrimidique, la cytosine (C) ou la thymine (T), d'un sucre à cinq carbones (un pentose), le désoxyribose et d'un phosphate sur le carbone 5′. On parle donc d'acide désoxyribonucléique (ADN) (Dimassi et *al.*, 2017), sous la forme d'une double hélice. La double hélice de l'ADN est enroulée autour de protéines, appelées histones, qui sont organisées et compactées, et ce compactage de l'ADN est appelé chromatine (Figure 01) (Lambert ,2014).

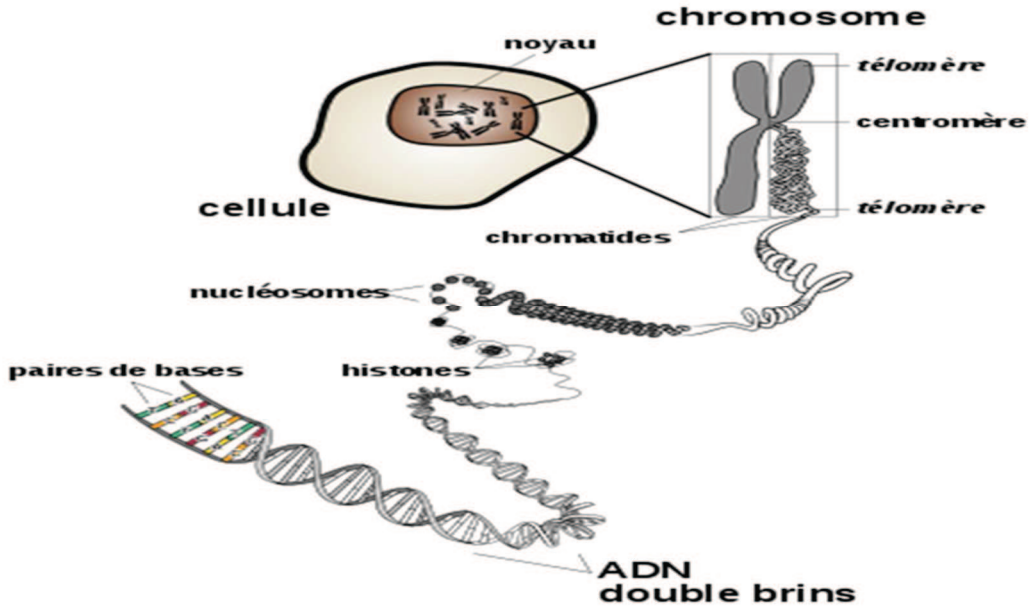

Figure 01 : Schéma montrant comment l'ADN dans nos cellules s'enroule autour des protéines structurelles (histones), qui sont empaquetées dans des nucléosomes. Les nombreux nucléosomes et l'ADN qui les entoure forment la chromatine, qui se condense pour former des chromosomes.

C'est seulement 20 ans après Mendel que les progrès de la microscopie optique ont permis de décrire les chromosomes et d'établir que chaque espèce eucaryote est munie d'un nombre spécifique de chromosomes désigné par le nombre diploïde (2n) (BOULDJADJ, 2020). Les chromosomes sont constitués d'ADN qui porte les gènes 20 000 environ (Kochko et *al*., 2000).

L'information génétique est répartie sur les 46 chromosomes (23 paires) (Ruffié, 1970). Pour chaque paire, il y a un chromosome d'origine paternelle et un chromosome

d'origine maternelle. Pour une même paire, les deux chromosomes ne seront donc pas identiques (Seitz, 2022).

Les 22 premières paires sont appelées autosomes. La 23ème paire est celle qui détermine le sexe de la personne. Il s'agit des chromosomes X et Y. Les femmes possèdent deux chromosomes X, alors que les hommes possèdent un chromosome X et un chromosome Y (figure 02) (Dubois, 1988). Les chromosomes sexuels structurellement distincts (X et Y) sont le mode le plus familier de détermination génétique du sexe et ont évolué indépendamment dans de nombreux taxons différents (Miller, 1991).

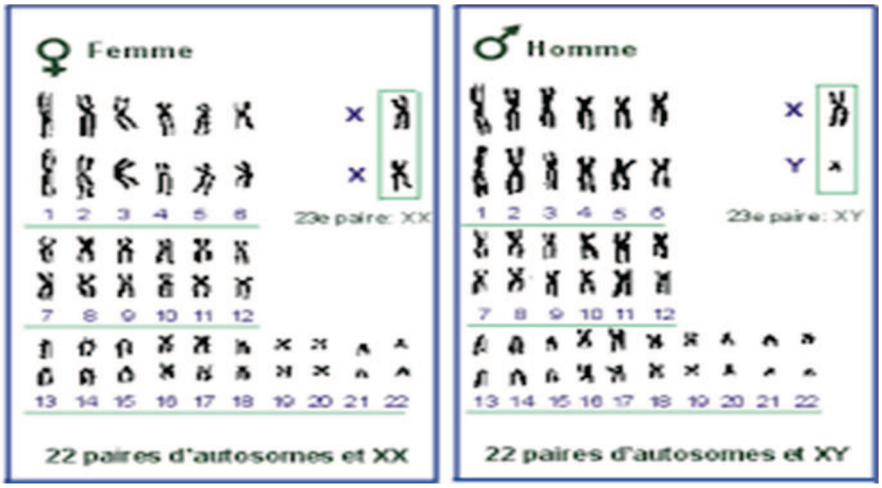

Figure 02 : L'ensemble des 23 paires de chromosomes du génome humain[1].

On peut distinguer les différents chromosomes selon deux critères (Gallais et Bannerot , 1992)

1. Chromosomes sont variables en taille : ainsi chez l'humain très grande différence entre chromosomes 1 et 21 (3 à la fois plus grands) (Larbi, 2021).

2. L'indice centromérique

Ces critères de classification permettent de distinguer 7 groupes de chromosomes (tableau 01) (Boudiaf, 2018) :

Tableau 01 : Classification des groupes de chromosomes.(Lesse, 1970).

Groupes	Chromosomes
A	Chromosomes 1,2 et 3.
B	Chromosomes 4 et 5.
B	Chromosomes 6 à 12 + le ou les chromosomes X dont la taille et voisine de celle d'un chromosome 6.
D	Chromosomes 13,14 et 15 ou chromosomes acrocentriques.
E	Chromosomes 16,17 et 18.
F	Chromosome 19 et 20.
G	Chromosomes 21 et 22, on adjoint le chromosome Y.

[1] (https://babel.csfoy.ca/profs/gbourbonnais/biotlm/genetiquetlm/notgenet3.htm).

Chez l'individu normal la formule chromosomique s'écria 46, XX XY (figure 03) :

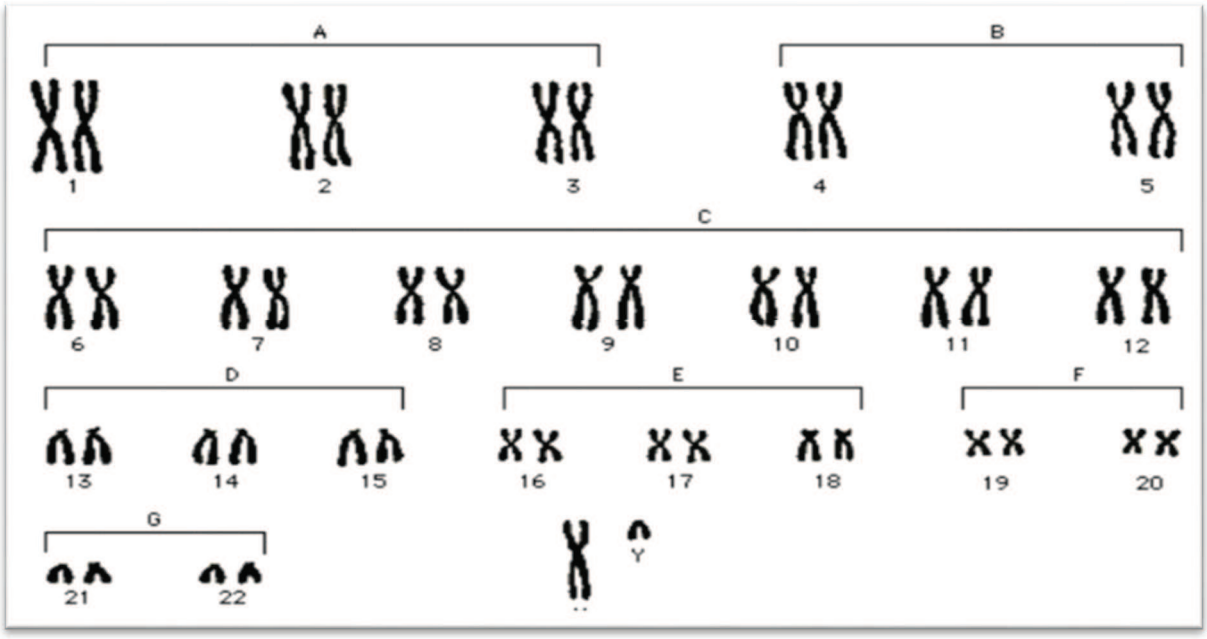

Figure 03 : caryotype humain normal (Hamerton,2013).

Le projet génétique (The Human Gnome Project) vise à séquencer la séquence de nucléotides qui composent l'ADN (Dheur et Saupe, 2021). Il a été discuté dès 1985. Le premier programme conçu en 1990 (figure 04). (Paslier et Bernot, 2001) qui nous a permis d'identifier et de caractériser les gènes qui interviennent dans plusieurs maladies génétiques. (Sfar et Chouchane, 2008). Il a un objectif du séquençage du génome entier. Afin de sélectionner également tous les gènes leur localisation chromosomique. (Guellaën et Andrologie, 1999).

Il devait permettre de résoudre deux problématiques (Mélançon et Lambert, 1992):

1. Avoir la séquence la plus complète possible du génome humain, représentant plus de 3 milliards de positions, réparties dans 22 paires de chromosomes "autosomaux" ainsi qu'une paire de chromosomes sexuels.

2. Localiser et définir les unités fonctionnelles de ce génome, et plus particulièrement les gènes.

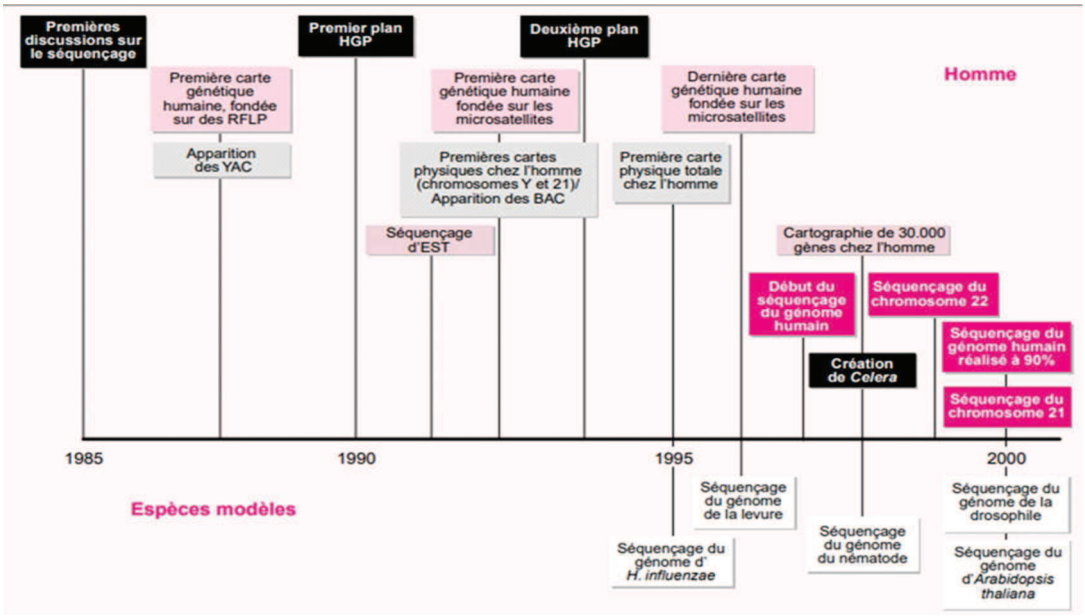

Figure 04 : Principales étapes du projet génome humain, et des projets des espèces modèles[2].

2. ANALYSER LA SÉQUENCE NUCLÉOTIDIQUE

Annotation du séquençage génétique est nécessaire pour identifier les gènes d'un organisme, c'est-à-dire pour trouver leur emplacement exact dans la séquence du génome et assigner une ou plusieurs fonctions biologiques à chacun de ces gènes hypothétiques (Figure 04) (Touchon et *al.*, 2009). L'identification des éléments fonctionnels le long de la séquence d'un génome, lui donnant ainsi un sens. L'annotation est nécessaire car le séquençage de l'ADN produit des séquences de fonction inconnue ((Cohen et al, 2019).

