**REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE**

**MINISTERE DE L'ENSEIGNEMENT SUPERIEURE ET DE LA RECHERCHE SCIENTIFIQUE**
**UNIVERSITE DE MEDEA**

**FACULTE DES SCIENCES**



# Mémoire

Présenté pour obtenir le diplôme de **Master**

En Sciences de la nature et de la vie

**Spécialité :** Biologie et Pathologie Cellulaire

**Par :**

**GUEZZI Khedidja**

**Thème :**

## Approche d'intelligence artificielle basée sur l'expression génique pour la classification des types de cancer

Soutenue publiquement le : 17/06 /2023. Devant le jury composé de :

| | | | |
|---|---|---|---|
| **CHAOUATI Karima** | **Grade : MCB** | Univ. Médéa | Président |
| **ZERGOUG Amina** | **Grade : MCB** | Univ. Médéa | Examinateur |
| **HOCEINI Amina** | **Grade : MCA** | Univ. Médéa | Encadreur |
| **BOUKELIA Abdelbasset** | **Grade : MRB** | CRBT Constantine | Co-Encadreur |

**Année universitaire 2022/2023**

# *Remerciements*

*Je tiens d'abord à exprimer mes remerciements et gratitude au Seigneur Dieu de m'avoir donné la force et le courage de mener à terme ce mémoire. J'aimerais adresser mes profonds respects et remerciements à Mr* ***Abdelbasset Boukelia****, de l'attention et du soutien qu'il a porté à mon travail.*

*Je tiens également à exprimer ma gratitude à Mme* ***Amina Hoceini*** *pour sa disponibilité et l'attention dont elle m'a entourée tout au long de ce travail. Je souhaite particulièrement adresser mes vifs remerciements aux membres du jury, Mme* ***K. Chaouati*** *et Mme* ***A. Zergoug****, d'avoir accepté d'examiner mon mémoire de master.*

*Que le personnel de l'Université Yahia Fares de Médéa trouve ici l'expression de ma sincère reconnaissance.*

*Mes plus vifs remerciements à mon mari, mes chers parents et surtout ma sœur Amel. Je ne leur serai probablement jamais assez reconnaissant de m'avoir soutenu tout au long de mes études.*

## *Résumé*

La classification des différents types de tumeurs est d'une grande importance dans le diagnostic du cancer et la découverte de médicaments. Cependant, la plupart des études antérieures de classification sont uniquement basées sur des données cliniques. Elles ont une capacité de diagnostic limitée. Ainsi, l'utilisation d'autres types de classification, à l'instar de la classification à l'aide des données d'expression génique, est devenue indispensables. Cette dernière est utilisée pour contenir les clés permettant de résoudre les problèmes fondamentauxliés au diagnostic du cancer et à la découverte de médicaments.

L'évolution des plateformes de séquençages a rendu possible la surveillance simultanée de milliers de profil d'expressions géniques. Avec cette abondance des données, les chercheurs ont commencé à explorer les possibilités d'utiliser ce type de méthodes pour la classification des tumeurs.

Nous avons utilisé dans notre contribution à la classification des maladies cancéreuses à partirde l'expression génique des techniques de machine learning et de deep learning. La méthode proposée dans cette étude repose sur une approche en trois étapes pour la classification des maladies cancéreuses à partir de l'expression génique, avec pour objectif principal d'améliorer la précision diagnostique et d'identifier des biomarqueurs informatifs. En mettant en œuvre notre approche et en évaluant les performances des différents modèles, nous avons constaté que le modèle CNN a obtenu les meilleurs résultats en termes de précision, sensibilité et spécificité. Ses capacités de traitement des données en 1D, en utilisant des filtres et des convolutions, ont permis de capturer des caractéristiques spécifiques à partir de l'expression génique, conduisant à une classification plus précise des types de cancers.



