الجمهورية الجزائرية الديمقراطية الشعبية

République Algérienne Démocratique et Populaire

وزارة التعليم العالي والبحث العلمي

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

جامعة الإخوة منتوري قسنطينة 1

Frères Mentouri Constantine 1 University

Université Frères Mentouri Constantine 1

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

كلية علوم الطبيعة والحياة
قسم علوم الطبيعة و الحياة

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*

N° d'ordre :
N° de série :

Intitulé :

Classification des hydrolases à partir des données protéomiques par approche Deep Learning

Présenté par : **BELBEKRI Amina** **Le 22/06/2022**

Jury d'évaluation :

Encadreur : HAMIDECHI Mohamed Abdelhafid (Professeur - Université Frères Mentouri, Constantine 1).

Examinateur 1 : DAAS Mohamed Skandar (MCA - Université Frères Mentouri, Constantine 1).

Examinateur 2 : BOULAHROUF Khaled (MCB - Université Frères Mentouri, Constantine 1).

Année universitaire
2021 - 2022

Remerciements

J'adresse, en premier, ma reconnaissance à notre **DIEU** tout puissant, de m'avoir permise d'arriver à ce stade de savoir Louange à **ALLAH**, Seigneur des mondes :

Papa et maman,

La pudeur m'a longtemps empêchée de vous adresser ces quelques lignes. Et puis le temps passe, il file même, et je réalise à quel point il est important de vous dire MERCI pour tout.

Mes sincères remerciements à monsieur HAMIDECHI Mohamed Abdelhafid pour m'avoir encadrée dans ce travail. Le regard critique, juste et avisé qu'il a porté sur mes travaux ne peut que m'encourager à être plus perspicace et engagée dans mes recherches.

Ma gratitude va au Dr DAAS Skandar (Président du jury) et au Dr BOULAHROUF Khaled (Examinateur) qui me font l'honneur d'évaluer mon travail et qui ont fort aimablement accepter cette lourde tâche compte tenu de leurs occupations respectives surtout en cette fin d'années.

Merci à tous mes professeurs de la spécialité Bio-informatique et à mes amies de Master de Bioinformatique pour leur enthousiasme et leur solidarité.

Enfin, je remercie toutes les personnes qui ont participé de près ou de loin à la réalisation de ce travail.

Dédicaces

Je dédie ce travail à :

Mon très cher père et ma très chère mère qui ont toujours cru en moi, pour les sacrifices et tous les efforts qu'il ont fait pour faire de moi la personne que je suis aujourd'hui. Que **DIEU** les récompense pour tous leurs bienfait et que les mots n'arriveront jamais à les décrire.

Mes frères Khaled et Zaki et mes belles sœurs Nihed et Nedjoua pour leur soutien, leur encouragement, leur affection et leur patience.

Ma grand-mère Mamani à qui je souhaite une bonne santé et une longue vie.

Toute ma famille.

Tous mes amis.

Résumé

IL existe actuellement sept classes d'enzymes après ajout de la famille des Translocases. Les hydrolases sont une classe qui utilisent l'eau pour rompre une liaison chimique, ce qui entraîne généralement la division d'une molécule plus grosse en molécules plus petites. L'objectif principal de ce travail est de proposer un système de classification d'enzymes basé sur leur structure primaire en utilisant une approche bioinformatique de Deep Learning. Les résultats de notre travail ont été concluants et ont abouti à une valeur Accuracy égale à 97,1 % ce qui est acceptable relativement aux publications internationales. Cependant, il est plus intéressant de travailler sur un nombre de séquences plus important afin d'aboutir à des conclusions plus pertinentes et plus précises. Ce travail a permis de mettre en évidence l'apport de cette approche et à améliorer la précision de la classification des protéines.

Mots clés : Hydrolases, Prédiction, Intelligence Artificielle, Deep Learning.

Abstract

There are currently seven classes of enzymes after adding the Translocase family. Hydrolases are a class that use water to break a chemical bond, which usually causes a larger molecule to split into smaller molecules. The main objective of this work is to propose an enzyme classification system based on their primary structure using a Deep Learning bioinformatics approach. The results of our work were conclusive and led to an Accuracy value equal to 97.1%, which is acceptable relative to international publications. However, it is more interesting to work on a larger number of sequences in order to reach more relevant and more precise conclusions. This work made it possible to highlight the contribution of this approach and to improve the precision of the classification of proteins.

Keywords: Hydrolases, Prediction, Artificial Intelligence, Deep Learning.

ملخص

يوجد حاليًا سبع فئات من الإنزيمات بعد إضافة عائلة .Translocase الهيدرولازات هي فئة تستخدم الماء لكسر رابطة كيميائية ، والتي عادة ما تتسبب في انقسام جزيء أكبر إلى جزيئات أصغر .الهدف الرئيسي من هذا العمل هو اقتراح نظام تصنيف الإنزيمات بناءً على هيكلها الأساسي باستخدام نهج التعلم العميق للمعلومات الحيوية .كانت نتائج عملنا حاسمة وأدت إلى قيمة Accuracy تساوي 97.1٪ ، وهو أمر مقبول بالنسبة للمنشورات الدولية .ومع ذلك ، فمن المثير للاهتمام العمل على عدد أكبر من التسلسلات من أجل الوصول إلى استنتاجات أكثر صلة وأكثر دقة .أتاح هذا العمل تسليط الضوء على مساهمة هذا النهج وتحسين دقة تصنيف البروتينات

.الكلمات المفتاحية : Hydrolases :، التنبؤ ، الذكاء الاصطناعي ، التعلم العميق.

TABLE DES MATIERES

INTRODUCTION

Les données biomoléculaires sont de plus en plus nombreuses et l'on assiste d'ailleurs à l'ère des BigData ou données massives nécessitant l'implication obligatoire de systèmes informatiques (Algorithmes) aidant à enregistrer, à organiser et surtout à traiter cette masse de données. La situation est devenue donc problématique face à ce flux de données de plus en plus en croissance exponentielle. A titre informatif[1], la version 2020_01 du 26 février 2020 d'UniProtKB/Swiss-Prot contient 561 911 entrées de séquences protéiques, comprenant un total de 202 173 710 acides aminés extraits de 270 528 références.

La contribution de la bioinformatique dans le traitement de ces données n'est plus à démontrer. Grâce à de multiples programmes informatiques et la mise en pratique de pipelines spécifiques, la tâche est devenue aisée pour les biologistes utilisateurs de ces données métagénomiques et métaprotéomiques.

L'objectif de notre travail se situe dans cette perspective d'Intelligence Artificielle en proposant un pipeline par approche Deep Learning (Apprentissage Approfondi) de traitement des données (séquences) protéiques et enzymatiques, de la base de données UniProt du portail Expasy, afin de classer les enzymes de la superfamille des hydrolases, et ce ne tenant compte que de la seule séquence en acides aminés (structure primaire) téléchargeable au format Fasta de SwissProt/UniProt et de EMBL (European Moleculary Biology Laboratory).

[1] https://ftp.uniprot.org/pub/databases/uniprot/previous_major_releases/release-2020_01/knowledgebase/UniProtKB_SwissProt-relstat.html

Partie 1

RECHERCHE BIBLIOGRAPHIQUE

Chapitre 1

CLASSIFICATION ET NOMENCLATURE DES ENZYMES

1. SYSTEME DE CLASSIFICATION

1.1 Définition d'une enzyme

Une enzyme est une protéine qu'on peut trouver chez tous les organismes Vivants, elle est responsable de la catalysassions de différentes réactions chimiques. Il existe un nombre très important d'enzymes et en découvre encore aujourd'hui.

Du point de vue informatique, nous considérerons une enzyme comme un triplé, de la même manière que pour toute autre protéine. Une protéine peut être considérée à plusieurs niveaux :

1. une séquence d'acides aminés (structure primaire),
2. une structure formée par cette séquence repliée dans l'espace (structure spatiale 3D),
3. une fonction associée [1, 2].

1.2 Nomenclature et classification des enzymes

Le nom de la majorité des enzymes a été formé à partir d'un suffixe ou plus exactement à partir de l'ajout d'un suffixe « ase » au nom de leur substrat ou à un mot ou une phrase décrivant son activité.

La nomenclature officielle des enzymes a été créée par la commission des enzymes (EC). La nomenclature officielle nous permet d'avoir une référence unique pour chaque enzyme. Celles-ci peuvent être classées en sept grandes catégories selon leur réaction catalysée [3 4 5] :

1) Oxydoréductases : qui catalysent des transferts d'électrons et de proton d'un donneur a un récepteur
2) Transférases : qui catalysent les transferts de groupement
3) Hydrolases qui catalysent des réactions d'hydrolase
4) Lyases : qui catalyse l'addition de groupe s a des liens doubles ou l'inverse
5) Isomérases : qui catalysent le transfert de groupes dans une même molécule pour produire des formes isomères

6) Ligases qui forment des liens C-C, C-S, C-O, C-N lors de réactions de condensation couplées à l'utilisation de l'ATP
7) Translocases : Ces enzymes catalysent le mouvement des ions ou des molécules à travers les membranes ou leur séparation à l'intérieur des membranes, la réaction est désignée comme un transfert du «côté 1» au «côté 2» parce que les désignations «in» et «out», qui étaient auparavant utilisées , peut être ambigu. Les sous-classes désignent les types de composants transférés et les sous-sous-classes indiquent les processus de réaction qui fournissent la force motrice de la translocation.

Enfin, une enzyme est une molécule (protéine ou même un ARN dans le cas des ribozymes) qui permet d'abaisser l'énergie d'activation d'une réaction et d'accélérer jusqu'à des millions de fois les réactions chimiques du métabolisme énergétique cellulaire. Ces réactions biochimiques se déroulent dans les milieux intra et extracellulaires sans aucune modification de la structure initiale de l'enzyme d'où son appellation de biocatalyseur. Les enzymes agissent à faibles concentrations. La première enzyme fut découverte par Anselme Payen et Jean-François Persoz en 1833.