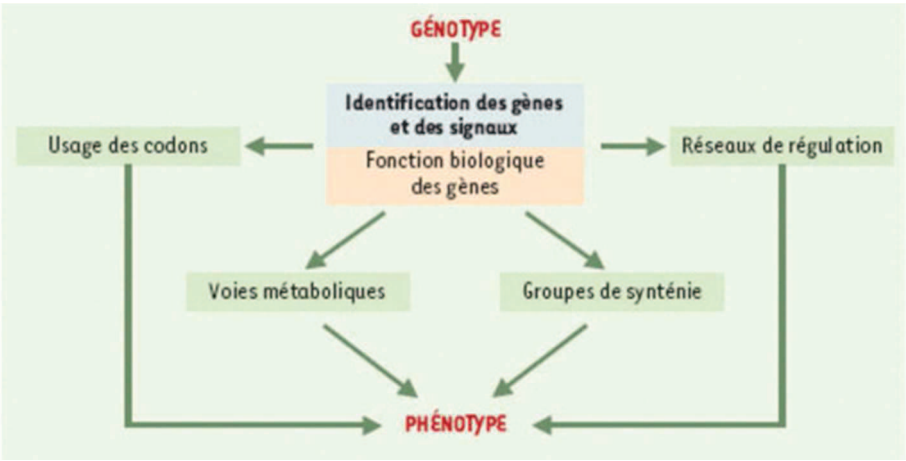

Figure05 : Stratégies d'annotation in silico des génomes.[3]

[2] https://www.researchgate.net/figure/Principales-etapes-du-projet-genome-humain-9_fig5_331149640

[3] https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/genome-annotation

L'annotation génomique représente les connaissances actuelles d'un génome sur laquelle les analyses bio-informatiques se basent. Il est donc essentiel qu'elle soit à jour, et organisée selon une structure fixe et explicite afin d'être reconnue par les divers programmes informatiques. Aujourd'hui, les annotations se conforment aux formats GTF (Gene Transfer Format) ou GFF (General Feature Format) qui sont agencés un peu différemment, mais contiennent la même information sur la structure des gènes (Bucci et *al.*, 2016).

2.1 Les différents niveaux d'annotation des génomes :

Il existe trois niveaux de complexité d'annotation du génome (Médigue, 2002 et Vincent, 2009).

- L'annotation syntaxique : C'est la recherche de gènes au sens large, c'est-à-dire consiste à prédire et localiser l'ensemble des séquences codantes ou gènes du génome et à identifier leur structure, leur fonction ainsi que les relations entre les entités biologiques relatives au génome (Gaudriault et Vincent,2009).

- L'annotation fonctionnelle : Est une tâche qui dépend principalement à des informations qui peuvent être associées à la protéine à annoter, et elle est réalisée manuellement (Souciet *et al.,* 2009).

- L'annotation relationnelle : regroupe différentes relations établies entre les séquences pour décrire des objectifs ou modules biologiques par exemple une voie métabolique (Beroud, 2010).

3. LES MUTATIONS ET LES POLYMORPHISMES GÉNÉTIQUES

3.1 Les mutations

Les mutations sont des changements dans la séquence d'ADN d'un organisme. Elles sont soit dû à des erreurs de transcription sont appelées mutations spontanées ou lors des divisions, soit causées par des agents mutagènes physiques ou chimiques sont des mutations induites (CHEIKH, 2020).

3.1.1 Les différents types de mutations

La plupart des mutations se limitent à un seul nucléotide ou à quelques-uns seulement, situés les uns à côté les autres sur le même gène. C'est ce qu'on appelle des mutations ponctuelles (Merlin, 2013) (figure06).

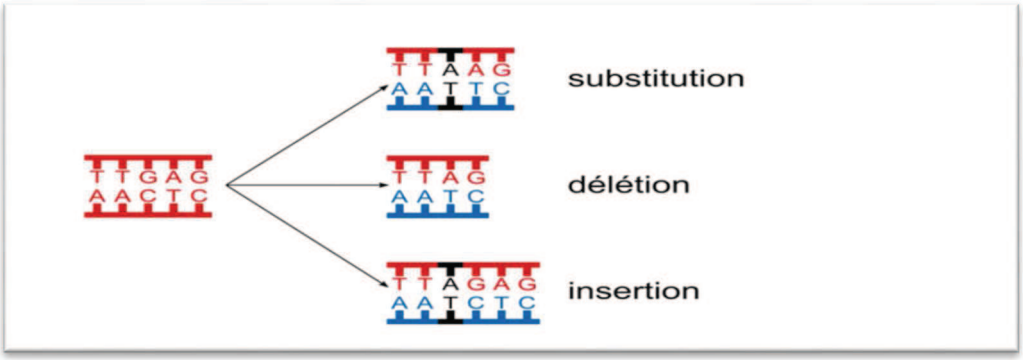

Figure 06 : 3 mécanismes de mutations peuvent affecter l'ADN.[4]

• Substitution : Dans cette mutation, l'anomalie se produit en remplaçant un nucléotide par un autre ou, dans notre exemple, par une lettre avec autre.
• Délétion : perte d'un ou plusieurs nucléotides.
• Addition : l'insertion d'un ou plusieurs nucléotides.
(Aguedach *et al.*, 2005).

Les délétions/insertions résultent un changement de cadre et conduisent généralement à une protéine incomplète qui se décompose rapidement.

Les mutations ponctuelles se distinguent selon leurs conséquences sur les protéines (tableau 02) (Sheikh et Halder, 2019).

Tableau 02 : Conséquences au niveau de la protéine (Saidi, 2020).

Mutation silencieuse	Le codon muté code pour le même acide aminé (AA) que le codon sauvage (normal). La protéine est normale/Fonction normale.
Mutation faux sens	Le codon spécifie un (AA) fonctionnellement Équivalent (Arg (AGA), lysine (AAA). La fonction peut être normale tout dépend de sa localisation.
Mutation frameshif	Mutations avec changement dans le cadre de lecture, protéine complètement différente. Pas de fonction.

3.2 Les polymorphismes

Un polymorphisme est une mutation et peut se situer en région codante ou non codante. La notion de polymorphisme repose à la fois sur le caractère non autogène de ces variations et la fréquence (> 1% dans la population) (CHEIKH. 2020).

3.2.1 Types de marqueurs

Il y a plusieurs types de polymorphisme, parmi lesquels :

• Polymorphisme de longueur de fragments de restriction ou RFLP.

[4] https://www.assistancescolaire.com/eleve/1re/sciences-de-la-vie-et-de-la-terre/reviser-le-cours/1_t_03/print?print=1&printSheet=1

• Polymorphisme de la séquence répétée ou SSRP.

• Polymorphisme d'insertion et de répétition d'un nombre variable de copie sou CNVs.

• Polymorphisme simple de nucléotide ou SNP.

Le séquençage du génome humain est à l'origine de la découverte de millions de variations de séquences dans le génome humain. (MORVAN, 2005).

La variation génétique se produit au sein et entre différentes populations et contribue à la variation des traits, y compris la susceptibilité aux maladies, entre les individus. La forme la plus fréquente de variation génétique chez l'homme sont : Les polymorphismes nucléotidiques (single nucléotide polymorphismes, SNP). (Korzeniewski, 2013).

3.2.2 Les SNPs (ou Single Nucléotide Polymorphisme)

Les SNPs sont la forme la plus courante de variation génétique De la séquence qui se distingue par une seule différence de nucléotide (Jurinke, 2022) par exemple, les deux séquences ADN de deux individus, AAGCCTA et AAGCTTA auront pour seule différence le nucléotide en cinquième position C ou T. Pour ces individus. Il n'y a alors que deux allèles possibles C ou T et trois génotypes CC, CT ou TT seule différence le nucléotide en cinquième position C ou T.

Les snp peuvent être trouvés dans des régions de gènes non codantes (introns), des régions codant pour des gènes exons ou des régions intergéniques (figure 07) (Dr SEMMAME.O).

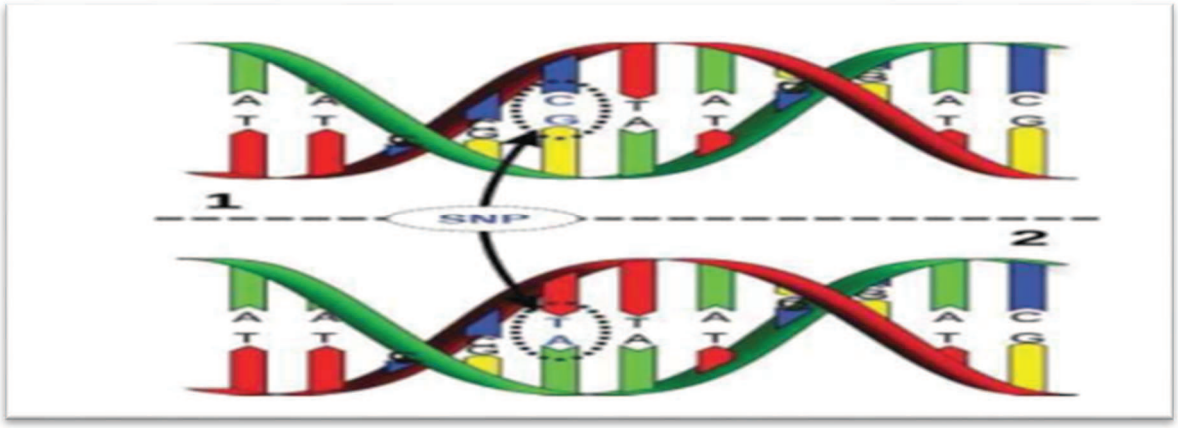

Figure 07 : Exemple d'un SNP. Ici, les molécules d'ADN 1 et 2 diffèrent à un locus donné d'une seule paire de base.[5]

L'accès aux bases de données de SNP ouvre la possibilité d'étudier l'impact de ces polymorphismes sur le risque de maladie ainsi que sur les réponses aux médicaments. (Morvan, 2005).

Les données sur le polymorphisme sont importantes dans la susceptibilité à certaines maladies. La base de données de référence concernant le polymorphisme chez l'homme est la base dbSNP, le site de cette base de données (Guillocheau, 2018).

[5] https://www.police-scientifique.com/adn/evolutions

3.2.3 Études d'association à l'échelle du génome (GWAS)

L'approche GWAS est une bonne stratégie pour détecter les variants génétiques récurrents (Pierre et *al*., 2016). Elle est conçue pour détecter des groupes de SNP associés à la maladie (Berrou, 2020). Elle consiste à comparer la fréquence de centaines de milliers de variantes génétiques répartis sur tous les chromosomes entre un groupe de cas malades et un groupe de témoins, à l'aide de techniques de génotypage à haut débit (Duvoux *et al*. , 2012).

4. MÉDECINE PRÉDICTIVE

Le terme « médecine prédictive » a été créé en France pour désigner un éventail d'innovations dans les technologies de prédiction médicale. C'est un système développé récemment qui vise à indiquer le plus précisément possible le risque de développer certaines maladies grâce à la génétique (séquençage du génome) (verdier, 1996).

L'objectif de la médecine prédictive est d'identifier les humains à risque de développer une maladie associée à une anomalie génétique, mais aussi de savoir quel risque des futurs parents qui peuvent transmettre un trait à leur progéniture. Il peut également prédire, par exemple, votre risque de développer la drépanocytose (Montgolfier, 2015).

La médecine prédictive est définie de deux manières différentes. Une définition large met l'accent sur sa méthode, à savoir l'utilisation de tests génétiques. Une définition plus stricte insiste sur les maladies qui constituent son champ d'investigation spécifique. Selon ce type de définition, dans le domaine de la recherche ou de la clinique, la médecine prédictive a pour objet d'étudier les maladies dites multifactorielles, pour lesquelles plusieurs gènes ainsi que des déterminants environnementaux constituent des facteurs de risque (Dekeuwer, 2020).

Le généticien Victor Makusik a décrit aux médecins praticiens le fonctionnement de la médecine clinique prédictive en ce sens que les diagnostics sont construits à partir de signatures génétiques personnalisées (polymorphismes mononucléotidiques, haplotypes et génotypes) qui se concentrent sur la capacité de prédiagnostiquer la maladie. Certains voient une médecine préventive fondée sur une approche rationnelle du risque (Aymé, 2001).

5. MALADIES GÉNÉTIQUES ET HÉRÉDITAIRES

La génétique moléculaire, après le succès du traitement des maladies héréditaires monogéniques, s'est donné pour objectif d'identifier les déterminants génétiques des maladies récurrentes à forte composante génétique (Rousseau, 2003).

Les maladies héréditaires sont des conditions médicales qui peuvent être transmises d'une génération à l'autre. Elles peuvent se manifester de différentes manières et affecter les personnes de tous les âges, allant des nouveau-nés aux adultes. Dans ce cas nous découvrirons certains d'exemple les plus courants de maladies héréditaires, et apprendrons pourquoi l'anémie falciforme est souvent caractérisée comme une maladie héréditaire. Finalement, nous verrons comment ces maladies sont diagnostiquées. Dans plusieurs maladies héréditaires, les globules rouges deviennent sphériques, ovoïdes ou en forme de faucille (drépanocytose) (Catonné et *al*.,).

Drépanocytose, également connue sous le nom d'anémie falciforme (AF) est une maladie héréditaire chronique qui touche l'hémoglobine, affectant l'homme c'est-à-dire

transmise par les parents et présente dès la naissance. Cela veut dire que les personnes qui en sont atteintes ont hérité d'un gène de l'hémoglobine S de chaque parent. L'anémie falciforme est aussi une maladie chronique, donc une maladie qui dure toute la vie, même si elle est traitée (Carolyn, 2022).

5.1 Drépanocytose

En 1904, James HERRICK, médecin de Chicago, fait la première description médicale de la drépanocytose (Medkour 2008). Le mot drépanocytose vient du mot grec drepanon, qui signifie faucille. Cette maladie se caractérise par une forme anormale des globules rouges (ou globules rouges) (Lainé , 2009), provoquant la drépanocytose (figure 08). Les globules rouges déformés empêchent le transport efficace de l'oxygène vers les organes du corps. Une mauvaise circulation est à l'origine des crises douloureuses caractéristiques de la drépanocytose. Cependant, la maladie est plus fréquente chez les personnes originaires d'Afrique, de la Méditerranée, des Caraïbes, du Moyen-Orient, de certaines parties de l'Inde et de l'Amérique du Sud. Près de 120 millions de personnes seraient porteuses d'une mutation drépanocytaire et 50 millions seraient touchées par cette maladie (Lemogoum, 2008).