***Mots clefs :*** *apprentissage automatique, apprentissage approfondi, bioinformatique, biomarqueur, expression génique, algorithmes d'optimisation, classification des cancers.*

## *Summary*

The classification of different types of tumors is of great importance in cancer diagnosis and drug discovery. However, most previous classification studies have been solely based on clinical data, which have limited diagnostic capabilities. Therefore, the use of alternative classification methods, such as classification using gene expression data, has become indispensable. Gene expression data contains the keys to solving fundamental issues related tocancer diagnosis and drug discovery.

The advancement of sequencing platforms has made it possible to simultaneously monitor thousands of gene expression profiles. With this abundance of data, researchers have begun exploring the possibilities of using this type of method for tumor classification.

In our contribution to the classification of cancerous diseases based on gene expression, we have used machine learning and deep learning techniques. The method proposed in this study is based on a three-step approach for the classification of cancer diseases based on gene expression, with the main objective of improving diagnostic accuracy and identifying informative biomarkers. By implementing our approach and evaluating the performance of the different models, we found that the CNN model performed best in terms of accuracy, sensitivity, and specificity. Its 1D data processing capabilities, using filters and convolutions, have enabled the capture of specific features from gene expression, leading to more accurate classification of cancer types.



***Keywords**: machine learning, deep learning, bioinformatics, biomarker, gene expression, optimization algorithms, cancer classification.*

# ملخص

يعتبر تصنيف أنواع الأورام المختلفة ذا أهمية كبيرة في تشخيص السرطان واكتشاف الأدوية. ومع ذلك ، فقد استندت معظم دراسات التصنيف السابقة فقط على البيانات السريرية ، والتي لها قدرات تشخيصية محدودة. لذلك ، أصبح استخدام طرق التصنيف البديلة ، مثل التصنيف باستخدام بيانات التعبير الجيني ، أمرًا لا غنى عنه. تحتوي بيانات التعبير الجيني على مفاتيح حل المشكلات الأساسية المتعلقة بتشخيص السرطان واكتشاف الأدوية.

جعل تقدم منصات التسلسل من الممكن مراقبة الآلاف من ملفات تعريف التعبير الجيني في وقت واحد. مع هذه الوفرة من البيانات، بدأ الباحثون في استكشاف إمكانيات استخدام هذا النوع من الأساليب لتصنيف الورم.

في مساهمتنا في تصنيف الأمراض السرطانية بناءً على التعبير الجيني ، استخدمنا تقنيات التعلم الآلي والتعلم العميق. تعتمد الطريقة المقترحة في هذه الدراسة على نهج من ثلاث خطوات لتصنيف أمراض السرطان بناءً على التعبير الجيني ، بهدف رئيسي هو تحسين دقة التشخيص وتحديد المؤشرات الحيوية المفيدة. من خلال تنفيذ نهجنا وتقييم أداء النماذج المختلفة ، وجدنا أن نموذج CNN كان أفضل أداء من حيث الدقة والحساسية والخصوصية. لقد مكنت قدرات معالجة البيانات أحادية الأبعاد ، باستخدام المرشحات والتلافيف ، من التقاط ميزات محددة من التعبير الجيني ، مما أدى إلى تصنيف أكثر دقة لأنواع السرطان.

**الكلمات المفتاحية:** التعلم الآلي ، التعلم العميق ، المعلوماتية الحيوية ، العلامات الحيوية ، التعبير الجيني ، خوارزميات التحسين ، تصنيف السرطان.

| | |
|---|---|
| **Introduction générale :…….** | **01** |
| **Références bibliographiques :…….** | **03** |
| **Chapitre I : Cancérogénèse** | |
| **I.1. Introduction :……...** | **04** |
| **I.2. Biologie moléculaire :……..** | **04** |
| **I.2.1. ADN (Acide désoxyribonucléique) :……...** | **05** |
| **I.2.2. ARN Acide ribonucléique :…….** | **07** |
| **I.2.3. Les protéines :……..** | **08** |
| **I.2.3.1. Structure des protéines :……...** | **09** |
| **I.2.3.2. Biosynthèse des Protéines :……..** | **10** |
| **I.2.3.3. Les protéines et la carcinogenèse :…….** | **11** |
| **I.2.4. Expression génétique :……** | **12** |
| **I.2.5. Gène :……..** | **14** |
| **I.2.6. Génome :…….** | **17** |
| **I.2.7. Biomarqueurs :…….** | **18** |
| **I.2.7.1. Caractéristiques d'un biomarqueur idéal :……** | **20** |
| **I.2.7.2. Les biomarqueurs en cancérologie :……** | **20** |
| **I.3. Cancérogénèse :……..** | **22** |
| **I.4. Conclusion :…….** | **23** |
| **I.5. Références bibliographiques :……..** | **25** |
| **Chapitre II : Bioinformatique** | |
| **II.1. Introduction :…….** | **28** |
| **II.2. Bioinformatique :……..** | **28** |
| **II.2.1. Objectifs :……** | **30** |