La nomenclature des enzymes n'est pas standardisée, elle se compose le plus souvent d'un radical proche du substrat ou du produit de la catalyse suivi d'un suffixe.

Une enzyme, comme toute protéine, est synthétisée par les cellules vivantes à partir des informations codées dans l'ADN ou dans l'ARN.

Il existe deux grandes catégories d'enzymes :

- Les enzymes purement protéiques (qui ne sont constituées que d'acides aminés) :
  Les « holoenzymes »
- Les enzymes en deux parties :
  Une partie protéique (« l'apoenzyme ») et une partie non protéique (« le cofacteur »), appelées « hétéroenzymes ».

Dans ce chapitre nous présenterons la classification et la nomenclature des enzymes ; en particulier la classe des hydrolases (structure et fonction).

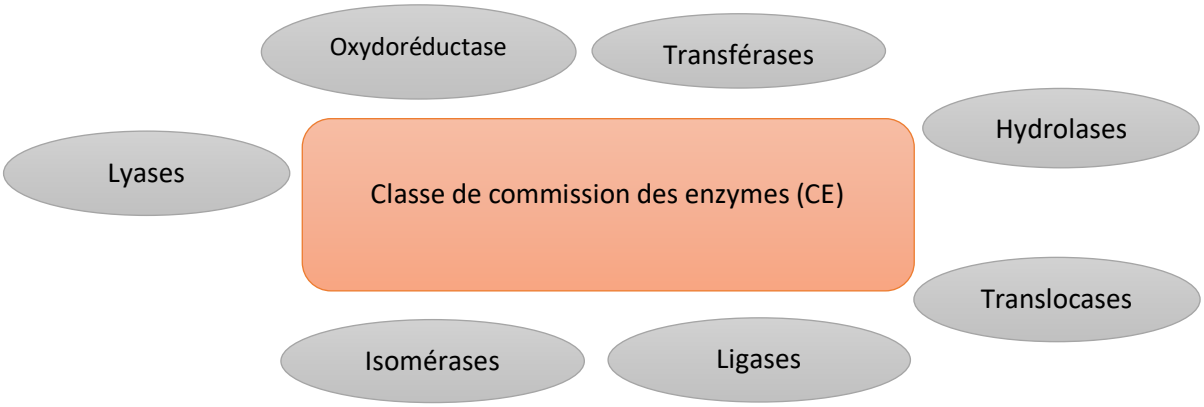


Figure 1 : Classification basique (selon UIPAC) des sept familles d'enzymes

2 DESCRIPTION DE LA CLASSE DES HYDROLASES

2.1 Définition de l'hydrolase

L'hydrolase est une classe d'enzymes hydrolytiques couramment utilisées comme catalyseurs biochimiques qui utilisent l'eau comme donneur de groupe hydroxyle lors de la dégradation du substrat. En termes simples, une hydrolase est une enzyme qui catalyse l'hydrolyse d'une liaison chimique dans les biomolécules. Ceci, à son tour, divise une grosse molécule en deux plus petites. Les hydrolases sont donc importantes pour l'environnement car elles digèrent les grosses molécules en petits fragments pour la synthèse de bio polymères ainsi que pour la dégradation des toxines [6].

Les hydrolases effectuent d'importantes réactions de dégradation dans l'organisme. Elles clivent les grosses molécules en fragments plus petits utilisés pour la synthèse, l'excrétion de déchets ou comme sources de carbone pour la production d'énergie. Ceux-ci sont impliqués dans les processus de digestion, de transport, d'excrétion, de régulation et de signalisation, etc. ; par exemple, les enzymes digestives comme le cholinestérase, la carboxylestérase, les hydrolases lysosomales, etc. L'hydrolase exprimée par *Lactobacillus spp*. dans l'intestin humain pourrait stimuler le foie à sécréter des sels biliaires qui facilitent la digestion des aliments [7].

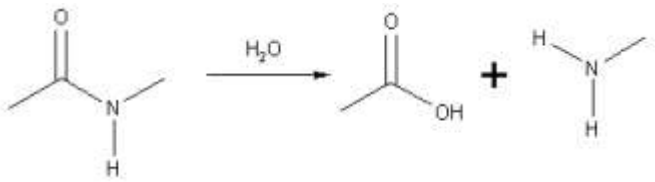


Figure 2 : coupure d'une liaison à l'aide d'une molécule d'eau

2.2 Structure et fonction des hydrolases

Étant la classe d'enzymes la plus vaste et la plus diversifiée, les hydrolases offrent une opportunité d'explorer la diversité conformationnelle qui constitue la base de leurs fonctions biologiques différentielles. Comme toutes les enzymes, les hydrolases (généralement extracellulaires) sont des protéines qui possèdent un site actif permettant la réaction enzymatique et un site de reconnaissance des molécules cibles, assurant la spécificité de la réaction [8].

Ces enzymes hydrolysent (figure 1) une gamme de groupes fonctionnels, notamment des esters, des amides et des nitriles, et nombre de ces enzymes peuvent être commodément actionnées dans le sens inverse pour permettre la formation de ces mêmes groupes fonctionnels. Elles catalysent le clivage d'une liaison covalente à l'aide d'eau. Les types d'hydrolase comprennent les estérases, telles que les phosphatases, qui agissent sur les liaisons ester, et les protéases ou les peptidases qui agissent sur les liaisons amide dans les peptides.

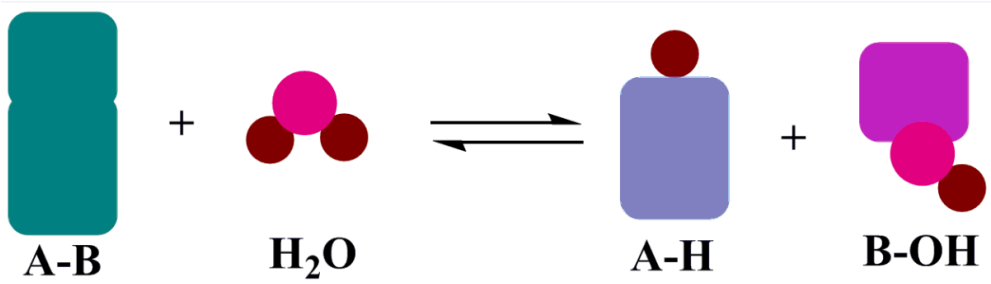


Figure 3 : Exemple basique d'une hydrolyse catalysée par une hydrolase[2]

[2] https://www.creative-enzymes.com/resource/hydrolase-introduction_21.html

2.3 Classification des hydrolases

Outre les noms communs donnés à certaines hydrolases, il existe des nomenclatures systématiques pour nommer ces enzymes.

2. 3.1 Basé sur les numéros de commission des enzymes (CE)

Les hydrolases appartiennent à la classe d'enzymes 3 (EC 3) et sont en outre classées en fonction du type de liaison qu'elles clivent [9].

Tableau 1 : Types de liaisons clivées par les hydrolases

Sous-classe (hydrolase agissant sur)	Exemple de sous-sous-classe	Exemple d'enzyme
3.1 Liaisons ester (estérases)	3.1.1 Lipases	3.1.1.3 Triacylglycérol lipase
3.2 Sucres	3.2.1 Glycosidase	3.2.1.1 α-amylase
3.3 Liaison éther	3.3.2 Ether hydrolase	3.3.2.6 Leucotriène-A4 hydrolase
3.4 Liaison peptidiques (peptidases)	3.4.21 Serine endopeptidase	3.4.21.1 Chymotrypsine
3.5 Liaison CN (autre que les peptidique)	3.5.1 Dans les amides linéaires	3.5.1.1 Asparaginase
3.6 Anhydrides d'acide	3.6.1 Dans les anhydrides contenant du phosphore	3.6.1.1 Diphosphatase inorganique
3.7 Obligation CC	3.7.1 Dans les substances cétoniques	3.7.1.1 Oxaloacétase
3.8 Liaisons halogénures	3.8.1 Dans les complexe CX	3.8.1.1 Alkylhalidase
3.9 Obligation PN	3.9.1 Sur les Obligations PN	3.9.1.1 Phosphoamidase
3.10 Obligation SN	3.10.1 Sur les Obligations SN	3.10.1.1 N-sulfoglucosamine sulfohydrolase
3.11 Obligation CP	3.11.1 Sur les Obligations CP	3.11.1.1 Phosphonoacétaldéhyde hydrolase
3.12 Obligation SS	3.12.1 Sur les Obligations SS	3.12.1.1 Trithionate hydrolase
3.13 Obligation CS	3.13.1 Sur les Obligations CS	3.13.1.1 UDP-Sulfoquinovose synthétase

2. 3.2 Basé sur le site actif

La géométrie du site actif des différentes hydrolases est différente, malgré la même méthode catalytique, c'est-à-dire l'hydrolyse. Ainsi, un système avec H : classification hiérarchique des hydrolases, C : catalyse et S site (HCS) a été proposé qui est basé sur les acides aminés impliqués dans la catalyse [11 ,12]. La relation entre une classe et sa sous-classe dans la hiérarchie est que le site catalytique de la sous-classe raffine le site catalytique de la classe

de base. La sérine hydrolase comme l'estérase, avec une dyade catalytique inhabituelle Ser-His, appartient à la classe S.01(sérine hydrolases avec dyade Ser-His) tandis que les hydrolases, telles que la trypsine ou la subtilisine, sont en outre classées dans la sous-classe S.01.01(hydrolases avec la triade Ser-His-Asp/Glu), c'est-à-dire que la sous-classe contient tous les résidus de la classe de base et quelques autres. Actuellement, seules les hydrolases sont incluses dans une telle classification car ce sont les enzymes les plus étudiées et les plus abondantes.