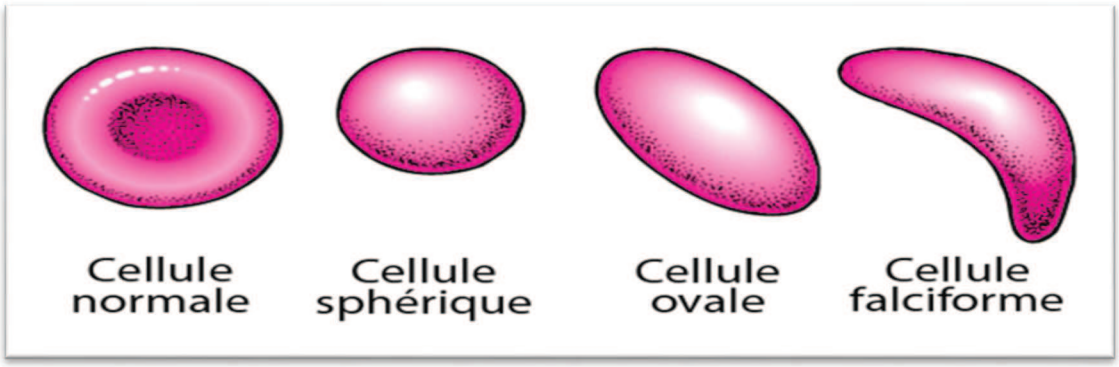

Figure 08 : Forme des globules rouges (Maura, 2022).

L'hémoglobine (HB) est un complexe protéique constitué de quatre protéines agencées entre elles (Cobessi *et al.* ,2010) :

• Deux sous-unités d'alpha-globine.

• Deux de bêta-globine.

5.1.1. Pathologie

Dossiers des patients suivis pour drépanocytose et ayant consulté en gastro-entérologie durant la période de janvier 2015 à décembre 2017. Collecte des données et prise en compte de différents critères d'étude dont: âge, sexe, motif de consultation, diagnostic, type d'atteinte vasculaire crise obstructive, et des examens cliniques ont été réalisés. Le résultat présent dans les 2 tableaux suivants (Banza ,2019) :

Tableau 03 : répartition des cas selon les motifs de consultation dans l'Afrique du Sud (Banza, 2019).

Motif de consultation	Effectif	Pourcentage (%)
Aucune plainte	90	43.68
Douleur abdominale	92	44.66
Fièvre	125	60.67
Troubles digestifs	62	30.09
Ballonnement abdominal	10	4.85
Hématémèse	9	4.36

Résultat 01 : L'étude de (Banza ,2019) a révélé une incidence plus élevée de maladies gastro-intestinales chez les patients atteints d'anémie falciforme, et il s'agit donc d'une maladie gastro-intestinale.

Tableau 04 : répartition des cas en fonction de l'âge et du sexe dans l'Afrique du Sud.(Banza, 2019).

Variables	Effectif	Pourcentage (%)
Tranche d'âge (ans)		
1-6	67	32.5
7-11	44	21.4
12-16	43	20.9
17-21	21	10.2
22-26	20	9.7
27-31	7	3.4
32-36	3	1.5
37-21	1	0.5
Sexe		
Masculin	99	48.1
Féminin	107	51.9

Résultat 02 : La tranche d'âge la plus touchée est celle de 1 à 6 ans dans laquelle le sexe féminin prédomine légèrement sur le sexe masculin.

5.1.2 Mutations génétiques dans la drépanocytose

La drépanocytose est due à une mutation unique et ponctuelle dans l'ADN du gène codant pour la bêta-globine, situé sur le chromosome 11 (Couque et Montalembert, 2013). Il lui confère une structure altérée qui permet à l'hémoglobine de former des chaînes (polymères) lorsque la concentration en oxygène dans le sang est faible (hypoxie) (Rodwell, 2002). Il déforme les globules rouges et leur donne la forme caractéristique de faucille. L'hémoglobine chez les personnes atteintes de drépanocytose est appelée hémoglobine S (Diop et Fall, 2022).

L'hémoglobine S résulte d'une mutation A en T du sixième codon de la région codant pour la β-globine. De nombreux changements dans la structure de l'hémoglobine sont dus à des mutations dans la population humaine qui impliquent souvent la substitution d'un acide aminé par un autre (figure 09) (Mojtahedi et *al.*, 2008).

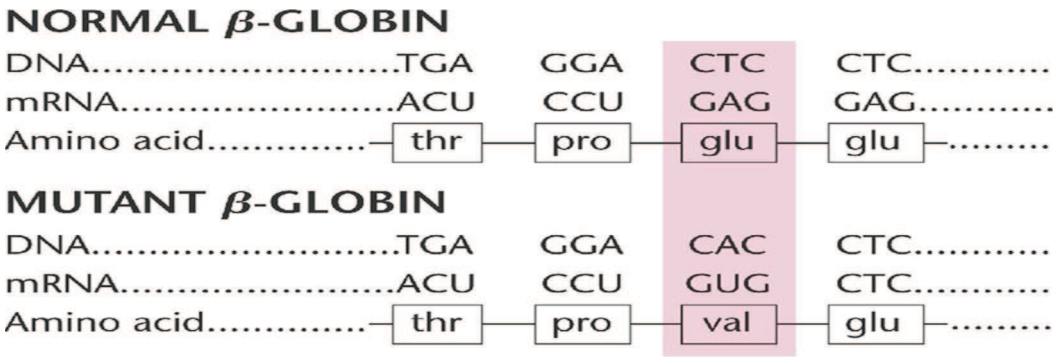

Figure 09 : Figure montrant un seul changement de nucléotide dans l'ADN codant pour la β-globine.[6]

6 http://sgugenetics.pbworks.com/w/page/61172304/Pathophysiology%20of%20Sickle%20Cell%20Anemia

CHAPITRE 2 : CONCEPTS D'INTELLIGENCE ARTIFICIELLE

1. INTELLIGENCE ARTIFICIELL (IA)

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 (Pastre, 2000).

Depuis longtemps, les être humains se sont intéressés à créer des machines capables de simuler la pensée humaine. Le terme « intelligence artificielle » a été créé en 1955 par John McCarthy (Perrot, 2019). L'émergence de l'intelligence artificielle (IA), dans le domaine médical est la conséquence de trois bouleversements radicaux : la numérisation des images médicales, le développement des algorithmes l'utilisation des données (Brunelle et Brunelle, 2019).

L'intelligence artificielle (IA) est le processus d'imitation de l'intelligence humaine (Pélissier, 2020). Elle est une branche de l'informatique dont le but est de réaliser des systèmes intégrants un grand nombre de connaissances et de traitements dans le monde du travail de l'être humain et ses applications quotidiennes, (dits systèmes intelligents) (Gabriel, 2019).

Les principaux thèmes d'étude de l'intelligence artificielle sont l'acquisition et la représentation des connaissances sous toutes leurs formes (Laurière, 1987). Le développement des dispositifs d'intelligence artificielle vise à amplifier les capacités cognitives et de réflexion des être humains et des scientifiques qui tentent de comprendre les mécanismes qui régissent les processus cognitifs des êtres vivants. L'IA permet aussi à des systèmes techniques de percevoir leur environnement, gérer ces perceptions, résoudre des problèmes et entreprendre des actions (Haiech, 2020).

L'IA est divisée en nombreuses sous-disciplines qui essaient chacune de traiter une partie du problème. Les principales sous-discplines de l'IA figurent dans (figure 10).

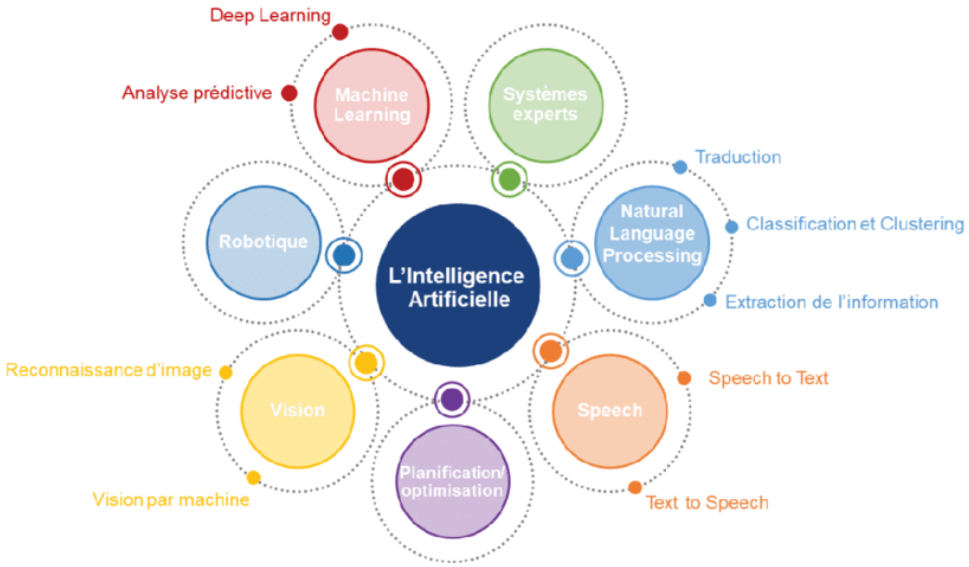

Figure 10 : Schéma de décomposition du domaine de l'intelligence artificielle (Artik Consulting, 2018).

2. APPRENTISSAGE AUTOMATIQUE

L'apprentissage automatique, ou aussi appelé machine learning (ML), est une branche de l'intelligence artificielle qui fait référence au développement de méthodes d'analyse et de mise en œuvre permettant à une machine d'évoluer à travers un processus d'apprentissage (Bonneuil, 2006). Il vise à développer des algorithmes basés sur l'apprentissage des données pour effectuer des analyses automatiques en détectant des modèles et des règles d'association (N'Guessan, 2020).

Le ML est idéal pour exploiter les opportunités cachées du Big Data. Cette technologie permet d'extraire de la valeur en provenance de sources de données massives et variées sans avoir besoin de compter sur un humain (Cadavid et *al*., 2021).

L'apprentissage automatique est également utilisé pour la traduction automatique des langues et pour convertir la parole prononcée à l'écran (parole en texte). Un autre cas d'utilisation est l'analyse des sentiments sur les réseaux sociaux, qui est également basée sur le traitement du langage naturel (NLP) (Calistru, 2022).

La plupart des données collectées par les entreprises ne sont pas structurées. Ces données ne correspondent pas à un modèle de données existant, comme les données structurées voire même semi-structurées. Pour de nombreuses entreprises, ces données non structurées sont inutiles, dans une certaine mesure (Labbe, 2019).

2.1 Les types d'apprentissage automatique

Les algorithmes d'apprentissage peuvent recentre le propos sur l'apprentissage automatique en différenciant apprentissage supervisé, apprentissage non-supervisé et apprentissage par renforcement (AR). (figure 11) :

Figure 11 : Différentes types pour le ML (Mishra, 2022).

2.1.1 L'apprentissage supervisé

Le terme d'apprentissage supervisé vient de l'idée qu'un algorithme apprend à partir d'un ensemble de données d'initiale, qui peut être considéré comme l'enseignant (Eensoo et Nouvel, 2016).

Dans l'apprentissage supervisé, on a deux types algorithmes :

Les algorithmes de régression et les algorithmes de classification (Ouladbrahim, 2022).

-La régression : est un type de méthode d'apprentissage supervisé qui utilise un algorithme pour comprendre la relation entre les variables dépendantes et indépendantes (Rakotomalala, 2011).

-Classification : Les problèmes de classification utilisent un algorithme pour affecter avec précision des données de test à des catégories spécifiques. Cet algorithme est très facile à comprendre, et fait partie des méthodes d'apprentissage supervisé. C'est-à-dire que les prédictions sont réalisées à partir de données historiques (Tanous, 2002).

2.1.2 L'apprentissage non supervisée

L'apprentissage non supervisé est utilisé pour réaliser une tâche importante et difficile en apprentissage. Ce processus intervient dans des contextes variés tels que la découverte de connaissances, analyser et regrouper des ensembles de données non étiquetées ou la description d'un ensemble de données sans nécessiter d'intervention humaine (Cleuziou, 2004).

L'utilisation de l'apprentissage non supervisé peut être réunie en problèmes de clustering et d'association (Goncharuk , 2004).

- Le clustering : est une procédure d'analyse multivariée. Il a été créé pour explorer la structure da nature inhérente des données, où les données d'un même ensemble sont aussi similaires que possible (Cleuziou, 2004).

La différence entre l'apprentissage supervisé et l'apprentissage non supervisé est montrée dans (tableau 05) :

Tableau 05 : Déférence entre deux types d'apprentissage automatique.(Duclayen et, Yvon, 2003).

	L'apprentissage supervisé	L'apprentissage non supervisé
Donnés d'entrée	Donnés connues enentrée	Donnés inconnues en entrée
Complexité informatique	Complexité	Moins complexité
Domaines d'activités	Classification et régression	Clustering et d'association
Précision	Produit des résultats précis.	Généré des résultats modérés.

3. APPRENTISSAGE APPROFONDI

L'apprentissage profond ou Deeplearning (DL) a révolutionné l'intelligence artificielle et il est très rapidement répandu dans de nombreux domaines d'activité (Ilyushchanka, 2021). Il est des principales techniques d'apprentissage automatique (figure 12). Il permet de traiter

les mégas données à l'aide d'un réseau de neurones artificiels (Artificial Neural Network : ANN) inspiré du réseau de neurones du cerveau humain. (Charlin, 2017).