| | |
|---|---|
| **II.2.2. Relation avec d'autres domaines……..** | **31** |
| **II.2.3. Les méthodes bioinformatiques par homologie :……** | **32** |
| **II.2.3.1. Alignement de séquences :……..** | **32** |
| **II.2.3.2. Recherche de similarité de séquences :…….** | **32** |
| **II.2.3.3. Recherche de domaines conservés :……..** | **33** |
| **II.2.3.4. Modélisation de la structure tridimensionnelle :……..** | **33** |
| **II.2.3.5. Phylogénie :……..** | **33** |
| **II.2.4. Les méthodes bioinformatiques par caractéristique :……** | **34** |
| **II.2.4.1. Recherche de motifs : …….** | **34** |
| **II.2.4.2. Prédiction de la structure secondaire : …….** | **34** |
| **II.2.4.3. Prédiction de la localisation subcellulaire : ……..** | **34** |
| **II.2.4.4. Annotation fonctionnelle : …….** | **35** |
| **II.2.4.5. Analyse des propriétés physico-chimiques : …….** | **35** |
| **II.3. L'intelligence artificielle (IA) :……** | **36** |
| **II.3.1. L'apprentissage automatique (Machine learning) : ……** | **36** |
| **II.3.1.1. Applications de l'apprentissage automatique en bioinformatique :……..** | **37** |
| **II.3.2. L'apprentissage profond : ……** | **39** |
| **II.4. Travaux connexes :……** | **42** |
| **II.5. Conclusion :……** | **45** |
| **II.6. Références bibliographiques :……** | **47** |
| **Chapitre III : Contribution** | |
| **III.1. Introduction :……** | **49** |

| | |
|---|---|
| **III.2. Méthode Proposée :…….** | **49** |
| **III.2.1. Traitement de données :…….** | **50** |
| **III.2.2. Modèle de classification CNN :……..** | **52** |
| **III.2.3. Modèle de classification avec ML :…….** | **53** |
| **III.2.4. Extraire les biomarqueurs :…….** | **55** |
| **III.3. Implémentation : …….** | **57** |
| **III.3.1. Dataset :……** | **57** |
| **III.3.2. Plateformes :…….** | **59** |
| **III.3.3. Métrique d'évaluation :……..** | **60** |
| **III.4. Résultats Expérimentaux :……..** | **62** |
| **III.5. Conclusion :……..** | **65** |

## Liste des figures

*Figure I.1 : L'ARN à double brin (ou ARNdb)* 8
*Figure I.2: Production d'une chaine polypeptidique à partir d'une chaine d'ADN* 10
*FigureI.3: le processus de synthèse protéique* 12
*Figure I.4: Le génome dans la cellule* 15
*Figure I.5: Le développement du processus tumoral* 22
*Figure I.6: Les étapes du processus de cancérogenèse* 23
*Figure II.1 : La bioinformatique est une combinaison de biologie, d'informatique, de mathématiques, de statistiques, de physique et de principes d'ingénierie* 26
*Figure II.2: Partie d'une page de résultats de recherche de similarité de séquences* 29
*Figure III.1 : captures d'écran montrant les différentes étapes suivies pour la, classification, des types de cancer* 58
*Figure III.2 : Graphe montrant les différents biomarqueurs extraits* 63