2.4 Fonction des hydrolases

Il existe de nombreux types d'hydrolases en fonction du type de liaison covalente qu'elles hydrolysent. Elles utilisent toutes une molécule d'eau pour leur réaction et portent généralement le nom du substrat :

- Les estérases (coupent des liaisons ester), comme les nucléases (qui coupent les acides nucléiques).
- Les protéases coupent des liaisons peptidiques.
- Les phosphatases coupent des liaisons impliquant un phosphate, etc.

Elles sont donc impliquées dans un grand nombre de processus cellulaires et physiologiques, permettant aussi bien la digestion que la régulation de l'expression des gènes [11].

2.5 Exemple d'hydrolases

2.5.1 Glycoside Hydrolase : L'hydrolyse des liaisons glycosidiques dans les sucres complexes est catalysée par des glycoside hydrolases (également appelées glycosidases ou glycosyl hydrolases). Ce sont des enzymes extrêmement courantes qui jouent divers rôles dans la nature, notamment la dégradation de la cellulose (cellulase), de l'hémicellulose (hémicellulose) et de l'amidon (amylase), les stratégies de protection antibactérienne (par exemple, le lysozyme), les mécanismes de pathogenèse (par exemple, les neuraminidases virales), et une fonction cellulaire normale (par exemple, rognage des mannosidases impliquées dans la biosynthèse des glycoprotéines N-liées). Les glycosidases, avec les glycosyltransférases, sont les enzymes clés impliquées dans la synthèse et la rupture des liaisons glycosidiques [11].

2.5.2 Hydrolase d'ester de cholestérol : Une stérol estérase est une enzyme qui catalyse la réaction chimique en enzymologie.

Ester de stérol + stérol H_2O +acide gras

Ainsi, l'ester de stérol et H_2O sont les deux substrats de l'enzyme, tandis que le stérol et l'acide gras sont les deux produits de l'enzyme [11].

2.5.3 Sérine Hydrolase : Les sérine hydrolases sont l'un des plus grands groupes d'enzymes connus, représentant environ 200 enzymes ou 1% des gènes du protéome humain. La présence d'une sérine nucléophile dans le site actif, qui est utilisée pour l'hydrolyse du substrat, est une caractéristique distinctive de ces enzymes. Via cette sérine, la catalyse commence par la formation d'un intermédiaire acyl-enzyme, suivie de la saponification de l'intermédiaire par l'eau/hydroxyde et de la régénération de l'enzyme. La sérine nucléophile de ces hydrolases, contrairement aux autres sérines non catalytiques, est normalement activée par un relais protonique impliquant une triade catalytique constituée de la sérine, d'un résidu acide (par exemple aspartate ou glutamate) et d'un résidu simple (généralement histidine) [12.13]

Chapitre 2

L’APPRENTISSAGE APPROFONDI

L'IA est une science dans laquelle les modèles mathématiques et l'informatique sont omniprésents (quel que soit le domaine d'application). Dans le cas de la biologie, l'IA s'appuie largement sur les ressources et les techniques de la bioinformatique. De plus en plus d'applications de l'IA sont développées, en particulier en robotique, en électronique, en communication et surtout dans le traitement de texte et dans la reconnaissance et l'analyse d'images et de la voix.

Le but de l'IA est double. D'une part, elle s'attache à résoudre des problèmes qui relèvent d'activités humaines ou animales de nature variée : Perception, planification, interprétation de données, diagnostic, prise de décision, compréhension du langage, conception. D'autre part, elle cherche à mieux comprendre et modéliser l'intelligence. Elle se rapproche ainsi des sciences cognitives dont elle s'inspire par ailleurs pour la conception de modèles (mémoire, raisonnement, apprentissage).

La nécessité de restreindre l'activité à un champ d'application limité et de s'appuyer sur des connaissances de natures diverses est apparue rapidement en IA. Cette approche symbolique de l'IA a donné lieu aux systèmes à base de connaissances.

Ce chapitre présente les différents model de l'IA, ainsi que les méthodes d'apprentissage associées. Nous allons partir du général qui est l'IA ou machine Learning jusqu'au plus précis qui l'apprentissage profond ou approfondi.

1 L' Intelligence Artificielle

1.1 Définition

Il n'existe pas de définition universelle de l'IA et chaque spécialiste la définit selon l'école qu'il suit. Yann Le Cun qui définit l'IA comme « un ensemble de techniques permettant à des machines d'accomplir des tâches et de résoudre des problèmes normalement réservés aux humains et à certains animaux ».

De plus, l'IA peut être définit comme un ensemble d'algorithmes relier à une machine qui a des capacités d'analyse et de décision qui lui permettent de s'adapter intelligemment aux situations en faisant des prédictions à partir de données déjà acquises [14].

Autrement dit, L'IA a pour objectif de construire des dispositifs simulant les processus cognitifs humains. On peut dire que l'IA a pour objectif de mimer les fonctions mentales de l'humain [14].

Suite à cette définition on peut citer trois types de l'IA :

- IA faible qui peut être inférieure ou équivalente à l'humain.
- IA moyenne qui peut être supérieure à la plupart des hommes.
- IA forte qui peut être supérieure à tout humain.

2. APPRENTISSAGE AUTOMATIQUE

2.1 Définition

La définition de l'apprentissage automatique selon Wikipédia (septembre 2020) est :

« L'apprentissage automatique (en anglais machine Learning, littéralement « apprentissage machine ») ou apprentissage statistique est un champ d'étude de l'IA qui se fonde sur des approches mathématiques et statistiques pour donner aux ordinateurs la capacité d'apprendre à partir de données, c'est à-dire d'améliorer leurs performances à résoudre des tâches sans être explicitement programmés pour chacune. Plus largement, il concerne la conception, l'analyse, l'optimisation, le développement et l'implémentation de telles méthodes. » [15]

Plus simplement l'apprentissage automatique ou machine Learning (ML) est une application courante de l'IA basée sur l'idée que nous devrions simplement donner aux machines l'accès à des données pour qu'elles apprennent par elles-mêmes et nous donner des résultats à la fin [16].

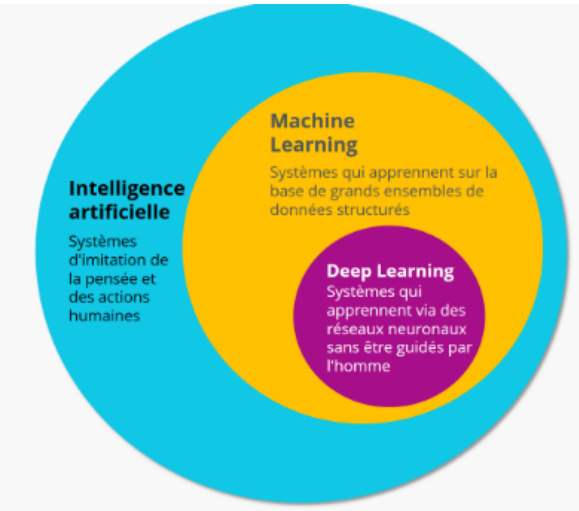


Figure 4 : la relation entre l'IA l'apprentissage automatique et l'apprentissage profond

2.2 L’apprentissage automatique, comment se fait-il ?

L’apprentissage se divise en deux parties :

2.2.1 La partie d’apprentissage : Dans cette première partie, notre machine commence par construire un modèle, qui sera la base du système de son raisonnement sans avoir des règles ou des limites. Pour cela on doit fournir à la machine des exemples qu’elle va à la suite analyser pour qu’elle puisse comprendre la logique du modèle suite à ça, l’intégration de l’algorithme de transformation.

2.2.2 La partie de prédiction : Après que le raisonnement et l’algorithme feront partie intégrée de la machine, le programme de l’apprentissage automatique déterminera la finalité de la situation donnée.

2.3 Types de l’apprentissage automatique

L’apprentissage automatique se divise en trois types :

- L’apprentissage supervisé
- L’apprentissage non supervisé
- L’apprentissage par renforcement

2.3.1 L’apprentissage supervisé :

C’est un système qui apprend à classer les données selon un modèle de classification.

Le processus se passe en deux phases :

La première phase (dite d'apprentissage), consiste à déterminer un modèle à partir des données étiquetées par un expert.

La seconde phase (dite de test) consiste à prédire l'étiquette d'une nouvelle donnée, en se basant sur le modèle préalablement appris. 17

Exemple : Dans notre cas la machine peut apprendre à reconnaitre la séquence d’hydrolase après lui avoir montré des milliers de cette dernière donc elle peut juger si c’est hydrolase ou non hydrolase.

2.3.2 apprentissage non supervisé :

La machine apprend par elle-même. Mais le terme d'apprentissage autonome reste très relatif. La machine peut faire des regroupements et donc de réaliser des classifications, elle n'est pas capable de définir par elle-même les différentes formules, car elle n'a pas conscience des données dont elle a la charge d'en faire l'apprentissage.