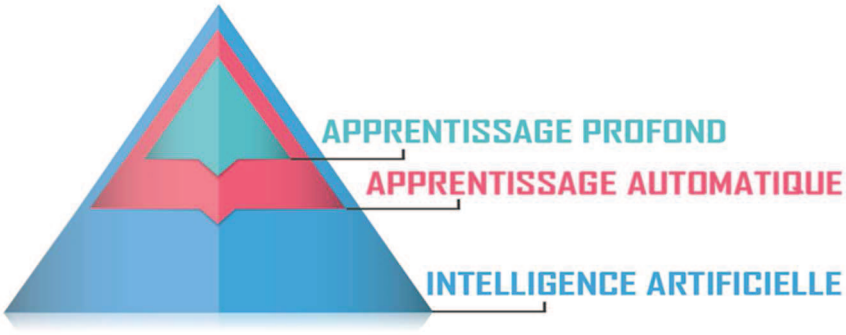

Figure 12 : la relation entre IA et ML et Deep Learning.[12]

3.1 Réseau de neurones artificiels

Les modèles de réseaux de neurones artificiels existent depuis longtemps (JRynkie, 2022) constitué de neurones artificiels. Ce réseau de neurones forme une application permettant d'aborder sous de nouveaux angles les problèmes de cognition, de mémoire, d'apprentissage et de raisonnement (Parizeau, 2004). Ce réseau est spécialisé dans le traitement, la réception et la transmission d'informations électrochimiques à d'autres neurones du cerveau central ou périphérique. Un neurone biologique est une cellule complexe, mais seule sa fonction de base a été une source d'inspiration pour les neurones computationnels (Clement et Russo, 2021).

Il se compose généralement de trois types de couches : couche d'entrée dans laquelle les unités représentent les champs d'entrée et une ou plusieurs couches cachées et la couche de sortie dans laquelle des unités représentent les champs cibles (figure 13).

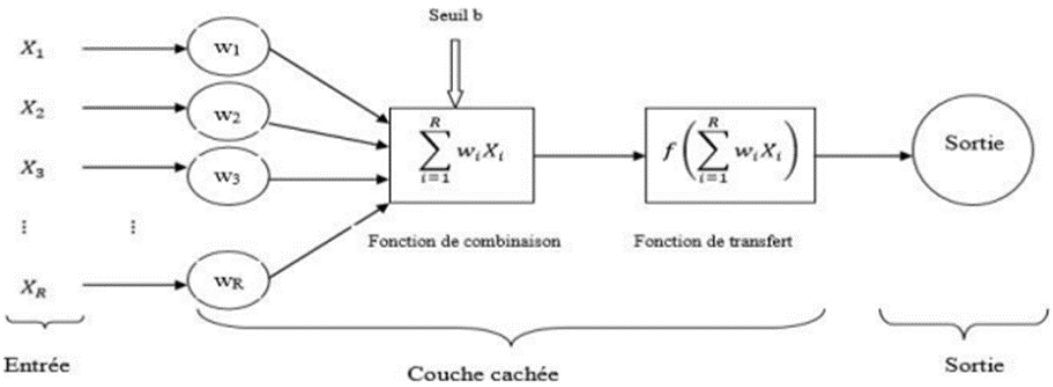

Figure 13 : Conception d'un Réseau de Neurones (Thévenet *et al.,* 2012).

[1] https://www.researchgate.net/figure/Fonctions-dactivation-dun-neurone-artificiel_fig2_322194356

[2] https://knowledgeone.ca/mini-glossaire-de-lintelligence-artificielle/?lang=fr

La fonction de transfert est en général, une fonction non linéaire, les fonctions de transfert sont de qualités diverses : elles peuvent être déterministes, continues, discontinues ou aléatoires (figure 14).

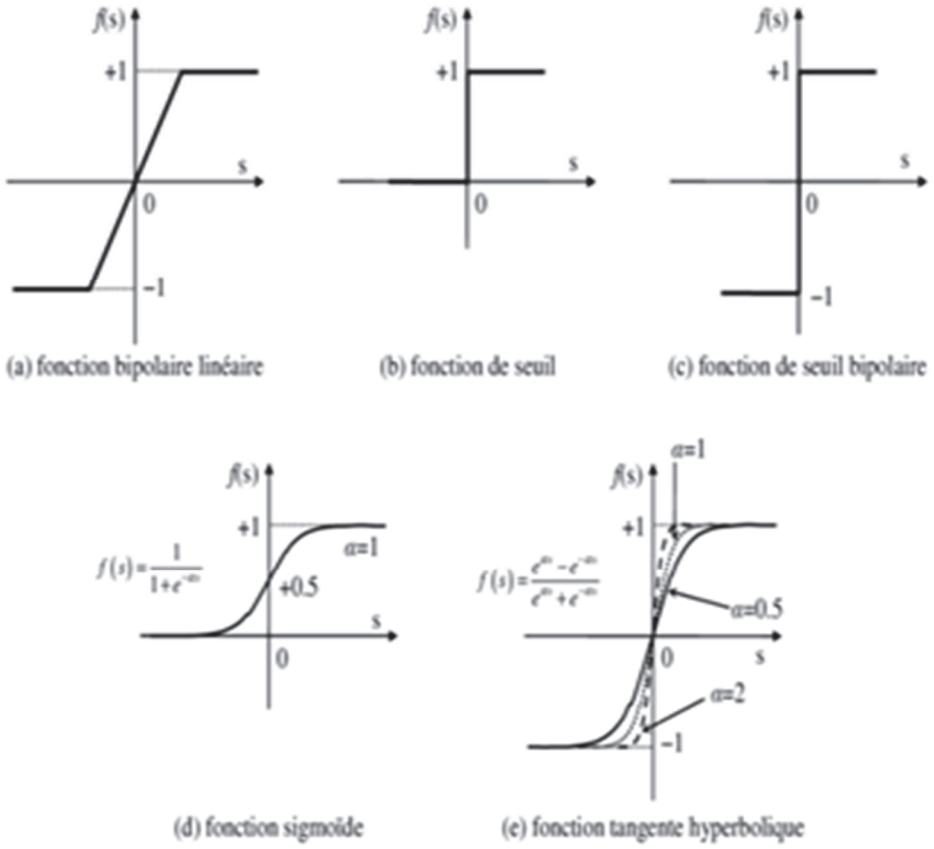

Figure 14 : Les modèles de fonctions d'activation.[3]

3.2 Le travail du réseau de neurones artificiels

Il simule le travail de nos propres neurones (Figure15) (Clément et Rousseau, 2021).

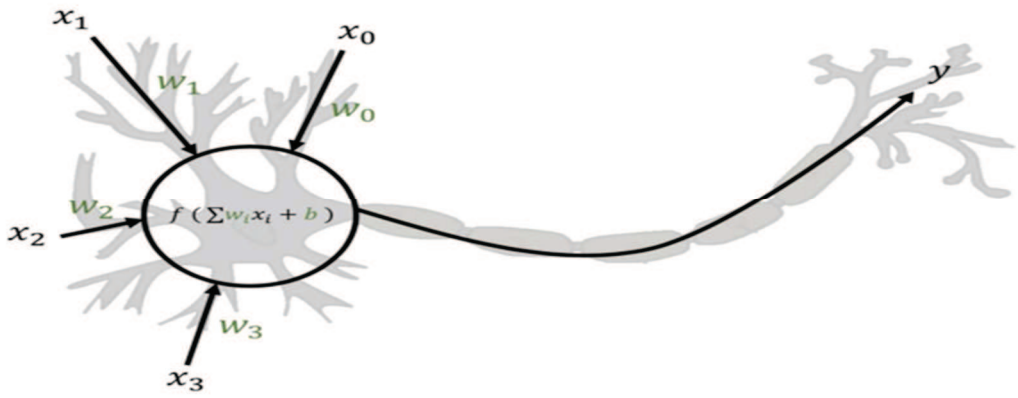

Figure 15 : Schéma d'un neurone informatique superposé à un schéma de neurone biologique.(Cadavid, 2021).

[3] https://www.researchgate.net/figure/Fonctions-dactivation-dun-neurone-artificiel_fig2_322194356

• (X0, x1, x2 et x3) : le neurone s'active à son tour en envoyant un signal aux neurones en aval. Ces signaux s'accumulent dans le corps du neurone. Ces signaux on appelle des signaux électriques.

•$f\ \sum\ (Wi\ xi + b)$: représente les signaux du neurone en amont et est émis en tant que valeur réelle de la sortie (y). Cette sortie est calculée à partir de l'entrée via cette équation.

•Wi : qui sont appelés poids, L'entrée sont multipliées par ces valeurs. Cela représente la force de la relation entre ces neurones et les neurones en amont.

•F : La fonction f est la fonction dite d'activation. C'est une fonction non linéaire croissante. F renvoie 1 si son argument est strictement positif et 0 sinon.

•B : que l'on appelle le biais, représente l'appétence ou la résistance du neurone à s'activer.

Les poids et le biais ce sont ces valeurs qui sont modifiées au cours d'un entraînement.

4. PROCESSUS DU ML/DL

Les étapes successives du ML/DL suivant :

4.1. Collection des données

Il s'agit de regrouper les données d'un problème à résoudre sous un format adéquat (construction d'une Dataset).

4.2. Prétraitement des données

Afin de rendre Dataset utilisable à l'apprentissage, il faut la nettoyer (suppression de données inutiles, répétées, incomplètes et manquantes, enrichissement par d'autres données, décomposition des données).

4.3. Choix du modèle

Selon le problème traité, on peut choisir :

- ✓ la régression : s'il s'agit d'un problème de prédiction,
- ✓ le Clustering (K-means ou le voisin K) : pour les problèmes tels que la détection d'anomalies, la classification d'images, etc.
- ✓ naïve bayes : s'il s'agit un problème de classification.

4.4. Entraînement du modèle

Les données de Dataset sont séparées en :

- 80 % pour entraîner l'algorithme choisi,

- 20 % pour tester et vérifier la performance du résultat

4.5. Évaluation

Les principales façons de vérifier le rendement du modèle d'apprentissage sont les suivantes (Chudasama, 2021) :

1. Matrice de confusion : pour vérifier les prédictions de classification où les résultats pourraient être :

- ✓ Le vrai positif (VP) : une sortie prédite qui appartienne à une classe et qui appartienne réellement à cette classe.
- ✓ Le vrai négatif (VN) : une sortie prédite qui n'appartienne pas à une classe et qui n'appartienne pas réellement à cette classe.
- ✓ Le faux positif (FP) : une sortie prédite qui appartienne à une classe et qui n'appartienne pas réellement à cette classe.
- ✓ Le faux négatif (FN) : une sortie prédite qui n'appartienne pas à une classe et qui appartienne réellement à cette classe.

2. Accuracy : c'est la valeur de la somme VP et VN divisée la somme des valeurs VP, VN, FP et FN combinées (Priyadarshini et Cotton, 2021)
3. Précision : c'est la valeur VP divisée la somme des valeurs VP et FP combinées (Priyadarshini et Cotton, 2021)
4. Sensibilité : c'est la valeur VP divisée la somme des valeurs VP et FN combinées (Priyadarshini et Cotton, 2021)
5. Spécificité : c'est la valeur VN divisée la somme des valeurs VN et FN combinées (Priyadarshini et Cotton, 2021)

PARTIE 2 : MATÉRIEL ET MÉTHODES

1. MATÉRIEL

1.1 Collection des données

Les données utilisées pour la suite de ce travail sont extraites à partir de la base NCBI (Le National Center for Biotechnology Information). L'objectif de ce travail est de développer de nouvelles technologies de l'information pour aider à comprendre les processus génétiques sous-jacents qui contrôlent la drépanocytose.

Dans ce travail, quelques séquences sont enregistrées sous format FASTA et d'autre sous format texte. Le format faste FASTA est un format de fichier texte utilisé pour stocker des séquences biologiques de nature (nucléaire). Afin de construire le dataset, nous avons collecté 133 séquences. Ces séquences sont organisées dans deux dossiers.

Tableau 05 : Description du contenu des dossiers

Nom du dossier	Description	Taille
Malade	Ce dossier contient les séquences nucléotides des personnes malades. Le nombre des séquences est 92. Chaque séquence est enregistrée dans un fichier soit sous format FASTA ou texte.	2.139MO
Saine	Ce dossier contient les séquences nucléotides des personnes qui ne sont pas malades. Le nombre des séquences est 41. Chaque séquence est enregistrée dans un fichier soit sous format FASTA ou texte.	1.396MO

1.2. Configuration de la machine

Les caractéristiques de la machine exploitées sont détaillées dans le tableau suivant :

Tableau 06 : Caractéristiques de la machine utilisée pour le ML.

Ordinateur	Caractéristiques
Processeur	Intel(R) Core (TM) i7-6500 UCPU@3.40GHz 2.6GHz
Mémoire installée RAM	4.00 Go
Stockage	Western Digital Blue Desktop 1 To SATA 6Gb/s 64 Mo
Système d'exploitation	Windows 10 professionnel
Type de système	Système d'exploitation 64 bits

1.3. Outils et bibliothèques

Nous décrivons brièvement les outils et bibliothèques utilisés pour effectuer ce travail dans cette section :

1.3.1. Outils

Ce travail a été réalisé avec le langage de programmation Python, via le notebook Jupyter d'Anaconda (Tableau 07).

Tableau 7 : Principaux outils utilisés

Outil	Description
Python 3.11.4	Python est le langage de programmation open source le plus utilisé par les informaticiens. Elle a propulsé ce langage au premier plan dans la gestion des infrastructures, l'analyse de données ou le développement de logiciels.
Anaconda 2.3.1	Anaconda est un outil distribué gratuit et open source. Il est destiné à la programmation dans l'environnement Python et R. Anaconda est largement utilisé dans la science des données, l'intelligence artificielle ou l'apprentissage automatique. Cette distribution scientifique de Python contient plusieurs packages nécessaires à l'analyse des données.
Jupyter notebook 6.3	Jupyter Notebook est une application Web open source que peut être utilisée pour créer et partager des documents contenant du code en direct, des équations, des visualisations et du texte. Jupyter Notebook est maintenu par les personnes de Project Jupyter.