## Liste des tableaux

*Tableau I.1: Exemples de quelques cancers et leurs biomarqueurs* 21
*Tableau III.1: Résultats Expérimentaux des 3 algorithmes utilisés* 60
*Tableau III.2: Les 20 biomarqueurs géniques pour chaque type de cancer* 62

## Liste des abréviations

***ADN : Acide désoxyribonucléique***

***ARN : Acide ribonucléique***

***EFGR : Récepteur du facteur de croissance épidermique***

***BLAST (Basic Local Alignment Search Tool)***

***IA : Intelligence artificielle***

***SVM : Machines à vecteurs de support***

***RNN : Réseaux de neurones récurrents***

***CNN : Les réseaux de neurones convolutifs***

***MLP : Perceptrons multicouches***

***CFS : Sélection de fonctionnalités basée sur la corrélation***

***ReLU : Rectified Linear Unit***

***BRCA : Cancer du sein***

***KIRC : Cancer du rein clair***

***LUAD : Adénocarcinome du poumon***

***PRAD : Cancer de la prostate***

***COAD : Adénocarcinome du côlon***

***IDE : Environnements de développement intégrés***

# Introduction Générale

# Introduction générale

Ces dernières années, la recherche en biologie et tout particulièrement en génétique a connu un formidable essor. Les avancées réalisées sont considérables. Les chercheurs utilisent d'autres techniques notamment avec l'évolution des plateformes de séquençages qui fournissent une très importante somme d'informations que nous ne pouvons traiter sans l'aide de l'informatique. A tel point, qu'il y a un peu plus d'une dizaine d'années, une nouvelle discipline a été créée nommée la bioinformatique, située au carrefour de la biologie moléculaire, des mathématiques, des statistiques et de l'informatique.

La classification et la prédiction du type de cancer à partir des données d'expression génique peuvent être effectuées à l'aide de diverses techniques de bioinformatique, d'apprentissage automatique et d'apprentissage profond [1-4].

Les données d'expression génique ont largement été utilisées dans la recherche sur le cancer pour identifier des biomarqueurs pour le diagnostic et le pronostic, ainsi que pour développer des modèles prédictifs pour la médecine personnalisée. Les approches d'intelligence artificielle (IA), telles que l'apprentissage automatique et l'apprentissage profond, ont également montré des promesses dans l'analyse des données d'expression génique pour prédire le type de cancer[5].

Une approche courante consiste à utiliser les données d'expression génique en entrée d'un algorithme d'apprentissage automatique, qui apprend ensuite à classifier les échantillons en différents types de cancer en fonction des modèles présents dans les données. Cette approche a été appliquée à divers types de cancer, tels que le cancer du sein, le cancer du poumon et la leucémie, avec des degrés de succès variables[6].

Une autre approche consiste à utiliser des algorithmes d'apprentissage profond, tels que les réseaux de neurones convolutionnels (CNN), pour analyser les données d'expression génique. Les CNN sont particulièrement adaptés aux tâches d'analyse d'images, mais ils peuvent également être appliqués aux données d'expression génique en traitant les niveaux

d'expression de chaque gène comme un "pixel" dans une image haute dimension. En entraînant le CNN sur de grands ensembles de données d'expression génique, il peut apprendre à identifier des modèles et des caractéristiques qui sont caractéristiques de types de cancer spécifiques[7].

Il existe également des approches qui combinent les données d'expression génique avec d'autres types de données omiques, tels que la méthylation de l'ADN et l'expression des miARN, pour améliorer la prédiction du type de cancer. Ces approches intégratives peuvent fournir une vue plus complète des caractéristiques moléculaires du cancer et améliorer la précision des modèles prédictifs[8].

Dans l'ensemble, les approches d'IA basées sur les données d'expression génique montrent des promesses pour la prédiction du type de cancer, mais une validation et un affinement supplémentaires sont nécessaires avant qu'elles puissent être appliquées dans des contextes cliniques[9].