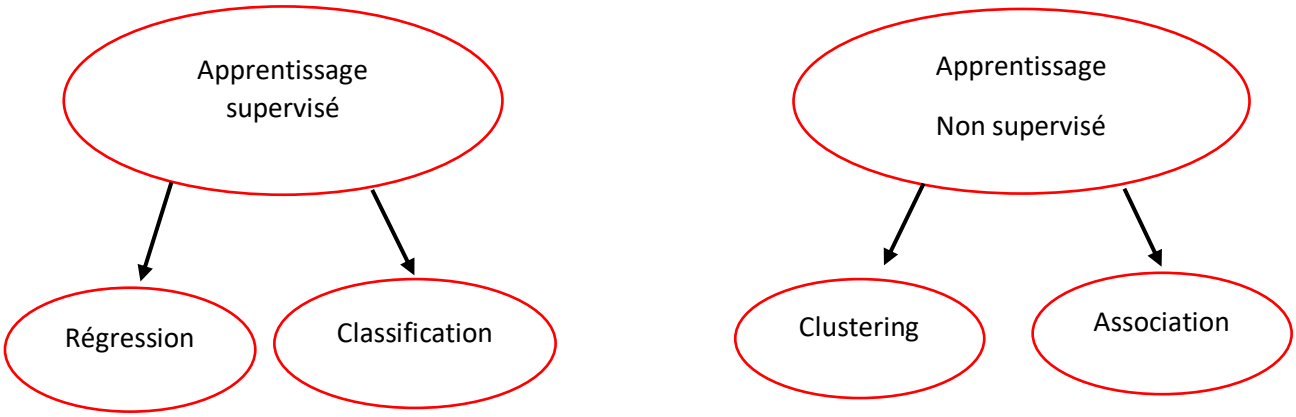


Figure 5 : apprentissage machine

Quand on dit apprentissage supervisé, on vise que les données étiquetées par l'homme sont transmises à un algorithme d'apprentissage machine pour apprendre à l'ordinateur une fonction, comme la reconnaissance d'une image.

Dans l'apprentissage non supervisé, les données non marquées sont transmises à un algorithme d'apprentissage automatique, qui lui a son tour tente de trouver une structure cachée aux données, comme l'identification d'une séquence, et puis c'est à l'opérateur de déduire un sens [17].

2.3.3 Apprentissage par renforcement :

On peut l'aligner avec l'apprentissage non supervisé, dont la machine apprend par elle-même, mais on peut considérer que le mode d'apprentissage est différent, le système s'appuie sur des récompenses ou des sanctions pour ajuster ses actions.

Plus simplement c'est L'évaluation de grandes quantités de données et la dérivation des données importantes.

2.4 les champs d'application de l'apprentissage automatique :

Le machine Learning ou apprentissage automatique comprend tous les secteurs d'activité, que sa sois l'industrie, le commerce, la santé et les sciences de la vie, le tourisme et l'hôtellerie, les services financiers, l'énergie, les matières premières et les services publics. Domaines d'utilisation [18] :

- Secteur industriel : maintenance prédictive et surveillance des équipements
- Commerce: upselling et marketing cross-canal
- Santé et sciences de la vie : diagnostic et réduction des risques
- Tourisme et hôtellerie : tarification dynamique
- Services financiers : analyse et régulation des risques
- Énergie : optimisation de la demande et de l'approvisionnement

3 APPRENTISSAGE APPROFONDI

3.1 définition

L'apprentissage profond ou Deep Learning (DL) est un ensemble de méthodes d'apprentissage automatique (ML), supervisés ou non supervisés, dans le but de modéliser avec un haut niveau après une récolte de données, en utilisant un modèle de réseaux de neurones artificielles (ANN) multicouche à d'énormes quantités de données.

L'apprentissage profond a permis des progrès importants dans les domaines :

- De l'analyse des signaux visuels ou sonores.
- De la reconnaissance faciale, de la reconnaissance vocale.
- De la vision par ordinateur MV.

4. PRINCIPE DU PERCEPTRON

C'est en 1957 que le Perceptron fut inventé par Frank Rosenblatt au laboratoire aéronautique de Cornell. En se basant sur les premiers concepts de neurones artificiels, il proposa la " règle d'apprentissage du Perceptron ".

Un Perceptron est un neurone artificiel, donc une unité de réseau de neurones. Il effectue des calculs pour détecter des caractéristiques ou des tendances dans les données d'entrée. Il s'agit d'un algorithme pour l'apprentissage supervisé de classificateurs binaires. C'est cet algorithme qui permet aux neurones artificiels d'apprendre et de traiter les éléments d'un ensemble de données.

Le Perceptron joue un rôle essentiel dans les projets de Machine Learning. Il est utilisé pour classifier les données, ou en guise d'algorithme permettant de simplifier ou de superviser les capacités d'apprentissage de classificateurs binaires.

En sachant que l'apprentissage supervisé consiste à apprendre à un algorithme à réaliser des prédictions. Pour y parvenir, on nourrit l'algorithme à l'aide de données déjà étiquetées correctement.

4.1 La règle d'apprentissage du perceptron

Selon la Perceptron Learning Rule (règle d'apprentissage du Perceptron), l'algorithme apprend automatiquement les coefficients de poids optimaux. Les caractéristiques des données d'entrée sont multipliées par ces poids, afin de déterminer si un neurone s'allume ou non.

Le Perceptron reçoit de multiples signaux d'entrée. Si la somme des signaux excède un certain seuil, un signal est produit ou au contraire aucun résultat n'est émis.

Dans le cadre de la méthode d'apprentissage supervisé de Machine Learning, c'est ce qui permet de prédire la catégorie d'un échantillon de données. Il s'agit donc d'un élément essentiel.

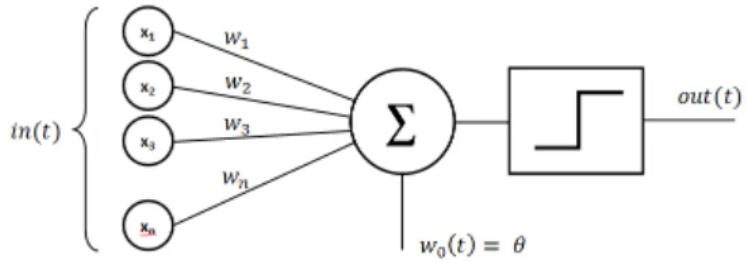


Figure 6 : formule du perceptron

4.2 Fonction du perceptron

Le Perceptron est une fonction mathématique. Les données d'entrée (x) sont multipliées par les coefficients de poids (w). Le résultat produit est une valeur.

Cette valeur peut être positive ou négative. Le neurone artificiel s'active si la valeur est positive. Il ne s'active donc que si le poids calculé des données d'entrée dépasse un certain seuil.

Le résultat prédit est comparé avec le résultat connu. En cas de différence, l'erreur est rétro-propagée afin de permettre d'ajuster les poids.

5 LES RESEAUX DE NEURONES ARTIFICIELS

5.1 Définition

Les réseaux de neurones, les réseaux de neurones artificiels sont des imitations simples des fonctions d'un neurone dans le cerveau humain pour résoudre des problématiques d'apprentissage de la machine (Machine Learning).

Le neurone (perceptron), inventé par Rosenblatt (1957) est un type de réseau de nueurones le plus simple, fonction avec une entrée et une sortie suite à une pondération de la somme des entrées (wixi) et l'application d'une fonction d'activation (ou de transfert). C'est une unité exprimée par une fonction généralement sigmoïde (figure 7), ou à seuil (Heaviside), ou linéaire, ou encore une fonction Tanh (ou Tangente hyperbolique)

$$f(x) = \frac{1}{1 + e^{-x}}$$

Figure 7 : Fonction sigmoïde d'une unité d'un neurone prenant en entrée des signaux xi et retournant une sortie y ou f(x).

Le réseau de neurones est habituellement défini par sa structure caractérisée par le nombre de couches, le nombre de nœuds par couche et le nombre de s orties

Les domaines d'application des réseaux neuronaux caractérisés par une relation entrée-sortie de la donnée d'information :

- La reconnaissance d'image
- Les classifications de textes ou d'images
- Identification d'objets
- Prédiction de données
- Filtrage d'un set de données

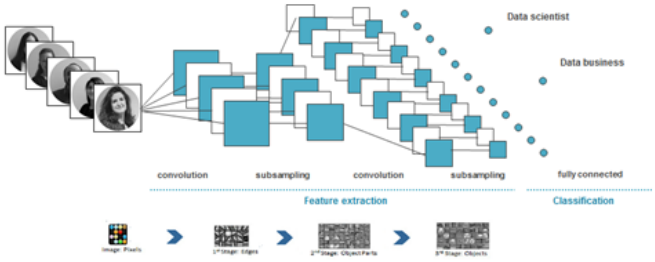


Figure 8 : réseau de neurone classique pour reconnaissance d'image

5.2 Types de réseaux neuronaux :

5.2.1 Les réseaux de neurones feed-forwarded

Feed-forwarded fait tout simplement référence à la procédure du traitement de la donnée par le réseau neuronal. En effet, feed-fowarded (propagation avant) signifie tout simplement que la donnée traverse le réseau d'entrée à la sortie sans retour en arrière de l'information.

Ce trouve uniquement dans la famille des réseaux à propagation avant, on distingue les réseaux monocouches (perceptron simple) et les réseaux multicouches (perceptron multicouche).

Le perceptron simple est dit simple parce qu'il ne dispose que de deux couches ; la couche en entrée et la couche en sortie. Le réseau est déclenché par la réception d'une information en entrée. Le traitement de la donnée dans ce réseau se fait entre la couche d'entrée et la couche de sortie qui sont toutes reliées entre elles. Le réseau intégral ne dispose ainsi que d'une matrice de poids. Le fait de disposer d'une seule matrice de poids limite le perceptron simple à un classificateur linéaire permettant de diviser l'ensemble d'informations obtenues en deux catégories distingues.

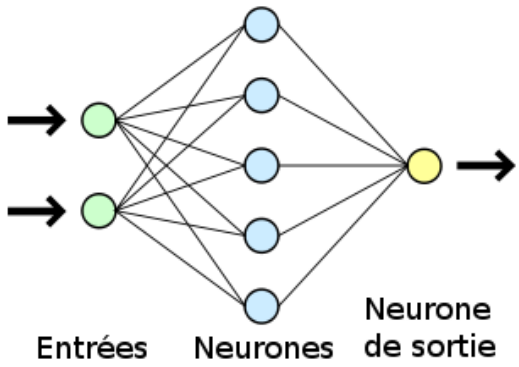


Figure 9: Distribution monocouche

Le perceptron multicouche se structure de la même façon. L'information entre par une couche d'entrée et sort par une couche de sortie. À la différence du perceptron simple, le perceptron multicouche dispose entre la couche en entrée et la couche en sortie une ou plusieurs couches dites « cachées ». Le nombre de couches correspond aux nombres de matrices de poids dont disposent le réseau. Un perceptron multicouche est donc mieux adapté pour traiter les types de fonctions non-linéaires.