1.3.2. Les bibliothèques

Les fonctions python utilisées proviennent de ces bibliothèques principales (tableau 08) :

Tableau 08 : Différents bibliothèques python utilisées

Bibliothèque	Description
Pandas 2.0.2	La bibliothèque de logiciels open source Pandas est spécialement conçue pour manipuler et analyser des données dans le langage Python. Il est puissant, flexible et facile à utiliser.
Numpy 1.24.3	Le terme Numpy est en fait l'abréviation de "Numerical Python". C'est une bibliothèque open source en langage Python. Cet outil est utilisé pour la programmation scientifique en Python, notamment la programmation en science des données, pour l'ingénierie, les mathématiques ou les sciences.
Sklearn 1.0.2	Scikit-Learn, également connu sous le nom de sklearn, est la bibliothèque d'apprentissage automatique la plus puissante de Python. Il fournit une sélection d'outils puissants pour l'apprentissage automatique et la modélisation statistique, y compris la classification, la régression et le clustering via une interface cohérente en Python. Cette bibliothèque, qui constitue une partie importante des tests en Python, est applicable à NumPy, SciPy et Matplotlib.

2. MÉTHODES

Le processus de travail est divisé en deux parties principales, la section de prétraitement des données et la section ML :

2.1. Pré-traitement des données

Plusieurs fonctions sont utilisées pour filtrer et nettoyer les données afin de les rendre plus pratiques. Nos principales étapes de nettoyage et de prétraitement sont les suivantes :

2.1.1. Nettoyage du fichier FASTA et texte

La lecture de fichier FASTA : D'abord, nous avons chargé les séquences de nucléotides FASTA (malade et saine), exemple la séquence suivante (figure 16) :

```
1  >lcl|LC742143.1_gene_1 [gene=HP] [location=<1..>735] [gbkey=Gene]
2  CTTCTTATCTCGACCTCTGGGCTTTCAGGACCATAAAGAACATTGGGGTTCCTGCCAGAAATGAGGGGAG
3  CTTGCCTTTCCATTGGCTTCTATTCGGGGTGGAAGGAGATTGATGTGCAGAGCAGCTCCCGCTCATCTGA
4  CTTTTCACGGTTCACTGGGAACAATTTCCAAATAGCAAACTCTCTGGCTTCTCTCTCTTTGCAGATGACG
5  GCTGCCCGAAGCCCCCCGAGATTGCACATGGCTATGTGGAGCACTCGGTTCGCTACCAGTGTAAGAACTA
6  CTACAAACTGCGCACAGAAGGAGATGGTAAGATGTGGACAACTGTCTCCATGCCCTACATACAACCCCCT
7  TCTCTGACATTTCCATGATGGGTGGTGCTGAGGTGATTCGCCAGAAAGTTCGTTGCTCTCCTTGGAGCCA
8  GGAGATTTAGATTCTAATAAGGGTTTTGTCGCCAGTAGCCATGGCCCTTTGGGCAGACTAACTTTTGTCA
9  GCCTCAAGTTTTCTGTTTTGTTAAGGGGAGGCGATGCCATGCAGCCTACCTCATGTAAATCTCAGAGTCA
10 GATTTACATCTCCAGCAGATGTGGGAAAAGAAGGAATGCTGATGATGATGTCACCCTCACCTAGTGAGTC
11 TTGCTGTCCTGGCACTGCTCTAAGGGCTTTATACTTATTTGCTCACTTAGTCCTCACAGTATCCCTCTGA
12 ACAGAGTTTATTGTTTTCACTTTGCTGATAAGGAA
```

Figure 16 : La lecture et l'affichage du fichier FASTA.

- ✓ Les fichiers FASTA commencent souvent par une ligne d'en-tête qui peut contenir des commentaires ou d'autres informations. Le reste du fichier contient les données de la séquence. Chaque séquence commence par un symbole ">", suivi du nom de la séquence. A partir du reste de la ligne, nous trouvons les bases de la séquence. Nous allons supprimer cet entête. nous retrouvons la séquence suivante (figure 17) :

```
1
2  CTTCTTATCTCGACCTCTGGGCTTTCAGGACCATAAAGAACATTGGGGTTCCTGCCAGAAATGAGGGGAG
3  CTTGCCTTTCCATTGGCTTCTATTCGGGGTGGAAGGAGATTGATGTGCAGAGCAGCTCCCGCTCATCTGA
4  CTTTTCACGGTTCACTGGGAACAATTTCCAAATAGCAAACTCTCTGGCTTCTCTCTCTTTGCAGATGACG
5  GCTGCCCGAAGCCCCCCGAGATTGCACATGGCTATGTGGAGCACTCGGTTCGCTACCAGTGTAAGAACTA
6  CTACAAACTGCGCACAGAAGGAGATGGTAAGATGTGGACAACTGTCTCCATGCCCTACATACAACCCCCT
7  TCTCTGACATTTCCATGATGGGTGGTGCTGAGGTGATTCGCCAGAAAGTTCGTTGCTCTCCTTGGAGCCA
8  GGAGATTTAGATTCTAATAAGGGTTTTGTCGCCAGTAGCCATGGCCCTTTGGGCAGACTAACTTTTGTCA
9  GCCTCAAGTTTTCTGTTTTGTTAAGGGGAGGCGATGCCATGCAGCCTACCTCATGTAAATCTCAGAGTCA
10 GATTTACATCTCCAGCAGATGTGGGAAAAGAAGGAATGCTGATGATGATGTCACCCTCACCTAGTGAGTC
11 TTGCTGTCCTGGCACTGCTCTAAGGGCTTTATACTTATTTGCTCACTTAGTCCTCACAGTATCCCTCTGA
12 ACAGAGTTTATTGTTTTCACTTTGCTGATAAGGAA
```

Figure 17 : Le fichier FASTA final.

Une séquence nucléotide ne contient que les caractères A, C, T et G. Parfois, nous trouvons le caractère N ça soit au niveau des séquences au format Fasta ou texte. Ce caractère remplace une base azotée quelconque. Dans la procédure de prétraitement, nous éliminons ce caractère afin d'obtenir des résultats corrects.

2.1.2. Organisation des séquences

Nous regroupons tous les séquences des personnes qui ne sont pas malades sous forme d'une liste dans un seul fichier que nous avons nommé saine. Nous regroupons tous les séquences des personnes malades sous forme d'une liste dans un seul fichier que nous avons nommé malade.

Dans le but de rendre les séquences homogènes et afin de faciliter le traitement, nous délimitaient la longueur de chaque séquence en 350.

Pour faciliter la lecture des résultats par la suite, nous avons proposé le code suivant :

- Numéro 2 : pour une séquence saine
- Numéro 3 : pour une séquence malade.

```
import os
files = os.listdir('dataset/saine')

ets = []
for file in files[:]:
    content = open("dataset/saine/"+file, "r").read()

    c = content.split('\n')

    c = [x for x in c if x.strip() != "" and ">" not in x ]
    c = "".join(c)
    if c[0:350] not in _alls and "N" not in c[0:350]:

        ets.append(2)
        _alls.append(c[0:350])


files = os.listdir('dataset/s')
for file in files[:]:
    content = open("dataset/s/"+file, "r").read()

    c = content.split('\n')

    c = [x for x in c if x.strip() != "" and ">" not in x ]
    c = "".join(c)
    if c[0:350] not in _alls and "N" not in c[0:350]:

        ets.append(2)
        _alls.append(c[0:350])
```

Figure 18 : Lecture et l'affichage des séquences saines

```
#lecture des sq malade
files = os.listdir('dataset/malade')

etm = []
for file in files[:]:
    content = open("dataset/malade/"+file, "r").read()

    c = content.split('\n')[:]

    c = [x for x in c if x.strip() != "" and ">" not in x ]
    c = "".join(c)
    if c[0:350] not in _alls and "N" not in c[0:350]:
        etm.append(3)
        _alls.append(c[0:350] )

files = os.listdir('dataset/m')


for file in files[:]:
    content = open("dataset/m/"+file, "r").read()

    c = content.split('\n')[:]

    c = [x for x in c if x.strip() != "" and ">" not in x ]
    c = "".join(c)
    if c[0:350] not in _alls and "N" not in c[0:350]:
        etm.append(3)
        _alls.append(c[0:350] )
#sequences = [seq for seq in sequences if seq != ""]
```

Figure 19 : Lecture et l'affichage des séquences malade

2.1.3. Numérisation des séquences :

Puis les techniques de ML sont destinées à traiter les données numériques, il faut numériser les séquences qui sont représentées comme des chaines des caractères. De ce fait, nous proposons le code suivant pour chaque base nucléotidique :

A = 0.25, T = 1, G = 0.75, C = 0.5.

```
import numpy as np
sequences = _alls
num = {"A":0.25, "T":1, "G":0.75, "C":0.5}


numeric_sequences = []

for seq in sequences:
    _seq =  []
    for n in seq:
        if n in num.keys():
            _seq.append(num[n])
        else:
            _seq.append(0)
    numeric_sequences.append(_seq)
numeric_sequences = np.array(numeric_sequences)
numeric_sequences
```

Figure 20 : Numérisation des séquences.

Nous présentons un exemple du résultat de la numérisation dans la figure suivante :

```
array([list([0.25, 1, 0.75, 1, 1, 1, 0.75, 1, 0.75, 1, 1, 0.5, 0.5, 1, 0.75, 0.75, 1, 0.75, 0.5, 1, 0.75, 0.5, 1, 0.75, 0.5,
0.5, 0.25, 0.5, 1, 0.75, 0.75, 1, 0.75, 1, 0.5, 0.5, 0.25, 0.75, 0.5, 0.5, 0.25, 0.75, 1, 0.75, 1, 0.75, 1, 0.75, 0.25, 0.25,
0.5, 0.5, 1, 0.75, 0.25, 0.5, 0.5, 0.25, 0.5, 0.5, 0.25, 0.75, 0.75, 0.25, 0.5, 0.5, 0.5, 0.25, 0.25, 0.5, 1, 1, 0.5, 0.5, 1,
0.5, 0.5, 1, 0.75, 0.5, 0.5, 1, 0.25, 0.5, 0.25, 0.5, 0.5, 0.25, 0.25, 0.5, 1, 0.5, 0.5, 1, 1, 0.5, 0.25, 0.5, 0.5, 0.25, 0.7
5, 0.75, 0.75, 0.75, 0.25, 0.75, 1, 0.5, 1, 0.25, 0.5, 1, 0.25, 0.5, 0.5, 0.5, 1, 0.75, 0.25, 0.5, 0.25, 0.25, 0.75, 0.75, 1,
0.75, 1, 1, 0.5, 0.25, 0.75, 0.75, 1, 0.5, 0.5, 1, 0.5, 1, 0.75, 1, 0.75, 0.5, 1, 0.75, 0.5, 0.25, 0.5, 0.25, 0.75, 0.5, 0.2
5, 0.5, 0.5, 0.5, 0.25, 0.75, 0.75, 0.25, 0.5, 0.5, 1, 0.75, 1, 1, 0.5, 0.5, 1, 0.75, 0.5, 0.5, 0.25, 1, 1, 0.5, 1, 1, 0.5,
0.25, 0.75, 0.5, 0.25, 0.25, 1, 0.75, 1, 0.75, 0.25, 0.5, 0.5, 1, 0.75, 0.75, 1, 1, 0.5, 0.5, 0.25, 1, 0.75, 0.5, 0.5, 0.25,
1, 0.5, 0.5, 0.25, 1, 0.75, 1, 0.75, 1, 0.5, 1, 0.75, 0.75, 0.5, 0.25, 0.5, 0.5, 0.25, 0.25, 1, 0.75, 0.75, 0.5, 0.25, 0.5,
0.5, 0.25, 0.25, 0.75, 0.25, 0.75, 0.75, 1, 1, 1, 0.75, 0.25, 0.5, 0.25, 0.25, 0.5, 0.5, 0.5, 1, 0.75, 1, 0.75, 0.5, 1, 0.75,
0.5, 0.5, 0.25, 1, 1, 0.5, 0.25, 0.25, 1, 0.75, 0.25, 1, 0.75, 0.75, 0.25, 0.75, 1, 0.5, 1, 0.25, 0.5, 1, 1, 1, 0.75, 0.5, 0.
5, 0.25, 0.75, 0.5, 0.25, 0.5, 0.25, 0.75, 0.25, 0.75, 0.25, 0.25, 0.75, 0.25, 0.75, 0.5, 0.25, 0.25, 0.5, 0.25, 1, 0.5, 0.2
5, 1, 0.5, 0.25, 0.75, 0.75, 0.75, 0.75, 0.5, 1, 0.75, 0.75, 0.25, 1, 1, 1, 1, 0.75, 0.75, 0.5, 0.25, 0.5, 0.5, 0.25, 0.5,
0.5, 0.5, 1, 0.75, 0.75, 0.25, 0.5, 0.25, 0.75, 0.5, 0.25, 0.25, 0.75, 0.25, 0.5, 0.5, 0.5, 0.25, 0.75, 1, 0.5, 0.5, 0.5,
1])
```

Figure 21 : Exemple d'une séquence numérisée.