Ce mémoire de master est composé de trois chapitre dont nous présentons une brève description ci-dessous :

Le Chapitre I présente l'essentiel de la biologie moléculaire. Ensuite, nous avons présenté quelques notions et terminologies biologiques afin de comprendre tous ce qui suit dans notre recherche. Nous avons montré dans cette section le dogme central de la biologie moléculaire qui est les trois macromolécules cellulaire (ADN, ARN et protéine). Finalement, nous avons élucidé le processus de carcinogénèse, et nous avons compris que la compréhension de la carcinogenèse repose sur l'étude de l'ADN, des gènes et du génome, ainsi que des mécanismes moléculaires impliqués dans la régulation de la croissance cellulaire et la progression tumorale.

Le Chapitre II présente un état de l'art sur la bioinformatique, l'apprentissage profond (deep learning) et l'apprentissage automatique (machine learning) ou on a montré que ces domaines sont essentiels pour la prédiction et la classification du cancer. Leurs approches permettent une analyse approfondie des données biologiques, une meilleure compréhension de la biologie du cancer et une personnalisation des traitements pour améliorer les résultats pour les patients atteints de cancer.

Dans le Chapitre III, on a présenté en détail notre approche pour la classification des maladies cancéreuses à partir de l'expression génique en utilisant des techniques de machine

learning et deep learning. Nous avons abordé trois principales étapes : le traitement des données, la classification à l'aide de différentes méthodes (CNN, SVM et MLP), et l'extraction des biomarqueurs.

# Chapitre I:
# **Cancérogenèse**

## I.1. Introduction

Le domaine de la cancérogenèse a rapidement évolué, notamment depuis la révolution de la biologie moléculaire dans les années 1970. Il s'agit d'un processus graduel qui peut prendre des années, voire des décennies, avant qu'un cancer ne se développe[10].

La carcinogenèse implique des changements dans le matériel génétique des cellules, qui peuvent être causés par des facteurs environnementaux ou des erreurs aléatoires lors de la division cellulaire[11].

La compréhension de la carcinogenèse est importante pour la prévention et le traitement des cancers. En effet, la plupart des cancers sont évitables et peuvent être prévenus en évitant les facteurs de risque connus et en adoptant un mode de vie sain. De plus, la compréhension des mécanismes de la carcinogenèse peut conduire au développement de nouveaux traitements pour le cancer[12].

Des études ultérieures ont révélé un lien entre l'exposition aux produits chimiques, le métabolisme et les dommages causés à l'acide désoxyribonucléique (ADN), fournissant une compréhension plus complète de l'étiologie de la cancérogenèse. Il est maintenant bien admis que la carcinogenèse implique de multiples facteurs allant des expositions à des produits chimiques, des mutations dans les proto-oncogènes et des suppresseurs de tumeurs dues à des étiologies exogènes et endogènes, et une régulation ou des activités discordantes de nombreuses voies de signalisation critiques nécessaires à l'homéostasie cellulaire normale[13].

## I.2. Biologie moléculaire

La biologie moléculaire est une discipline scientifique ancienne. Elle a commencé avec l'évolution de la vie. Peu de temps après la cellule a été reconnue comme l'unité de base de la vie et sa structure a été déchiffrée (la théorie cellulaire), la biologie moléculaire a commencé après la théorie cellulaire et elle s’intéresse à étudier cette unité au niveau moléculaire et particulièrement les macromolécules d’ADN, ARN et protéine[14, 15]. Il est devenu clair que les constituants cellulaires ne sont que des composés chimiques. Bien que, certains de ces composés sont des substances chimiques simples comme les sucres et de nombreuses autres petites molécules, certaines autres molécules sont très grandes ayant une structure chimique complexe comme les acides nucléiques (ADN, ARN), des protéines, des glucides et des lipides complexes. Ceux-ci sont donc appelés des macromolécules. La biologie moléculaire

s'intéresse principalement à étudier la cellule au niveau moléculaire c'est-à-dire la compréhension des interactions entre les différents systèmes moléculaires d'une cellule, y compris les interactions entre ces macromolécules (l'ADN, l'ARN et la biosynthèse des protéines) ainsi que l'apprentissage de la façon dont ces interactions sont réglementées[16]. Selon Michel Morange [17], la biologie moléculaire est « l'ensemble des techniques et découvertes qui ont permis l'analyse moléculaire des processus les plus intimes du vivant, de ceux qui en assurant la pérennité et la reproduction ». La naissance de la biologie moléculaire moderne liée à la découverte de l'ADN et leur structure en double hélice en 1953. Ensuite, cette science a connu un essor très rapide selon les avancées réalisées dans la découverte des molécules cellulaire.