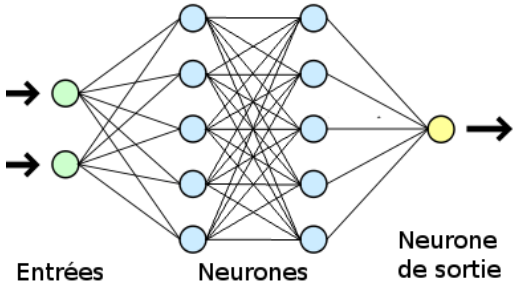


Figure 10 : Distribution multicouches

5.2.2 Les réseaux de neurones récurrents :

Les Réseaux de Neurones récurrents traitent l'information en cycle. Ces cycles permettent au réseau de traiter l'information plusieurs fois en la renvoyant à chaque fois au sein du réseau.

La force des Réseaux de neurones récurrents réside dans leur capacité de prendre en compte des informations contextuelles suite à la récurrence du traitement de la même information. Cette dynamique auto-entretient le réseau.

5.2.3 Les réseaux de neurones à résonance :

L'appellation du réseau neuronal fait encore une fois référence à son fonctionnement. En effet, au sein des réseaux de neurones à résonance, l'activation de tous les neurones est renvoyée à tous les autres neurones au sein du système. Ce renvoi provoque des oscillations, d'où la raison du terme résonance.

5.2.4 Les réseaux de neurones auto-organisés :

Les Réseaux de neurones auto-organisés sont surtout adaptés pour le traitement de d'informations spatiales. Par des méthodes d'apprentissage non-supervisé, les réseaux neuronaux auto-organisés sont capables d'étudier la répartition de données dans des grands espaces comme par exemple pour des problématiques de clustérisation ou de classifications.

Le modèle le plus connu de ce type de réseaux de neurones est sans doute la carte auto-organisatrice de Kohonen :

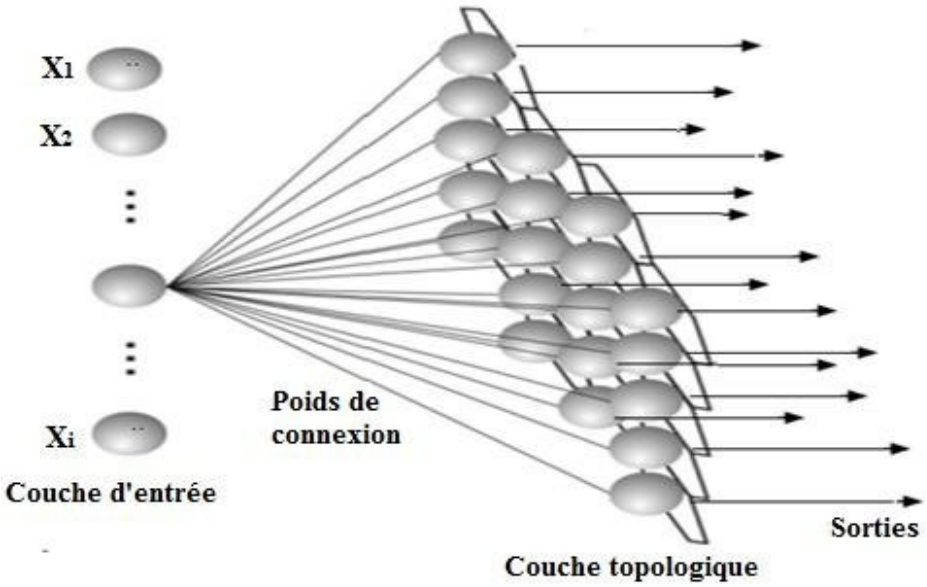


Figure 11 : Model de Kohonen

Partie 2

Partie expérimentale

MATERIEL ET METHODES

MATERIEL

1.1. Données biologiques

Les données utilisées sont des données protéiques téléchargées à partir de la base de données Uniprot. Ces données sont téléchargeables en un seul fichier sous format TSV (Tabulation-Separated Values). La taille du fichier est de 209 Mo, téléchargé en date du 01 juin 2022. Ce dataset est formé de cinq colonnes : Entry, Status, Length, EC number, Sequence. Il contient un total de 567 483 séquences protéiques.

1.2. Outils et bibliothèques informatiques

1.2.1. Environnement de travail

- Python : Python est un langage de programmation open source créé par Guido van Rossum en 1991 multiplateformes et orienté objet. Grâce à des bibliothèques spécialisées, Python s'utilise pour de nombreuses situations comme le développement logiciel, l'analyse de données, ou la gestion d'infrastructures « futura-sciences. ».

- Anaconda : Plateforme la plus populaire et la plus fiable au monde pour la science des données, l'apprentissage automatique et l'IA avec les langages Python.

La Distribution qui inclut le langage Python contient des bibliothèques de calcul scientifique : Jupyter, spyder, jupyterlab… (« définitions-digital. »)

- Jupyter notebook : Jupyter Notebook est une application Web open source que vous pouvez utiliser 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. Jupyter Notebooks est un projet dérivé du projet IPython, qui avait lui-même un projet IPython Notebook. Le nom, Jupyter, vient des principaux langages de programmation pris en charge : Julia, Python et R. Jupyter est livré avec le noyau IPython, qui permet d'écrire des programmes en Python, mais, actuellement, il existe plus de 100 autres noyaux utilisables.

1.2.2. Bibliothèques python

- Pandas : Pandas DataFrame est une structure de données tabulaire bidimensionnelle variable en taille, potentiellement hétérogène, avec des axes étiquetés (lignes et colonnes). Une trame de données est une structure de données bidimensionnelle, c'est-à-dire que les données sont alignées de manière tabulaire en lignes et en colonnes. Pandas DataFrame se forme de trois composants principaux, les données, les lignes et les colonnes. (« geeksforgeeks.org »).
- NumPy : Un projet open source permettant le calcul numérique avec Python. Il a été créé en 2005, en s'appuyant sur les premiers travaux des bibliothèques Numerical et Numarray. Il est toujours 100% open source, gratuit pour les utilisateurs (« NumPy »).
- Sklearn : Scikit-learn est une bibliothèque clé pour le langage de programmation Python qui est généralement utilisé dans les projets d'apprentissage automatique. Scikit-learn se concentre sur les outils d'apprentissage automatique, y compris les algorithmes mathématiques, statistiques et à usage général qui constituent la base de nombreuses technologies d'apprentissage automatique. En tant qu'outil gratuit, Scikit-learn est extrêmement important dans de nombreux types de développement d'algorithmes pour l'apprentissage automatique et les technologies associées (« techopedia »).
- Tensorflow : TensorFlow est un framework open source, développé par des chercheurs de Google pour exécuter l'apprentissage automatique, l'apprentissage en profondeur et d'autres charges de travail d'analyse statistique et prédictive. Comme des plates-formes similaires, il est conçu pour rationaliser le processus de développement et d'exécution d'applications d'analyse avancées pour les utilisateurs tels que les scientifiques des données, les statisticiens et les modélisateurs prédictifs. Le logiciel TensorFlow gère les ensembles de données qui sont organisés en nœuds de calcul sous forme de graphe. Les arêtes qui relient les nœuds d'un graphe peuvent représenter des vecteurs ou des matrices multidimensionnelles, créant ce que l'on appelle des tenseurs. Parce que les programmes TensorFlow utilisent une architecture de flux de données qui fonctionne avec des résultats intermédiaires généralisés

des calculs, ils sont particulièrement ouverts aux applications de traitement parallèle à très grande échelle, les réseaux de neurones étant un exemple courant.

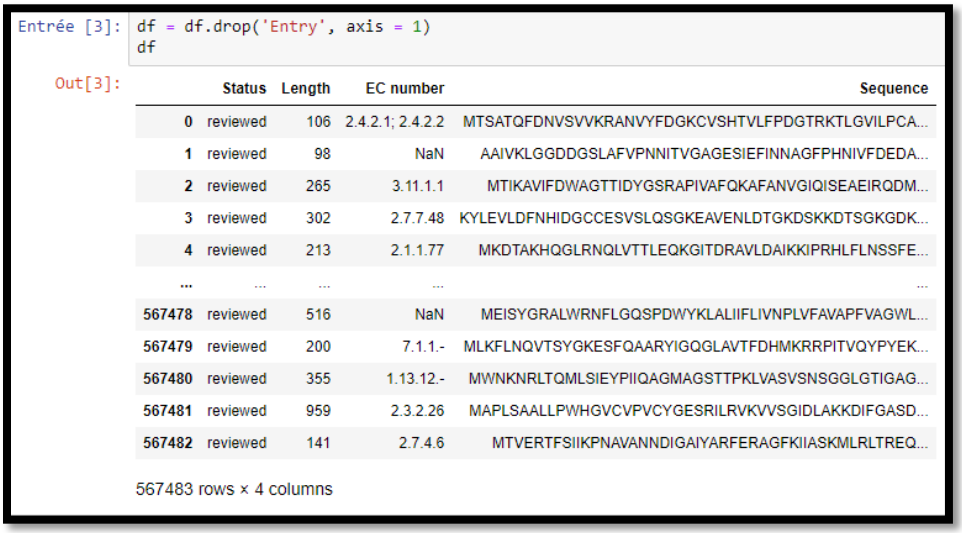
2. METHODES

Cette partie représente les méthodes utilisées afin de parvenir à la prédiction de la classification enzymatique relative aux hydrolases.