2.1.4. Création du fichier csv :

- Enregistrement des fichiers sains et malades numérisés sous format csv.
- Ajout de noms de chromosomes.
- Sélection des nucléotides du chromosome 11.
- Remplacement de bases azotées manquantes par des zéros.
- Supprimé l'entête :

```
def Rempli(matrice, longueur):
    sortie = []
    for m in matrice:
        if len(m) < longueur:
            m = m + [0 for i in range(longueur - len(m))]
        sortie.append(m)

    return np.array(sortie)

longueur = max(len(m) for m in numeric_sequences)
Sortie = Rempli(numeric_sequences, longueur)
print("columns : ", len(Sortie[0]))
print("lines :", len(Sortie))
```

```
columns :  350
lines : 133
```

```
import pandas as pd
data = []
for i in range(len(Sortie)):
    list_ = list(Sortie[i])
    list_.insert(0, et[i])
    data.append(list_)


csv = pd.DataFrame(data)
csv
```

	0	1	2	3	4	5	6	7	8	9	...	341	342	343	344	345	346	347	348	349	350
0	2	0.25	1.00	0.75	1.00	1.00	1.00	0.75	1.00	0.75	...	0.50	0.50	0.50	0.25	0.75	1.00	0.50	0.50	0.50	1.00
1	2	0.25	1.00	0.75	0.75	0.25	0.75	0.25	0.75	0.50	...	1.00	0.25	0.50	0.50	0.25	0.75	1.00	0.75	0.75	0.50
2	2	0.25	1.00	0.75	1.00	1.00	1.00	0.75	1.00	1.00	...	0.50	0.50	0.50	0.25	0.75	1.00	0.50	0.50	0.50	1.00
3	2	0.25	1.00	1.00	0.25	1.00	1.00	0.25	0.25	0.50	...	0.25	0.75	0.25	0.50	0.25	1.00	1.00	0.50	0.25	0.25
4	2	1.00	0.25	0.50	0.25	0.25	0.25	0.25	0.50	0.25	...	1.00	0.75	1.00	0.25	0.25	0.50	0.25	1.00	1.00	0.25
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
128	3	0.50	0.25	0.75	1.00	0.50	1.00	0.75	0.50	0.50	...	0.25	0.75	0.75	0.50	0.50	0.25	0.25	0.50	0.50	0.50
129	3	1.00	0.50	0.50	1.00	0.50	1.00	0.75	0.25	0.75	...	0.50	0.50	0.25	0.25	0.50	0.50	0.50	0.50	0.25	0.50
130	3	0.75	0.25	1.00	1.00	1.00	1.00	0.25	1.00	0.75	...	0.25	0.75	0.50	0.50	0.50	0.25	0.25	0.75	0.75	0.50
131	3	0.75	0.25	1.00	1.00	1.00	1.00	0.25	1.00	0.75	...	0.25	0.75	0.50	0.50	0.50	0.25	0.25	0.75	0.75	0.50
132	3	0.75	0.50	0.50	0.50	1.00	1.00	0.50	0.25	0.75	...	0.50	0.50	0.25	0.25	0.50	0.50	0.50	0.50	0.25	0.50

133 rows × 351 columns

Figure 22 : Lecture des fichiers malade et saine csv.

Le contenu de chaque fichier csv est représenté sous forme d'une matrice où chaque ligne représente une séquence et chaque colonne représente une base d'une séquence codée selon le code :

0.25: la base A
0.5 : la base C
0.75 : la base G
1 : la base T

```python
import csv

fichier_entree = 'dataset/csv.csv'
fichier_sortie = 'dataset/csv_.csv'

with open(fichier_entree, 'r') as csv_entree, open(fichier_sortie, 'w', newline='') as csv_sortie:
    lecteur_csv = csv.reader(csv_entree)
    ecrivain_csv = csv.writer(csv_sortie)

    lignes = list(lecteur_csv)  # Convertir le lecteur CSV en une liste de lignes

    if len(lignes) > 1:  # Vérifier s'il y a plus d'une ligne dans le fichier
        lignes = lignes[1:]  # Supprimer la première ligne (l'en-tête des colonnes)

    ecrivain_csv.writerows(lignes)  # Écrire les lignes dans le fichier de sortie
print("terminé")
```

Figure 23 : Enregistrement du fichier du dataset final sous format csv.

2.2. Apprentissages :

2.2.1. Répartition des données pour l'apprentissage, le test et l'évaluation

Les données (133 séquences nucléotidiques pour individus) on était divisées en deux ensembles. Le premier représente 80% (106 individus) des données. Elles ont servi à entraîner le modèle. La deuxième fraction du dataset représente les 20% (27 individus) pour tester le modèle. La division a été effectuée aléatoirement en utilisant la fonction train_test_split de sklearn.

```
from sklearn.model_selection import train_test_split
import csv
import numpy

sequences = np.array(list(csv.reader(open('dataset/csv_.csv', "r"))))
# Assuming you have your sequence data stored in a variable called 'sequences'
# and the corresponding labels in a variable called 'labels'
labels = et
# Splitting the data into training and testing sets
train_sequences, test_sequences, train_labels, test_labels = train_test_split(sequences, labels, test_size=0.2, random_state=42)

# Print the sizes of the training and testing sets
print("Training set size:", len(train_sequences))
print("Testing set size:", len(test_sequences))

print(train_labels)
```

```
Training set size: 106
Testing set size: 27
[3, 3, 3, 3, 3, 2, 3, 3, 2, 3, 3, 3, 2, 3, 3, 2, 3, 3, 3, 2, 3, 3, 2, 2, 3, 2, 3, 2, 3, 3, 3, 3, 2, 3, 3, 2, 3, 2, 2, 3, 2, 3,
3, 3, 3, 2, 2, 3, 2, 2, 2, 3, 3, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 2, 3, 3, 3, 3, 2, 2, 3, 3,
2, 3, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 2, 3, 3, 3, 2, 3, 3]
```

Figure 24 : Répartition des données via la fonction train_test_spli.

2.2.2. Construction du modèle :

- Le problème traité sert à classifier les séquences en deux classes : malade ou sain. De ce fait le problème est un problème de classification.
- La taille du dataset est petite. Pour cela, nous exploitons une des techniques de Machine Learning (ML) et pas Dep Learning (DL).
- Le modèle de classifieur qui a été exploité est le modèle RandomForestClassifier (forêt aléatoire).
- Cet algorithme de forêt aléatoire est connu pour être l'un des classificateurs prêts à l'emploi les plus efficaces.
- l'instruction clf.fit(train_sequences, train_labels sert à entrainer le modèle.
- Après l'apprentissage, le modèle est évalué avec les données de de validation. L'instruction clf.predict(test_sequences) sert à tester le modèle.

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score
clf = RandomForestClassifier(n_estimators=100)

# Entraîner le classifieur
clf.fit(train_sequences, train_labels)

# Faire des prédictions sur l'ensemble de test
y_pred = clf.predict(test_sequences)
print(y_pred)
# Calculer l'exactitude des prédictions
accuracy = accuracy_score(test_labels, y_pred)
print("Exactitude : {:.2f}".format(accuracy))
```

```
[3 3 2 3 3 3 3 3 3 2 3 3 3 3 3 3 3 2 3 3 3 2 3 3 3 3 3]
Exactitude : 0.70
```

Figure 24 : création d'un classifieur de forêt aléatoire.

2.2.3 Code de la Prédiction d'une séquence

La classification d'une séquence implique la prédiction d'une étiquette de classe pour cette séquence d'entrée.

```
def ToNumeric(seq):

    num = {"A":0.25, "T":1, "G":0.75, "C":0.5}

    numeric_sequences = []
    _seq = []

    for n in seq:

        if n in num.keys():
            _seq.append(num[n])
        else:
            _seq.append(0)

    if len(_seq) < 351:
        _seq = _seq + [0 for i in range(351 - len(_seq))]
    elif len(_seq) > 351:
        _seq = _seq[0:351]
    numeric_sequences.append(_seq)


    numeric_sequences = np.array(numeric_sequences)
    return numeric_sequences

    print(len(numeric_sequences))
    print(numeric_sequences)
```

Figure 25 : Code de prédiction, avec séquences.

2.2.4. Visualisation des résultats :

Étant donné une séquence d'ADN de valeurs ACGT :

- En premier lieu, l'application développée va numériser cette séquence,
- Ensuite, le modèle va décéder s'elle est appartenue à la classe saine ou malade : si le résultat est 2, ça signifie que la séquence appartient à la classe saine. Si le résultat est 3 ça signifie que la séquence appartient à la classe malade.
-

```
# Faire des prédictions sur les nouvelles séquences
y_new_pred = clf.predict(X_new)
print(y_new_pred)