### I.2.1. ADN (Acide désoxyribonucléique)

Est un type de macromolécule connue sous le nom d'acide nucléique. Il a la forme d'une double hélice torsadée et est composé de longs brins de sucres et de groupes phosphate alternés, ainsi que de bases azotées (adénine, thymine, guanine et cytosine). L'ADN est organisé en structures appelées chromosomes et logé dans le noyau de nos cellules. L'ADN se trouve également dans les mitochondries cellulaires[18, 19].

L'ADN contient l'information génétique nécessaire à la production de composants cellulaires, les organites, et à la reproduction de la vie. La production de protéines est un processus cellulaire vital qui dépend de l'ADN. Les informations contenues dans le code génétique sont transmises de l'ADN à l'ARN aux protéines résultantes lors de la synthèse des protéines[20].

Les dernières découvertes indiquaient que la majorité d'ADN humain est transcrit à des différents ARN-transcrits ; mais seulement l'ARN messager (ARNm) parmi ces ARNtranscrits qui est traduit à la suite à une protéine (ADN codant).

Les changements génétiques peuvent entraîner le cancer s'ils modifient la façon dont les cellules se développent et se propagent. La plupart des modifications de l'ADN cancérigènes se produisent dans les gènes, qui sont des sections d'ADN qui portent les instructions pour fabriquer des protéines ou de l'ARN spécialisé tel que le microARN.

L'ADN joue un rôle crucial dans la carcinogenèse, car les altérations génétiques sont à l'origine des mutations qui conduisent à la transformation des cellules normales en cellules cancéreuses. Les altérations génétiques peuvent être causées par des facteurs

environnementaux, tels que les produits chimiques cancérigènes, les radiations ionisantes, les virus et les bactéries, ou par des erreurs lors de la réplication de l'ADN[21].

Les mutations de l'ADN peuvent être détectées par plusieurs techniques de diagnostic, notamment le séquençage de l'ADN, qui permet d'analyser l'ADN pour détecter les mutations. Cette information peut aider à déterminer la stratégie thérapeutique la plus appropriée pour un patient atteint de cancer, en ciblant les altérations génétiques spécifiques associées à son cancer[22].

Pour qu'une cellule saine devienne cancéreuse, les scientifiques pensent que plus d'un changement d'ADN doit se produire. Les personnes qui ont hérité d'un changement génétique lié au cancer ont besoin de moins de changements supplémentaires pour développer un cancer. Cependant, ils peuvent ne jamais développer ces changements ou développer un cancer[23].

Au fur et à mesure que les cellules cancéreuses se divisent, elles acquièrent plus de modifications de l'ADN au fil du temps. Deux cellules cancéreuses dans la même tumeur peuvent avoir des changements d'ADN différents. De plus, chaque personne atteinte de cancer présente une combinaison unique de modifications de l'ADN dans son cancer.

Des modifications génétiques cancérigènes peuvent également se produire lorsque des segments d'ADN, parfois très volumineux, sont réarrangés, supprimés ou copiés. C'est ce qu'on appelle les réarrangements chromosomiques.

Par exemple, la plupart des leucémies myéloïdes chroniques (un type de cancer du sang) sont causées par un réarrangement chromosomique qui place une partie du gène BCR à côté du gène ABL. Ce réarrangement crée une protéine anormale, appelée BCR-ABL, qui provoque une croissance incontrôlée des cellules leucémiques.