2.1. Prétraitement des données

La phase de prétraitement représente diverses étapes parcourues afin de parvenir à un dataset propre et prêt à être introduit au model d'apprentissage, ces étapes sont :

- Suppression des colonnes Entry et Status, n'étant d'aucune utilité pour l'apprentissage.



Entrée [3]:
```
df = df.drop('Entry', axis = 1)
df
```

Out[3]:

	Status	Length	EC number	Sequence
0	reviewed	106	2.4.2.1; 2.4.2.2	MTSATQFDNVSVVKRANVYFDGKCVSHTVLFPDGTRKTLGVILPCA...
1	reviewed	98	NaN	AAIVKLGGDDGSLAFVPNNITVGAGESIEFINNAGFPHNIVFDEDA...
2	reviewed	265	3.11.1.1	MTIKAVIFDWAGTTIDYGSRAPIVAFQKAFANVGIQISEAEIRQDM...
3	reviewed	302	2.7.7.48	KYLEVLDFNHIDGCCESVSLQSGKEAVENLDTGKDSKKDTSGKGDK...
4	reviewed	213	2.1.1.77	MKDTAKHQGLRNQLVTTLEQKGITDRAVLDAIKKIPRHLFLNSSFE...
...	...	...	...	...
567478	reviewed	516	NaN	MEISYGRALWRNFLGQSPDWYKLALIIFLIVNPLVFAVAPFVAGWL...
567479	reviewed	200	7.1.1.-	MLKFLNQVTSYGKESFQAARYIGQGLAVTFDHMKRRPITVQYPYEK...
567480	reviewed	355	1.13.12.-	MWNKNRLTQMLSIEYPIIQAGMAGSTTPKLVASVSNSGGLGTIGAG...
567481	reviewed	959	2.3.2.26	MAPLSAALLPWHGVCVPVCYGESRILRVKVVSGIDLAKKDIFGASD...
567482	reviewed	141	2.7.4.6	MTVERTFSIIKPNAVANNDIGAIYARFERAGFKIIASKMLRLTREQ...

567483 rows × 4 columns

Figure 12 : suppression de la colonne entry

	Length	EC number	Sequence
0	106	2.4.2.1; 2.4.2.2	MTSATQFDNVSVVKRANVYFDGKCVSHTVLFPDGTRKTLGVILPCA...
1	98	NaN	AAIVKLGGDDGSLAFVPNNITVGAGESIEFINNAGFPHNIVFDEDA...
2	265	3.11.1.1	MTIKAVIFDWAGTTIDYGSRAPIVAFQKAFANVGIQISEAEIRQDM...
3	302	2.7.7.48	KYLEVLDFNHIDGCCESVSLQSGKEAVENLDTGKDSKKDTSGKGDK...
4	213	2.1.1.77	MKDTAKHQGLRNQLVTTLEQKGITDRAVLDAIKKIPRHLFLNSSFE...
...	...	...	...
567478	516	NaN	MEISYGRALWRNFLGQSPDWYKLALIIFLIVNPLVFAVAPFVAGWL...
567479	200	7.1.1.-	MLKFLNQVTSYGKESFQAARYIGQGLAVTFDHMKRRPITVQYPYEK...
567480	355	1.13.12.-	MWNKNRLTQMLSIEYPIIQAGMAGSTTPKLVASVSNSGGLGTIGAG...
567481	959	2.3.2.26	MAPLSAALLPWHGVCVPVCYGESRILRVKVVSGIDLAKKDIFGASD...
567482	141	2.7.4.6	MTVERTFSIIKPNAVANNDIGAIYARFERAGFKIIASKMLRLTREQ...

567483 rows × 3 columns

Figures 13 suppression de la colonne status

- Conversion de la colonne EC number en chaine de caractère, ce qui remplira les cellules vides (permettant de distinguer les protéines non enzymatiques des protéines enzymatiques) par la valeur nan (en chaine de caractères).

```
1 df
```

	Length	EC number	Sequence
0	106	2.4.2.1; 2.4.2.2	MTSATQFDNVSVVKRANVYFDGKCVSHTVLFPDGTRKTLGVILPCA...
1	98	nan	AAIVKLGGDDGSLAFVPNNITVGAGESIEFINNAGFPHNIVFDEDA...
2	265	3.11.1.1	MTIKAVIFDWAGTTIDYGSRAPIVAFQKAFANVGIQISEAEIRQDM...
3	302	2.7.7.48	KYLEVLDFNHIDGCCESVSLQSGKEAVENLDTGKDSKKDTSGKGDK...
4	213	2.1.1.77	MKDTAKHQGLRNQLVTTLEQKGITDRAVLDAIKKIPRHLFLNSSFE...
...	...	...	...
567478	516	nan	MEISYGRALWRNFLGQSPDWYKLALIIFLIVNPLVFAVAPFVAGWL...
567479	200	7.1.1.-	MLKFLNQVTSYGKESFQAARYIGQGLAVTFDHMKRRPITVQYPYEK...
567480	355	1.13.12.-	MWNKNRLTQMLSIEYPIIQAGMAGSTTPKLVASVSNSGGLGTIGAG...
567481	959	2.3.2.26	MAPLSAALLPWHGVCVPVCYGESRILRVKVVSGIDLAKKDIFGASD...
567482	141	2.7.4.6	MTVERTFSIIKPNAVANNDIGAIYARFERAGFKIIASKMLRLTREQ...

Figure 14 : Conversion de la colonne EC number en chaine de caractère

- Suppression de toutes les séquences ne disposant pas de EC number grâce à la valeur nan attribué par l'étape précédente, ce qui revient à supprimer toutes les protéines non enzymatiques.
- Suppression des séquences ayant plusieurs EC number.

```
3 df
```

	Length	EC number	Sequence
2	265	3.11.1.1	MTIKAVIFDWAGTTIDYGSRAPIVAFQKAFANVGIQISEAEIRQDM...
3	302	2.7.7.48	KYLEVLDFNHIDGCCESVSLQSGKEAVENLDTGKDSKKDTSGKGDK...
4	213	2.1.1.77	MKDTAKHQGLRNQLVTTLEQKGITDRAVLDAIKKIPRHLFLNSSFE...
5	588	3.1.1.5	MYKNRVELTTTAPVNRALPNAPDGYTPQGETCPSKRPSIRNATALS...
7	700	2.7.7.8	MNPIVKSFEYGQHTVTLETGVIARQADAAVLASMGDTTVLVTVVGK...
...	...	...	...
567477	151	2.7.4.6	MSTEQTFIAVKPDAVQRGLIGYIISKFELKGYKLRALKFLVPSRDL...
567479	200	7.1.1.-	MLKFLNQVTSYGKESFQAARYIGQGLAVTFDHMKRRPITVQYPYEK...
567480	355	1.13.12.-	MWNKNRLTQMLSIEYPIIQAGMAGSTTPKLVASVSNSGGLGTIGAG...
567481	959	2.3.2.26	MAPLSAALLPWHGVCVPVCYGESRILRVKVVSGIDLAKKDIFGASD...
567482	141	2.7.4.6	MTVERTFSIIKPNAVANNDIGAIYARFERAGFKIIASKMLRLTREQ...

Figure 15 : Suppression des séquences ayant plusieurs EC number

- Récupération du premier code du EC number dans une nouvelle colonne nommé first_level, qui sera ensuite utilisée pour l'apprentissage.

	Length	EC number	Sequence	first_level
2	265	3.11.1.1	MTIKAVIFDWAGTTIDYGSRAPIVAFQKAFANVGIQISEAEIRQDM...	NaN
3	302	2.7.7.48	KYLEVLDFNHIDGCCESVSLQSGKEAVENLDTGKDSKKDTSGKGDK...	NaN
4	213	2.1.1.77	MKDTAKHQGLRNQLVTTLEQKGITDRAVLDAIKKIPRHLFLNSSFE...	NaN

Figure 16 : Création d'une colonne first_level

- Suppression de la colonne EC number.
- Modification des classes en système binaire (hydrolase et non hydrolase)

0	MTIKAVIFDWAGTTIDYGSRAPIVAFQKAFANVGIQISEAEIRQDM...	hydrolase
1	KYLEVLDFNHIDGCCESVSLQSGKEAVENLDTGKDSKKDTSGKGDK...	non_hydrolase
2	MKDTAKHQGLRNQLVTTLEQKGITDRAVLDAIKKIPRHLFLNSSFE...	non_hydrolase
3	MYKNRVELTTTAPVNRALPNAPDGYTPQGETCPSKRPSIRNATALS...	hydrolase
4	MNPIVKSFEYGQHTVTLETGVIARQADAAVLASMGDTTVLVTVVGK...	non_hydrolase
...	...	...