labels_ = {2:"saine", 3:"malade"}


# Afficher les prédictions
for sequence, prediction in zip(new_sequences, y_new_pred):
    print("Séquence d'ADN : ", sequence)
    print("Prédiction : ",prediction, labels_[prediction])
```

Figure 26 : Code de prédiction pour une séquence.

PARTIE 3 : RÉSULTATS ET DISCUSSION

Résultats

En fin d'apprentissage, le modèle est testé pour vérifier son efficacité. Le test a été réalisé sur 20 % (c'est-à-dire 27 individus) des données rapportées dans la section Méthodes et réalisé à l'aide de la matrice de confusion illustrée à la figure 27. Le modèle prédit :

- 15 individus appartiennent à la classe malade et réellement ces personnes sont malades. Donc il y a 15 cas Vrai Positif (VP).
- 4 individus n'appartiennent pas à la classe malade et réellement ces personnes ne sont pas malades. Donc il y a 4 cas Vrai Négatif (VN).
- 8 individus appartiennent à la classe malade et réellement ces personnes ne sont malades pas. Donc il y a 8 cas Faux Positif (FP).
- 0 individu n'appartient pas à la classe malade et réellement ces personne sont malades. Donc il y a 0 cas Faux Négatif (FN).

Les pourcentages de VP, VN, FP et FN sont représentées comme suit :

VP = 55%
VN = 15%
FP = 30%
FN = 0%

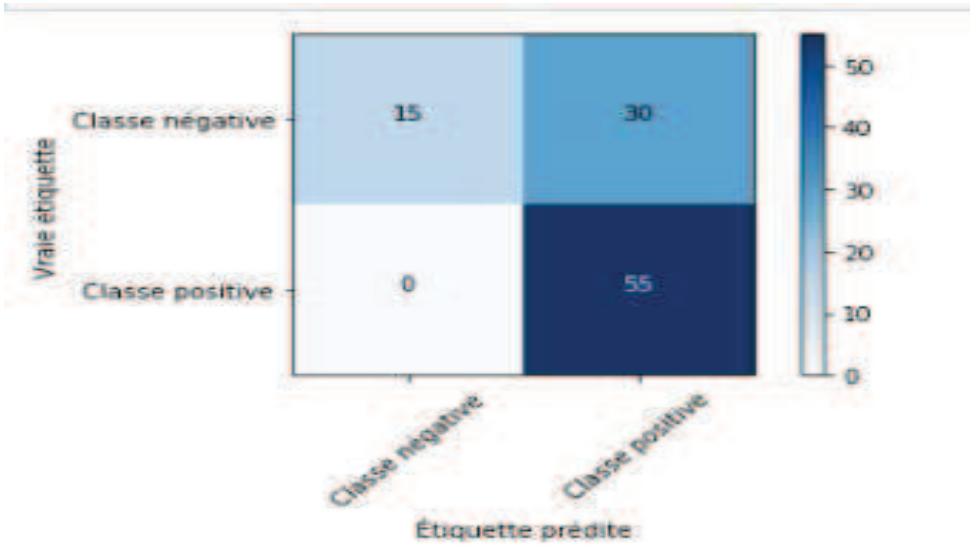

Figure 27 : Matrice de confusion du test du modèle FASTA.

- Accuracy = (VP+VN) / (VP+VN+FP+FN) 70%
- Précision = VP / (VP+FP)
- Sensibilité = VP/ (VP+FN)
- Spécifié = VN/(VN+FN)

Cette matrice a permis de calculer les valeurs suivantes :

- Accuracy = 70%
- Précision = 100%

- Sensibilité = 30%
- Spécifié (ROC_AUC) = 91.62%

Le récepteur Receiver Operating Characteristics (ROC) est une mesure d'évaluation des problèmes de classification binaire. ROC représente probabilité qui du taux de vrais positifs par rapport au taux de faux positifs. Area Under Curve (AUC) est la mesure de la capacité d'un classificateur à distinguer les classes et est utilisée comme résumé de la courbe ROC. Plus l'AUC est élevée (dans notre cas il est égal à 0.91, voire figure 28), plus le modèle est efficace pour distinguer les classes positives des classes négatives.

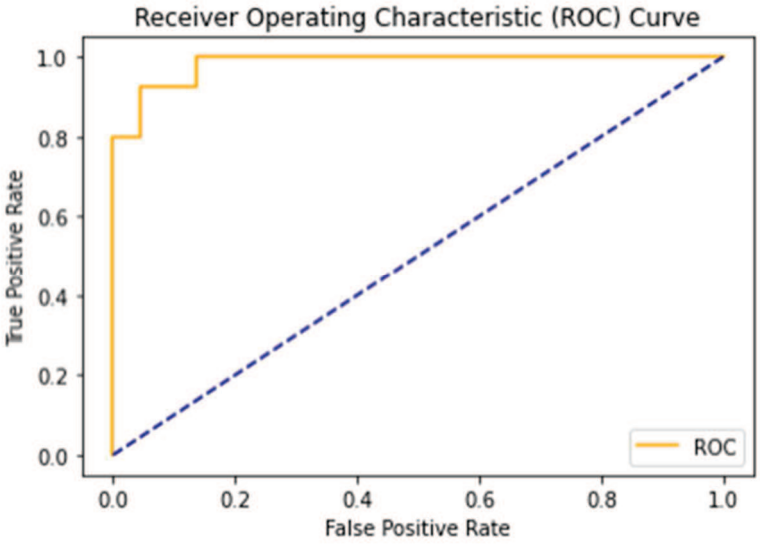

Figure 28 : La courbe de caractéristique de fonctionnement du récepteur.

DISCUSSION

Dans notre travail, nous avons identifié 133 nucléotides localisés dans le chromosome 11 avec des positions bien précises responsables de la AF. A l'aide de notre modèle RandomForestClassifier génétique prédictif adapté. Ce modèle serait suffisant pour informer de la présence ou de l'absence de l'AF pour un patient donné.

La méthode proposée a réussi à classer les personnes non malades et malades et a montré son efficacité avec une précision et une accuracy de 70% un score ROC_AUC de 91.62%. Les résultats obtenus durant ce travail démontrent que notre modèle RandomForestClassifier pourrait jouer un rôle efficace dans la prédiction et la classification de la AF.

CONCLUSION

CONCLUSION

Notre étude a confirmé la faisabilité et surtout l'importance de prédire la FA à partir des séquences FASTA, en utilisant un modèle ML qui prend en charge les nucléotides de l'ensemble de données et utilise le classificateur RandomForestClassifier. Nous avons validé notre algorithme d'apprentissage en utilisant 133 personnes de l'ensemble de données ; qui a produit un classeur dont la précision était très significative.

De plus, nous nous sommes concentrés uniquement sur les nucléotides situés dans le chromosome 11. Le fait que notre étude ait produit des résultats statistiquement significatifs, malgré ces limitations statistiques en nombre et en qualité, démontre le potentiel de cette approche ML dans le contexte de la prédiction de maladies.

Dans nos travaux futurs, nous prévoyons d'appliquer cette méthode sur tous les chromosomes et sur un échantillon humain plus grand, d'élaborer des modèles de classification plus précis et plus crédibles. .

CONCLUSION

RÉFÉRENCES BIBLIOGRAPHIQUES

RÉFÉRENCES

Aguedach A, Brosillon S., Morvan J., Lhadi E. Morvan D, 2005 « Photocatalytic degradation of azo-dyes reactive black 5 and reactive yellow 145 in water over a newly deposited titanium dioxide » *Applied Catalysis B: Environmental* , vol. 57, no. 1, pp. 55-62

Audet S. 2022 « identification de gènes impliqués dans les ataxies épisodiques par combinaison de séquençages génomique ettranscriptomique », mémoire présenté en vue de l'obtention du grade maîtrise ès sciences (m.sc.) Enneurosciencesuniversité de montréal, québec. (https://papyrus.bib.umontreal.ca/xmlui/handle/1866/27202)

Aymé C, 2001 « Complexe de cytochromes essentiel pour l'oxydation photosynthétique du thiosulfate et du sulfure chez Rhodovulum sulfidophilum » *Journal of Bacteriology*, vol 183, no 20

Bagnoud G. 2009 « l'innovation médicale et son intégration dans les assurancessociales », mémoire de master en ligne, https://serval.unil.ch/resource/serval:bib_89cd8855c829.p001/ref)

Balacheff N. 1994 « Didactique et intelligence artificielle ».*Recherches en didactique des mathématiques.* Vol. 14, pp9-42.

Banza M. I., Mulefu J. P., Lire I. L., Yannick T. Ben N'., Israel T. B., Vincent de P. K. Cabala, 2019 « Pathologies digestives associées à la drépanocytose à Lubumbashi: aspects épidémiologiques et cliniques » *Pan African Medical Journal*, vol. 33 :253 [doi: 10.11604/pamj.2019.33.253.18017]

Benammi S, Bakali Y, Alaoui M, Sebbah F. 2020 « La technologie va-t-elle accompagner ou remplacer le chirurgien : introduction à l'intelligence artificielle (AI) et perspectives d'avenir ».*Journal de Chirurgie Viscérale,* Vol. 157, No. 3 page S181.

Beroud C., 2010 « niveaux annotations Pathologie Biologie », vol 58, pp 387-395

Berrou J, 2020 « Une révolution mobile en Afrique subsaharienne » Réseaux *: communication, technologie, société*, no 219, pp 11-38, doi : 10.3917/res.219.0011

Besnard X., Mathieu B., Nadège C., Delphine R. (DREES), 2019 « Les proches aidants des seniors et leur ressenti sur l'aide apportée - Résultats des enquêtes 'CARE ' auprès des aidants (2015-2016) » *LES DOSSIERS DE LA DREES* N° 45 https://drees.solidarites-sante.gouv.fr/publications/les-dossiers-de-la-drees/les-proches-aidants-des-seniors-et-leur-ressenti-sur-laide

Boichard D., Le R. P., Levéziel H., Elsen J. M... 1998 « utilisation des marqueurs moléculaires en génétique animale », *productions animales*, vol. 11 no 1 (1998).

Bonneuil C, Demeulenaere E, Thomas F, Joly P-B, Allaire G et Goldringer I, 2006 « Innover autrement? La recherche face à l'avènement d'un nouveau régime de Production et de régulation des savoirs en génétique végétale » Dossier de l'environnement de l'INRA n° 30

Bonneuil C., Demeulenaere E., Thomas F., Joly Pierre-Benoit, Allaire G. et Goldringer I., 2006 « Innover autrement ? La recherche face à l'avènement d'un nouveau régime de production et de régulation des savoirs en génétique végétale ». Dossier de l'environnement de l'INRA n° 30. https://www.researchgate.net/publication/32232478_Innover_autrement_La_recherche_face_a_l'avenement_d'un_nouveau_regime_de_production_et_de_regulation_des_savoirs_en_genetique_vegetale

Boudiaf K, Bouhemadou A, 2018 « Electronic and thermoelectric properties of the layered» aFAgCh (Ch = S, Se and Te): First-principles study » *Journal of Alloys and Compounds,* vol. 759, pp. 32-43 DOI: 10.1016/j.jallcom.2018.05.142

BOULDJADJ R. 2020 « cours génétique » support du cours, université frère Mentouri, Constantine, Algerie

Bucci, V., Tzen, B., Li, N. *et al*. Stein, 2016 «MDSINE: Microbial Dynamical Systems INference Engine for microbiome time-Series analyses », *Genome Biol,* vol. 17, no. 121. https://doi.org/10.1186/s13059-016-0980-6

Cadavid J. P., LAMOURI S., GRABOT B., FORTIN A., 2021 « L'Apprentissage Automatique dans la planification et le contrôle de la production: un état de l'art » Conférence Internationale Génie Industriel Qualita, Jun 2019, Montréal, Canada. ⟨hal-03267867⟩

Calistru D, 2022 « Utilisation d'une hybridation de la recherche opérati onnelle et de L'apprentissage automatique pour injecter de l'émotion dans un agent Conversationnel » Mémoire de maîtrise, Polytechnique Montréal, Canada

Carolyn M, 2022 « Geology and Stratigraphic Correlation of the Murray and Carolyn Shoemaker Formations Across the Glen Torridon Region, Gale Crater, Mars vol 127, doi: 10.1029/2022JE007408

Catonné Y, 2013 « Fracture instable de l'odontoïde : stratégie chirurgicale dans une série de 22 cas et revue de la littérature » *Revue de Chirurgie Orthopédique et Traumatologique* vol 99, no 5, pp 509-517, doi : 10.1016/j.otsr.2013.03.049

Cazals F, Cazals C. 2020« intelligence artificielle : l'intelligence amplifiée par la technologie ». Édition de boeck : bruxelles (belgique).

Charlin L, 2017 « A Hierarchical Latent Variable Encoder-Decoder Model for Generating DialoguesProceedings of the AAAI» Conference on Artificial Intelligence, vol 31, no 1, doi: 10.1609/aaai. v31i1.10983

Cheikh M, 2020 « mutations gènètique et polymorphisme » *Journal européen de génétique médicale*, vol 49, pp 481-486

Chudasama V., Nighania K., Upla K., Raja K., Ramachandra R., Busch C., 2021 «E-comsupresnet: Enhanced face super-resolution through compact network » *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 2, pp. 166-179, , doi: 10.1109/TBIOM.2021.3059196.

Clement T, Russo M, 2021 « Psychodynamic Psychotherapy for Children as a Trauma-Informed Intervention. » *Le Journal officiel de l'Académie américaine de psychiatrie psychodynamique et de psychanalyse*, vol 51, pp. 133-250, doi :10,1521

Cleuziou G, 2004 « Une méthode de classification non-supervisée pour l'apprentissage de règles et la recherche d'information » THESE De doctorat, l'Universite d'Orleans, Orléans Cedex 2 - France

Cobessi D, Meksem A, Brillet K ,2010 « Structure of the heme/hemoglobin outer membrane receptor ShuA from Shigella dysenteriae: Heme binding by an induced fit mechanism" *PROTEINS: Structure, Function, and Bioinformatics* Vol78, no 2, PP 286-294

Cohen E. E. W., Soulières D. *et al.*, 2019«Pembrolizumab versus methotrexate, docetaxel, or cetuximab for recurrent or metastatic head-and-neck squamous cell carcinoma (KEYNOTE-040): a randomised open-label, phase 3 study » *Lancet,* Vol. 393, no. 10167, pp.156-167. doi: 10.1016/S0140-6736(18)31999-8.

Couque N, Montalembert M, 2013 « Diagnostic d'une hémoglobinopathie » *Feuillets de Biologie*, vol 311, no5, pp 777-780

Damon M, Oswald I, Luronjourn I. Rech, 2003 « apport des nouveaux outils de la postgénomique aux recherches en physiologie chez le porc » journées de recherche porcine, vol. 35, pp. 339-354.

Dekeuwer C, 2020 «Philosophie pratique de terrain : quelle posture de recherche ? », *Éthique, politique, religions*, vol. 02, no. 15, pp. 131-145. DOI : 10.15122/isbn.978-2-406-10144-4. p.0131

Dheur S. et Saupe S. J., 2021 « Mutations du mythe de l'ADN, étapes de matérialisation du gène ». *Matières vivantes*, vol 40, pp. 85-107

Dimassi S, Tilla M, Sanlaville D. 2017 « Anomalies chromosomiques » *Journal de Pédiatrie et de Puériculture*,Vol. 30,No 5/6, pp. 249-270.

Diop M, Fall M, 2022 « Intercontinental Spread of Eurasian Highly Pathogenic Avian Influenza A(H5N1) to Senegal", *Emerg Infect Dis*, vol 28, no 1, pp 234–237.doi: 10.3201/eid2801.211401

Dubois M, 1988 « chromosome sexuel Inactivation and reactivation of sex-linked steroid sulfatase gene in murine cell culture vol 14, pp113–121

Duclayen C, Yvon V, 2003 « La surexpression du ligand Notch, Jagged-1, induit des cellules T régulatrices humaines spécifiques de l'alloantigène, *blood*, vol 102, pp 3815–3821, doi : 10.1182/blood-2002-12-3826

Duvoux C, Roudot–Thoraval F, Decaens T, Pessione F, Badran H, Piardi T, Francoz C, Compagnon P, Vanlemmens C, Dumortier J, Dharancy S, Gugenheim J, *et al.*, 2012 «Liver transplantation for hepatocellular carcinoma: a model including α-fetoprotein improves the performance of Milan criteria » *Gastroenterology*Vol. 143, no. 4, pp. 986-994.e3

Eensoo E, Nouvel D, *et al*. 2016 « Combiner analyses textométriques, apprentissage supervisé et représentation vectorielle pour l'analyse de la subjectivité » 11e Défi Fouille de Texte (DEFT'2015), Caen (France), Jun 2016, Caen, France. ffhal-01335127f

Gabriel J, 2019 « Marvin Minsky : un des cerveaux de l'intelligence artificielle » [site en ligne] [consulté le 13 Mai 2023] : http://interstices.info

Gallaiset A., Bannerot H, 1992 *« Amélioration des espèces végétales cultivées. Objectifs et critères desélection »* INRA : France.

Gaudriault S, Vincent R. 2009 « Génomique », édition de boeck : Bruxelles (Belgique).

Gibson G, Muse SV. 2004 « précis de génomique »édition de boeck : bruxelles (belgique).

Goncharuk V, Zeng Z, Wang R, MacTavish D, Jhamandas J H,2004 «Distribution of the neuropeptide FF1 receptor (hFF1) in the human hypothalamus and surrounding basal forebrain structures: immunohistochemical study » *J Comp Neurol*, vol. 474, no. 4, pp. 487-503. doi: 10.1002/cne.20132.

Guellaën G, Andrologie, 1999 « Le projet " Génome Humain " et la caractérisation des étiquettes (E.S.T.) de testicule humain , *Génétique et Infertilités*, vol 9, no 3, pp 342-346

Guerd B., 2020« profil étiologique des anomalies congénitales du rein et du tractus urinaire (congenital anomalies of the kidney and urinarytract) (cakut)chezles enfants de 0-4 ans » thèse de doctorat, universite d'alger, algerie. (http://193.194.83.98/jspui/handle/1635/15489)

Haiech J, 2020 « Parcourir l'histoire de l'intelligence artificielle, pour mieux la définir et là Comprendre » *médecine/sciences* vol.36, pp.913-915