D'autres modifications de l'ADN, appelées modifications épigénétiques, peuvent également causer le cancer. Contrairement aux variants génétiques, les modifications épigénétiques (parfois appelées épimutations) peuvent être réversibles et n'affectent pas le code ADN. Au lieu de cela, les changements épigénétiques affectent la façon dont l'ADN est emballé dans le noyau. En modifiant la façon dont l'ADN est conditionné, les changements épigénétiques peuvent modifier la quantité de protéines fabriquées par un gène. Certaines substances et produits chimiques dans l'environnement qui provoquent des modifications génétiques

peuvent également provoquer des modifications épigénétiques, telles que la fumée de tabac, des métaux lourds comme le cadmium et des virus comme le virus d'Epstein-Barr[24].

### I.2.2. ARN Acide ribonucléique

Les molécules d'ARN sont des acides nucléiques simples brins composés de nucléotides. L'ARN joue un rôle majeur dans la synthèse des protéines car il est impliqué dans la transcription, le décodage et la traduction du code génétique pour produire des protéines.

L'ARN joue également un rôle important dans la carcinogenèse, car il régule l'expression des gènes et la synthèse des protéines dans les cellules. Les altérations de l'ARN peuvent entraîner des perturbations dans la régulation de l'expression des gènes, ce qui peut conduire à la formation de cellules cancéreuses[25].

Il existe plusieurs types d'ARN impliqués dans la carcinogenèse, notamment l'ARN messager (ARNm), l'ARN non codant (ARNnc), l'ARN ribosomique (ARNr) et l'ARN de transfert (ARNt)[26].

L'ARNm est une copie de l'ADN qui sert de modèle pour la synthèse des protéines. Les altérations de l'ARNm peuvent affecter la quantité ou la qualité des protéines produites dans les cellules, ce qui peut contribuer à la carcinogenèse. Par exemple, certaines mutations dans l'ARNm peuvent conduire à la production de protéines oncogènes qui stimulent la croissance cellulaire. Les ARNnc ; près de 90 % de l'ADN nucléaire est transcrit en ARNs non codants, ils ne sont pas traduits en protéines, mais ils régulent l'expression des gènes en modulant la stabilité et la traduction de l'ARNm[27]. Des altérations dans les niveaux ou les fonctions des ARNnc peuvent également contribuer à la carcinogenèse. Par exemple, certains ARNnc sont connus pour agir comme des régulateurs négatifs des gènes suppresseurs de tumeurs, ce qui peut conduire à une prolifération cellulaire accrue.

Ces ARNs sont de différents types : microARNs (miARNs), longs ARNs non codants (IncARNs) ou ARNnc circulaires (circARNs)[28].

Les microARNs sont une nouvelle classe de molécules découvertes depuis une quinzaine d'années chez l'homme. Ce sont de petits ARN non-codants d'une vingtaine de nucléotides dont la fonction et de réguler négativement l'expression des gènes, essentiellement au niveau post-transcriptionnel, jouant le rôle de suppresseurs de tumeurs ou d'oncogènes selon le type et le contexte cellulaires. En s'appariant à la région 3 non traduite de leurs ARN messagers (ARNm) cibles, ils provoquent la dégradation de ces ARNm ou bloquent leur traduction. L'ARNr et l'ARNt sont également importants dans la synthèse des protéines, car ils constituent les composants structuraux des ribosomes et interviennent dans la traduction de

l'ARNm en protéines. Des perturbations dans la synthèse des protéines peuvent également contribuer à la formation de cellules cancéreuses[29, 30].

La compréhension de la régulation de l'ARN pendant la carcinogenèse est importante pour le développement de nouvelles thérapies ciblées. Des approches thérapeutiques basées sur l'ARN, telles que la thérapie antisens et la thérapie génique, sont en cours de développement pour cibler spécifiquement les altérations de l'ARN associées à différents types de cancers[30].