Figure 17 : Modification des classes en hydrolase et non-hydrolase

- Suppression des séquences dont la taille est supérieure à 1000 acides aminés.

```
# Supression des séquences dont la taille est supérieure à 1000 aa
df = df[df['Length'] <= 1000]
df
```

	Length	Sequence	first_level
0	265	MTIKAVIFDWAGTTIDYGSRAPIVAFQKAFANVGIQISEAEIRQDM...	hydrolase
1	302	KYLEVLDFNHIDGCCESVSLQSGKEAVENLDTGKDSKKDTSGKGDK...	non_hydrolase
2	213	MKDTAKHQGLRNQLVTTLEQKGITDRAVLDAIKKIPRHLFLNSSFE...	non_hydrolase
3	588	MYKNRVELTTTAPVNRALPNAPDGYTPQGETCPSKRPSIRNATALS...	hydrolase
4	700	MNPIVKSFEYGQHTVTLETGVIARQADAAVLASMGDTTVLVTVVGK...	non_hydrolase
...	...	...	...
255854	151	MSTEQTFIAVKPDAVQRGLIGYIISKFELKGYKLRALKFLVPSRDL...	non_hydrolase
255855	200	MLKFLNQVTSYGKESFQAARYIGQGLAVTFDHMKRRPITVQYPYEK...	non_hydrolase
255856	355	MWNKNRLTQMLSIEYPIIQAGMAGSTTPKLVASVSNSGGLGTIGAG...	non_hydrolase
255857	959	MAPLSAALLPWHGVCVPVCYGESRILRVKVVSGIDLAKKDIFGASD...	non_hydrolase
255858	141	MTVERTFSIIKPNAVANNDIGAIYARFERAGFKIIASKMLRLTREQ...	non_hydrolase

Figure 18 : Suppression des séquences dont la taille est supérieure à 1000 aa

- Suppression des séquences dont la taille est inférieure à 50 acides aminés.
- Suppression de la colonne Length.
- Suppression des séquences qui contiennent les acides aminés non commun/ambiguës qui sont : X, O, J, U, Z, B afin de réduire l'ambiguïté de l'apprentissage.

```
df = df[~df.Sequence.str.contains("O")]
df = df[~df.Sequence.str.contains("J")]
df = df[~df.Sequence.str.contains("U")]
df = df[~df.Sequence.str.contains("Z")]
df = df[~df.Sequence.str.contains("B")]
```

Figure 19 : Suppression des séquences avec des acides aminés non commun

- Enregistrement du dataset propre.

2.2. Apprentissage

Afin de pouvoir introduire les données au model d'apprentissage, les données doivent d'abord être transformées, car le modèle d'apprentissage ne prend pas en charge les chaines de caractères, deux transformations ont eu lieu :

2.2.1. One hot encoding :

Le One hot encoding est procédé par le One hot encoder qui est une fonction de la bibliothèque Sklearn, il permet de convertir les classes en vecteur comme suit :

	Hydrolase	Non-Hydrolase
Hydrolase	1	0
Non-Hydrolase	0	1

Ce processus convertis les hydrolases en vecteur [1,0] et les non-hydrolase en [0,1].

```
from sklearn.preprocessing import OneHotEncoder

encoder = OneHotEncoder(sparse=False)
classes_labels = Y.reshape(((len(Y)), 1))
Y = encoder.fit_transform(classes_labels)
Y
```

```
array([[1., 0.],
       [0., 1.],
       [0., 1.],
       ...,
       [0., 1.],
       [0., 1.],
       [0., 1.]])
```

Figure 20 : Application du One Hot Encoding

2.2.2. Tokenizing

Le tokenizing est procédé par Tokenizer qui est une fonction de la bibliothèque Tensorflow, il permet la conversion des séquences en vecteur de numéro suivant la même codification :

Séquence	Token
MKDT	[1, 2, 3, 4]
MTKD	[1, 4, 2, 3]

2.2.3. Padding

Le padding est procédé par pad_sequences qui est une fonction de la bibliothèques Tensorflow, il permet d'unifier la taille des vecteurs, en ajoutant des zéros (valeurs nulles) aux petits vecteurs et en tronquant les longs vecteurs, la taille unifié est choisie selon le dataset.

Séquence	Token
MKDT	[1, 2, 3, 4]
MTK	[1, 4, 2, 0]

Séquence	Token
MKDT	[1, 2]
MT	[1, 4]

2.2.4. Construction du model

Une fois les données prêtes, on procède à la construction du model qui est de type Sequential. Le model se compose des couches suivantes :

- La couche Embedding qui convertira chaque vecteur en une matrice en attribuant des valeurs numériques à chaque élément du vecteur. Elle prend comme option le nombre de valeurs possibles pour chaque élément du vecteur, la taille de l'embedding ainsi que la taille du vecteur.
- La couche Conv1D qui représente le réseau de neurones Convolutionnelle (CNN), qui est composée de 256 filtres, qui dispose d'un kernel d'une taille de 4 unités et d'une fonction d'activation de type Relu.
- La couche MaxPolling1D effectue un regroupement maximal. C'est une opération de regroupement qui sélectionne le maximum d'éléments significatifs pour la classification. Ainsi, son output est une matrice contenant les caractéristiques les plus importantes de l'output de la couche CNN.
- La couche Flatten qui convertit l'output de la couche MaxPooling1D qui est une matrice en un vecteur.

```
embedding_dim = 64
model = tf.keras.Sequential([
    tf.keras.layers.Embedding(len(tokenizer.word_index)+1, 64, input_length=1000),
    tf.keras.layers.Conv1D(filters=256, kernel_size=4, padding='same', activation='relu'),
    tf.keras.layers.MaxPooling1D(pool_size=8),
    tf.keras.layers.Conv1D(filters=128, kernel_size=4, padding='same', activation='relu'),
    tf.keras.layers.MaxPooling1D(pool_size=8),
    tf.keras.layers.Dropout(.4, input_shape=(4,)),
    tf.keras.layers.Flatten(),
    tf.keras.layers.Dropout(.3, input_shape=(3,)),
    tf.keras.layers.Dense(32, activation='relu'),
    tf.keras.layers.Dense(2, activation='softmax'),])
model.compile(loss='binary_crossentropy', optimizer='adam', metrics=['accuracy'])
print(model.summary())
```

Figure 21 : Construction du model du deep learning

- La couche Dense qui représente un réseau de neurone artificiel standard, qui se chargera de la prise de décision de la prédiction.
- La couche Dense finale dont la sortie sera la probabilité de la prédiction. Etant donné que la classification est binaire, une valeur inférieure à 50% signifie que la classe prédite est non-hydrolase (absence du caractère) et une valeur supérieure à 50% signifie que la classe prédite est hydrolase (présence du caractère).

2.2.5. Répartition des données

Les données sont réparties en deux parties : 80% des données seront utilisées pour l'apprentissage et 20% de données qui seront utilisées pour le test.

Les 20% des données du test sont à leur tour divisées en deux parties : 50% pour le test lors de l'apprentissage et 50% pour l'évaluation du model après la fin de l'apprentissage.

2.2.6. L'apprentissage

L'apprentissage se lance en utilisant la fonction fit du Tensorflow, elle prend comme option les données d'apprentissage et les données du test ainsi que le nombre d'epochs (un epoch correspond à un apprentissage sur toutes les données) que le model doit parcourir afin de terminer son entrainement.

3. RESULTATS ET DISSCUSSION

Les résultats de ce travail est scindé en deux principales parties :

1-Résultats du prétraitement : Plusieurs étapes ont eu lieu lors du prétraitement, dont l'objectif était de nettoyer notre dataset et de ne garder que les données pertinentes pour l'apprentissage, les résultats de chaque étape sont les suivants :

- Le dataset brute contenait 567 483 séquences protéiques : enzymes et non-enzymes.

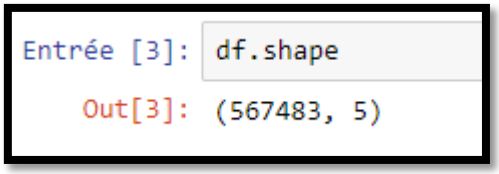


Figure 22 : Le nombre de séquences du dataset initial

- La suppression des séquences protéiques non enzymatiques a abouti à 273 690 séquences protéiques enzymatiques.

Figure 23 : Le nombre de séquences après suppression des séquences non enzymatiques

- La suppression des enzymes ayant plusieurs EC number a abouti à 255 859 séquences.

- La suppression des séquences trop courtes ou trop longues a abouti à 247 657 séquences.

- La suppression des séquences avec des acides aminés non communs a abouti à 246 992 séquences.

Ce prétraitement a donné au final un dataset composé de 52 752 séquences représentant la classe des hydrolases et 194 240 séquences représentant le reste des classes.

```
Out[26]: non_hydrolase    194240
         hydrolase         52752
         Name: first_level, dtype: int64
```

Figure 24 : Le nombre de séquences du dataset final

2- Résultats de l'apprentissage profond : Une fois l'apprentissage terminé, le modèle a été testé afin de vérifier son efficacité et sa précision. Le premier test que le modèle a subi, a été fait à la fin de l'apprentissage, en prédisant si l'enzyme est une hydrolase ou non pour chacune des séquences des données du test (soit 12 350 séquences). La prédiction a abouti aux résultats suivants :

- Accuracy = 97,1 %. Cette valeur est calculée en divisant le nombre des bonnes prédictions sur le nombre total des séquences, ce qui veut dire que sur les 12 350 séquences, la prédiction était correcte pour 11 992 séquences. L'apprentissage, a donc, abouti à la construction d'un modèle de DL capable de prédire avec une haute précision, si l'enzyme est une hydrolase ou non, se basant uniquement en connaissant la séquence de cette enzyme.