Hamerton L., McNamara L. T., Howlin B. J., Smith P. A., Cross P., Ward S., 2013 «Examining the initiation of the polymerization mechanism and network development in aromatic polybenzoxazines» *Macromolecules,* vol. 46, no. 13, pp. 5117–5132

HOC JM. 1986 « psychologie, intelligence artificielle et automatique »editions mardaga : bruxelles (belgique).

Ilyushchanka A, Kusin R, Manoila Y, Charniak I, Kusin A, Yurchanka S, Staselka A, Maiseyeva A, Kurylchyk I, Semenov V, 2021 «Application of sprayed bronze powders of the BrSn10Ph1 and BrSn5Zn5Pb5 grades for applying protective coatings by gas-flame spraying» *Materials Science. Non-Equilibrium Phase Transformations.* vol. 7, no 2, p. 67-69

Jurinke C, 2005 « Une approche basée sur le polymorphisme d'un seul nucléotide pour l'identification et la caractérisation de la modulation de l'expression génique à l'aide de MassARRAY » *Recherche sur les mutations/Mécanismes fondamentaux et moléculaires de la mutagenèse* Vol 573, no1, p 83-95

Jurinke C, 2022 « Une approche basée sur le polymorphisme d'un seul nucléotide pour l'identification et la caractérisation de la modulation de l'expression génique à l'aide de MassARRAY » *Recherche sur les mutations/Mécanismes fondamentaux et moléculaires de la mutagenèse* Vol 573, no1, p 83-95

Kochko A.2000 « De l'utilisation des chromosomes Artificiels de plantes comme outil pour la conservation et l'exploitation des Ressources génétiques végétales » Agricultures, Vol. 9, No4

Korzeniewski k, Nitsch-Osuch A, Chciałowski A, Korsak J, 2013 «Environmental factors, immune changes and respiratory diseases in troops during military activities » *Respir Physiol Neurobiol* Vol. 187, no. 1, pp. 118-22. doi: 10.1016/j.resp.2013.02.003.

Labbe C, Grima N, Gautier T, Favier B, Byrne JA, 2019 « Semi-automated fact-checking of nucleotide sequence reagents in biomedical research publications: The Seek & Blastn tool » *PLoS ONE,* vol.14, no 3. https://doi.org/10.1371/journal.pone.0213266

Lainé F, 2009 « Comparison of blood tests for liver fibrosis specific or not to NAFLD » *Journal of Hepatology* , vol 50, pp 165-170

Lambert N.. 2014. « Génétique et transmission transgénérationnelle »,*Cahiers de psychologie clinique* »,Vol. 2,No 43, pp 11-28.

Larbi A, 2022 « Génomique fonctionelle » support du cours Génétique Master 1 Génomique fonctionelle, Université Constantine 1, Constantine, Algérie.

Laurière J, 1987 « *Intelligence Artificielle Résolution de problémes par l'Homme et la machine* » Editions Eyrolles 2e éd.,

Ledoux T., et Cointe P. 1995 « Les Metaclasses Explicites comme Outil pour Ameliorer la Conception des Bibliotheques de Classes ».In Actes des Journées du GDR de Programmation, Grenoble, France, November 1995.

Lemogoum D., Bortel L V, Van de Borne P., 2008 « Aspects vasculaires de la drépanocytose» *Sang Thrombose Vaisseaux,* vol 20, no 4, pp. 191-6

Lesse H.1970 « les nombres de chromosomes dans le groupe de lysandracoridonet leur incidence sur sa taxonomie [lep. Lycaenidae] », *annales de la société entomologique de france,* vol. 75, no 3 / 4, pp64-68

Levesque S., 2004« identification de gènes de susceptibilité de la prééclampsie". Thèsede doctorat, université laval, canada.

Li, Y., Chao H., Lizhong D., Zhongxiao L., Yijie P., et Xin G.. 2019. « deep learning in bioinformatics: introduction, application, and perspective in the big data era»*methods,* vol. 166, no. 15, pp.4-21. Doi: 10.1016/j.ymeth.2019.04.008.

Lucy pattersonP.. 2010 « maîtriser les maladies génétiques » consulté 30 mars 2023, (https://www.scienceinschool.org/fr/article/2010/insight- fr/)

Maura F, Boyle EM, Rustad EH, Ashby C, Kaminetzky D, Bruno B, Braunstein M, Bauer M, Blaney P, Wang Y, Ghamlouch H, Williams L, Stoeckle J, Davies FE, Walker BA, Maclachlan K, Diamond B, Landgren O, Morgan GJ., 2022 «Chromothripsis as a pathogenic driver of multiple myeloma » *Semin Cell Dev Biol*, vol. 123, pp. 115-123. doi: 10.1016/j.semcdb.2021.04.014.

Medkour S, 2008 « Consommation quotidienne de cannabis : un nouveau facteur de risque de gravité de la stéatose chez les patients atteints d'hépatite C chronique » vol 134, pp 432-439

Mélançon et Lambert, 1992 « résoudredeuxproblématiques »

Merlin C, 2013 « Les usines de traitement des eaux usées urbaines comme points chauds pour la propagation de bactéries et de gènes résistants aux antibiotiques dans l'environnement : une revue » *Science of The Total Environnent* Vol 447, pp. 345-360

Miller, G. A., Charles, W. G., 1991 « Contextual correlates of semantic similarity. » *Language and Cognitive Processes*, vol. 6, no. 1, pp 1–28. https://doi.org/10.1080/01690969108406936 Charles G, 1991 «Contextual correlates of semantic similarity» journal toxicology science pp 1-8 doi:01690969108406936

Mishra S., Molinaro R., 2022 « Estimates on the generalization error of physics-informed neural networks for approximating PDEs » *IMA Journal of Numerical Analysis* vol 43, no. 01 pp 1–43

Mojtahedi et Parastar, Jalali-Hera, 2008 « Mutations génétiques dans la AF » (http://sgugenetics.pbworks.com/w/page/61172304/Pathophysiology%20of%20Sickle%20Cell%20Anemia.)

Montgolfier S, 2015 « La médecine prédictive en oncologie, une discipline en construction : quels changements dans les questions éthiques, quel regard sur la vie des individus concernés » *SHS Web of Conferences* , vol 21, doi : 10.1051/shsconf/20152102003

Moscatelli M. 2020 « étude de la régulation et de la fonction du long arn non codant xact chez l'humain durant le développement précoce et dans le système hématopoïétique », thèse de doctorat, université de Paris, France.

N'Guessan A C, Vahou K M, 2020 « Analyse morphosémantique des créativités lexicales relatives au concept de démocratie employées par les Acteurs politiques lors des élections générales en Côte d'Ivoire de 2015 à 2020 » actes du premier colloque scientifique international du labodylcal en hommage au professeur flavien gbeto, les éditions labodylcal, pp.515-531, 2022, 978-99982-65-30-1.

N'GUESSAN Affoué Cécile, Vahou Marcel, 2021 « des créativités lexicales relatives au concept de démocratie employées par les Acteurs politiques lors des élections générales en Côte d'Ivoire de 2015 à 2020 » , Actes du premier colloque scientifique international du labodylcal en hommage au professeur flavien gbeto, Calavi, Bénin, 17, 18 et 19 Octobre 2021

Ouladbrahim A., Belaidi I., Khatir S., Magagnini E., Capozucca R., Abdel Wahab M., 2022 «Experimental crack identification of API X70 »steel pipeline using improved Artificial Neural Networks based on Whale Optimization Algorithm » *Mechanics of Materials, Volume 166, March 2022, 104200*

Parizeau M, 2004 « réseaux de neurones GIF-21140 et GIF-64326 » Université de Laval : Canada

Paslier D, Bernot A, 2001« Le Projet Génome Humain : quinze ans d'efforts » *Med Sci* (Paris) vol 17, no 3 pp 294-8

Pastre F *et al.* ,2000 « Le Tardiglaciaire des fonds de vallée du Bassin Parisien »n *Quaternairen* vol 11, no 2, pp.107-122, doi : 10.3406/quate.2000.1660

Pélissier D, 2020 « La coconstruction ambiguë de l'intelligence artificielle (IA), analyse de conception de l'intervention d'ouverture de chat bots de recrutement » *Communication & management* vol. 17, no 2020/2, pages 67 - 82

Perrot A, Lauwers-Cances V, Tournay E, Hulin C, Chretien ML, Royer B, Dib M, Decaux O, Jaccard A, Belhadj K, Brechignac S, Fontan J, Voillat L, Demarquette H, Collet P, Rodon P, Sohn C, Lifermann F, Orsini-Piocelle F, Richez V, Mohty M, Macro M, Minvielle S, Moreau P, Leleu X, Facon T, Attal M, Avet-Loiseau H, Corre J., 2019«Development and validation of a cytogenetic prognostic index predicting survival in multiple myeloma » *J Clin Oncol*. vol.37, no 19, pp. 1657-1665. doi: 10.1200/JCO.18.00776.

Pierre S et al, 2016 «New Empiricisms and New Materialisms: Conditions for New Inquiry » *Cultural Studies ↔ Critical Methodologies* Vol 16, no. 02, doi: 10.1177/1532708616636147

Pius T. Mpiana, Koto-te-NyiwaNgbolua, ShaTshibey D. Tshibangu 2019 « Les alicaments et la drépanocytose : une mini-revue »*Comptes Rendus Chimie*. Vol. 19, No7, pp. 884-889.

Poirier - Sherbrooke MS, 2016 « Perception des impacts, acceptation et Acceptabilité de dispositifs nanotechnologiques utilisés en médecine : le cas De la prévention et du traitement des plaies. »Thèse de doctorat, Université de Sherbrooke, Canada

Priyadarshini S, Cotton M, 2021« Un nouveau réseau neuronal profond basé sur la recherche LSTM-CNN-grid pour l'analyse des sentiments » *The Journal of Supercomputing* vol 77, p 13911–13932

Rakotomalala R, 2011 « Régression Logistique Binaire et Polytomique », *eric.univ-lyon2*, consulté le 10/05/2023, [en ligne] : https://eric.univ-lyon2.fr/ricco/cours/cours/pratique_regression_logistique.pdf

RATT H.. 2006 « Diagnostic biologique de hemoglobinopathies au chu-hjra», Mémoire de Diplôme d'Etudes de Formations Spécialisées (D.E.F.S) de Biologie Médicale, universite d'Antananarivo,Madagascar. (Biblio.univ-antananarivo.mg) « Physiopathologie de la drépanocytoseSicklecellPathophysiology » Transfusion Clinique et Biologique, Vol. 21, No 4-5, pp. 178-181.

Rodwell J, 2002 « La diversité de la végétation européenne Un aperçu des alliances phytosociologiques et de leurs relations avec les habitats EUNIS » An overview of phtosociological alliances and relationships to EUNIS habitats. Report EC-LNV 54

Rousseau F, 2003 « Le rôle de la passion dans le bien-être subjectif des aînés Bulletin de la Société » *Revue Québécoise de Psychologie*, vol. 24, no. 3, pp. 197–211.

Ruffié, J, Quilici, J C Fernet P, 1970 « Sur l'expression phénotypique des chromosomes Rz du systemeRhesus (étude de la population Chipada des hautes-Andes) » *C R Acad Hebd Seances Acad Sci D,* vol. 270, no. 20, pp. 2489-91.

Saidi O., 2020 « Bioéthique » support du cours L3 Génétique, université Oran. Algerie.

Sauret N, 2022 « Intelligence artificielle & Sciences humaines et sociales (SHS): opportunités, défis et perspectives ».*I2D–Information, données & documents.* Vol. 1, No1, pp.97-103.

Seitz N, Traynor Ki S., Steinhauer N *et al. 2022* «A national survey of managed honey bee 2014–2015 annual colony losses in the USA » *Journal of Apicultural Research,* vol54, no4, pp. 292-304

Sfar S, Chouchane L, 2008 « Le projet génome humain » *Pathologie Biologie* Vol 56, PP 170-175

Sheikh M. N. A., Halder M., 2019 « SDN-Based Approach to Evaluate the Best Controller: Internal Controller NOX and External Controllers POX, ONOS, RYU » *Global Journal of Computer Science and Technology*, vol.19, no. E1, pp. 21-32.

Souciet J *et al.* 2009 « Exploration génomique des levures hémiacomycètes : 1. Un ensemble d'espèces de levures pour les études d'évolution moléculaire [Genomic exploration of the hemiascomycetous yeasts: 1. A set of yeast species for molecular evolution studies] » *FEBS Lett* vol 487, no 1 pp 3-12, doi : 10.1016/S0014-5793(00)02272-9

Tanous C, Kieronczyk A, Helinck S, Chambellon E, Yvon M, 2002 «Glutamate dehydrogenase activity: a major criterion for the selection of flavour-producing lactic acid bacteria strains » *Antonie Van Leeuwenhoek,* vol. 82, no 1-4, pages 271-8.

Thévenet P., Shen Y., Maupetit J., Guyon F., Derreumaux P., TufféryP. 2012 «PEP-FOLD: an updated de novo structure prediction server for both linear and disulfide bonded cyclic peptides» *Nucleic Acids Res,* vol. 40, no. Web Server, pp.W288-93. doi: 10.1093/nar/gks419.

Thomas M. Morgan, Harlan M. Krumholz, Richard P. Lifton, John A. Spertus. 2007 «Nonvalidation of reported genetic risk factors for acute coronary syndrome in a large-scale replication study »*JAMA*, Vol. 297, No. 14, pp.1551-61

Till I, Valdeyron G, Gouyon Ph., 1989«polymorphisme pollinique et polymorphisme génétique » *canadian journal of botany,* vol. 67, no 2, pp. 538-543

Touchon M., Hoede C., Tenaillon O., 2009 «Organized Genome Dynamics in the Escherichia coli Species Results in Highly Diverse Adaptive Paths » *PLoS Genet,* vol. 5, no.1: e1000344. doi: 10.1371/journal.pgen.1000344.

Verdier L, 1996 « Des catégories dérivées des catégories abéliennes » Astérisque : Paris (France).

Exploitation des techniques d'apprentissage automatique pour expliquer le séquençage spontané de l'ADN de la drépanocytose héréditaire/maladies génétiques

Mémoire pour l'obtention du diplôme de Master en Bioinformatique

La drépanocytose ou L'anémie falciforme (AF) est une maladie autosomique récessive causée par une mutation ponctuelle du gène de la globine situé sur le chromosome 11. La mutation du codon 6 entraîne le remplacement de l'acide glutamique 6 par la valine. Les globules rouges changent leur forme d'une boule ronde à une forme de croissant lorsqu'un trouble survient dans les gènes responsables de la formation de l'hémoglobine. Afin de trouver les mutations qui se produisent dans les polymorphismes sur ce gène spécifique, notre travail a mis en évidence l'importance d'utiliser FASTA_format pour étudier et analyser l'AF, et a développé une approche bio-informatique d'apprentissage automatique pour classer la maladie et les stades sains. Le but de cette approche est de développer un modèle explicatif afin de déterminer l'existence de maladie chez les individus notamment avec l'absence des symptômes ou d'estimer son stade, à l'aide d'un classifieur automatique, de ce fait, nous avons classé les données génétiques fonctionnelles avec une précision de test de 70%.

Mots-clés : Maladie Drépanocytose ; FASTA ; Apprentissage automatique.

Président : Dr GHERBOUDJ AMIRA (Université Frères Mentouri, Constantine 1).
Encadreur : OUAHIBA DJAMA (Université Frères Mentouri, Constantine1).
Examinateur : Dr TAMAGOULT MAHMOUD (Université Frères Mentouri, Constantine 1).