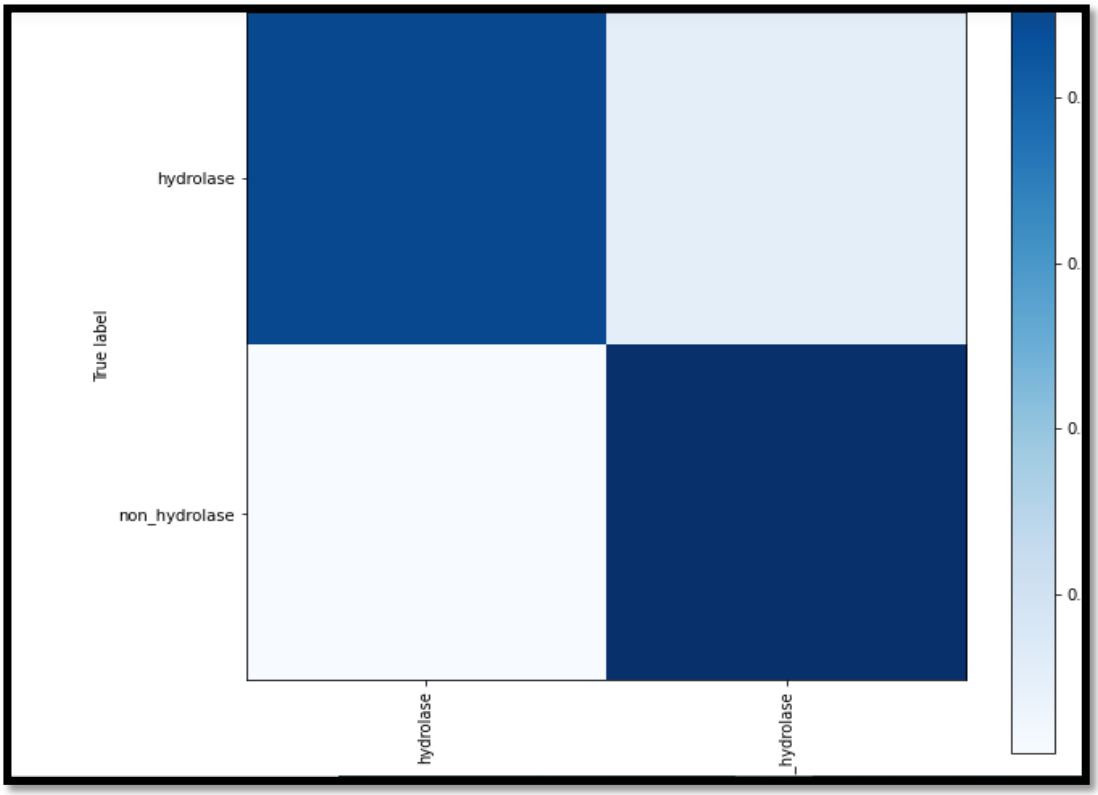


Figure 25 : Matrice de confusion du test du model

	precision	recall	f1-score	support
hydrolase	0.97	0.90	0.93	2618
non_hydrolase	0.97	0.99	0.98	9732
accuracy			0.97	12350
macro avg	0.97	0.94	0.96	12350
weighted avg	0.97	0.97	0.97	12350

Figure 26 : Rapport de classification du model

CONCLUSION

L'un des défis majeurs de la bioinformatique est l'annotation des génomes et des protéomes voire toutes les disciplines omiques car elle constitue un dénominateur commun à toutes les sciences et les disciplines des sciences de la nature et de la santé. Ce travail soutient partiellement cette implication de cette science émergeante en prédisant la classification des enzymes hydrolases.

L'approche adoptée par ce travail est de développer un modèle de prédiction basé sur NLP et DL pour prédire si une séquence d'enzyme fait partie ou non de la classe des hydrolases et ce, grâce à sa séquence primaire d'acide aminés.

Le travail réalisé et les résultats obtenus montrent que le DL joue un rôle efficace dans les prédictions et les classifications des protéines, particulièrement les protéines enzymatiques.

Cependant, malgré la précision apportée par cette approche, ce travail reste incomplet voire imprécis, et mérite des recherches plus approfondies sur des problématiques actuelles intéressant les Datasciences. Ainsi, les perspectives d'avenir corrélées à notre initiation en IA sont :

- Utiliser des techniques de DL plus sophistiquées pour améliorer la précision
- Faire l'apprentissage sur des bases de données plus variées
- Permettre au modèle d'effectuer l'apprentissage sur toutes les classes d'enzymes
- Utiliser d'autres données que la séquence de l'enzyme pour effectuer l'apprentissage.

REFERENCES

1 BENRAHHAL A. 2021. *Les applications de l'intelligence artificielle dans le domaine de la santé*. Université Mohammed de Rabat. p : 46.

2 BENRAHHAL A. 2021. *Les applications de l'intelligence artificielle dans le domaine de la santé*. Université Mohammed de rabat. pp : 46-49.

3 BENRAHHAL A. 2021. *Les applications de l'intelligence artificielle dans le domaine de la santé*. Université Mohammed de rabat. p : 50.

4 Bairoch A. 2000. La base de données ENZYME en 2000. *Nucleic Acids Research*. **28** (1) : 304 – 305.

5 Chand D., Avinas V. C., Yadav Y., Pundle A. V., Suresh C. G., Ramasamy S. 2017. Caractéristiques moléculaires des hydrolases des sels biliaires et pertinence pour la santé humaine. *Biochimica et Biophysica Acta* - Sujets généraux. **1861** : 2981-2991

6 AKSAS K. 2016. *Cours d'enzymologie*. 2ème année pharmacie. pp : 5-10. URL : pharm2an16_bioch-enzymologie.pdf. Consulté en Mai 2022.

7 Frederic S. 2021. *Introduction à l'apprentissage automatique*. Ecole des Mines de Nancy. p.11

8 GARET G. 2014. Classification et caractérisation des familles enzymatiques à l'aide de méthodes formelles. Université européenne de Bretagne.Thèse de Doctorat. France.

9 Gariev A., Varfolomeev S.D. 2006. Classification hiérarchique des sites catalytiques des hydrolases. Bioinformatique. 22 (20) : 2574

10 http://www.enzyme-database.org/downloads/ec3.pdf

11 https://www.futura-sciences.com/sante/definitions/biologie-hydrolase-9070/

12 https://www.futura-sciences.com/sante/definitions/biologie-hydrolase-9070/

13 https://www.vedantu.com/chemistry/hydrolase?fbclid=IwAR3SjYqbZkVSknh_FqI7xI37pfSTCSzz2w-UZDxSEl_2ywrBJ31wLtRvIDw

14 LANTEIGNEROCH L. M. 2010. *Utilisation des enzymes lipases et lactases pour améliorer la blancheur d'une pate désancrée de papier journal*. Mémoire de recherche. Université du Québec. p : 36.

15 McDonald A. G., Boyce S., Tipton K. F. 2009. ExplorEnz : La principale source de la liste des enzymes de l'IUBMB. *Recherche sur les acides nucléiques*. **37** : D593-D597

16 Anonyme. URL : https://www.netapp.com. Consulté en Mai 2022.

17 LAKHDAR O. 2021. *Prédiction de la toxicité des produits chimiques*. Thèse pour l'obtention du diplôme de Docteur en pharmacie. Université Mohamed de Rabat faculté de médecine et de pharmacie. pp : 46-48.

18 Anonyme. https://slideplayer.fr

19 Boukerche T. T. 2007. *Synthèse de nouvelles molécules lenciactives non ioniques par voie enzymatique*. Mémoire de magister. Université d'Oran ES-senia. p : 29.

Année universitaire : 2021-2022	**Présenté par :** BELBEKRI Amina

Classification des hydrolases à partir des données protéomiques par approche Deep Learning

Mémoire pour l'obtention du diplôme de Master en BioInformatique

Résumé

IL existe actuellement sept classes d'enzymes après ajout de la famille des Translocases. Les hydrolases sont une classe qui utilisent l'eau pour rompre une liaison chimique, ce qui entraîne généralement la division d'une molécule plus grosse en molécules plus petites. L'objectif principal de ce travail est de proposer un système de classification d'enzymes basé sur leur structure primaire en utilisant une approche bioinformatique de Deep Learning. Les résultats de notre travail ont été concluants et ont abouti à une valeur Accuracy égale à 97,1 % ce qui est acceptable relativement aux publications internationales. Cependant, il est plus intéressant de travailler sur un nombre de séquences plus important afin d'aboutir à des conclusions plus pertinentes et plus précises. Ce travail a permis de mettre en évidence l'apport de cette approche et à améliorer la précision de la classification des protéines.

Mots clés : Hydrolases, Prédiction, Intelligence Artificielle, Deep Learning.

Abstract

There are currently seven classes of enzymes after adding the Translocase family. Hydrolases are a class that use water to break a chemical bond, which usually causes a larger molecule to split into smaller molecules. The main objective of this work is to propose an enzyme classification system based on their primary structure using a Deep Learning bioinformatics approach. The results of our work were conclusive and led to an Accuracy value equal to 97.1%, which is acceptable relative to international publications. However, it is more interesting to work on a larger number of sequences in order to reach more relevant and more precise conclusions. This work made it possible to highlight the contribution of this approach and to improve the precision of the classification of proteins.

Keywords: Hydrolases, Prediction, Artificial Intelligence, Deep Learning.

ملخص

يوجد حاليًا سبع فئات من الإنزيمات بعد إضافة عائلة .Translocase الهيدرولازات هي فئة تستخدم الماء لكسر رابطة كيميائية ، والتي عادة ما تتسبب في انقسام جزيء أكبر إلى جزيئات أصغر .الهدف الرئيسي من هذا العمل هو اقتراح نظام تصنيف الإنزيمات بناءً على هيكلها الأساسي باستخدام نهج التعلم العميق للمعلومات الحيوية .كانت نتائج عملنا حاسمة وأدت إلى قيمة Accuracy تساوي 97.1٪ ، وهو أمر مقبول بالنسبة للمنشورات الدولية .ومع ذلك ، فمن المثير للاهتمام العمل على عدد أكبر من التسلسلات من أجل الوصول إلى استنتاجات أكثر صلة وأكثر دقة . أتاح هذا العمل تسليط الضوء على مساهمة هذا النهج وتحسين دقة تصنيف البروتينات

.الكلمات المفتاحية : الهيدرولازات ، التنبؤ ، الذكاء الاصطناعي ، التعلم العميق.

Mots-clefs : Hydrolases, Prédiction, Intelligence Artificielle, Deep Learning

Laboratoires de recherche :

Laboratoire de Génie Microbiologique et Applications (Université Frères Mentouri, Constantine 1)

Encadreur : HAMIDECHI Mohamed Abdelhafid (Professeur - Université Frères Mentouri, Constantine 1)

Examinateur 1 : DAAS Mohamed Skandar (MCA - Université Frères Mentouri, Constantine 1)

Examinateur 2 : BOULAHROUF Khaled (MCB - Université Frères Mentouri, Constantine 1)