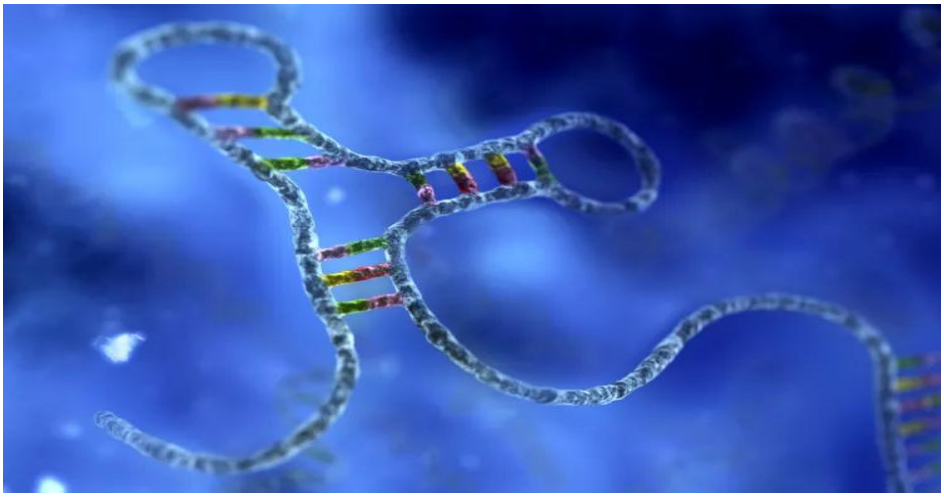


***Figure II.1 : L'interférence par l'ARN (ARNi) est un mécanisme d'expression génique qui implique de l'acide ribonucléique (ARN) à double brin. L'ARN à double brin (ou ARNdb), comme on le voit ici, peut être utilisé pour bloquer l'expression de gènes spécifiques. La voie de l'ARNi s'est développée chez les organismes unicellulaires primitifs pour empêcher la réplication réussie des virus à ARN. Les chercheurs mènent actuellement des expériences avec l'idée que l'introduction artificielle d'ARNdb provenant d'un gène de maladie peut déclencher le processus d'ARNi pour réduire au silence ce gène, tout comme il réduirait au silence les gènes d'un virus acquis naturellement. De plus, c'est un outil de recherche important, car l'introduction d'ARNdb synthétique dans les cellules peut induire la suppression de gènes spécifiques d'intérêt [31].***

### I.2.3. Les protéines

Les protéines sont de grandes molécules complexes qui jouent de nombreux rôles critiques dans le corps. Ils effectuent la majeure partie du travail dans les cellules et sont nécessaires à la structure, à la fonction et à la régulation des tissus et des organes du corps[32].

Les protéines sont constituées de centaines ou de milliers d'unités plus petites appelées acides aminés, qui sont liées les unes aux autres par de longues chaînes. Il existe 20 types d'acides aminés différents qui peuvent être combinés pour former une protéine. La séquence d'acides

aminés détermine la structure tridimensionnelle unique de chaque protéine et sa fonction spécifique. Les acides aminés sont codés par des combinaisons de trois éléments constitutifs de l'ADN (nucléotides), déterminés par la séquence des gènes[33].

### I.2.3.1. Structure des protéines

Les protéines sont composées de polymères de résidus d'acides aminés. Les acides aminés sont reliés entre eux par des liaisons peptidiques. Chaque protéine est un polymère linéaire construit à partir de différents acides aminés. Le type et la séquence des acides aminés dans une protéine sont spécifiés par l'ADN de la cellule qui les produit. Le code génétique spécifie généralement 20 acides aminés standard. Cependant, certains organismes, comme les archées, ont un code génétique qui en précise davantage. Cette séquence d'acides aminés est essentielle car elle détermine la structure globale et la fonction d'une protéine. Certaines protéines peuvent former un complexe avec une autre protéine. D'autres forment un complexe avec d'autres biomolécules autres que le peptide. Certains de ces groupes non peptidiques dans une protéine sont appelés cofacteurs ou groupes prosthétiques[34].

Il existe quatre types distincts de structure protéique : (1) structure primaire, (2) structure secondaire, (3) structure tertiaire et (4) structure quaternaire.

**La structure primaire** concerne la séquence d'acides aminés dans une chaîne polypeptidique.

**La structure secondaire** d'une protéine fait référence aux structures locales qui se répètent régulièrement et qui sont stabilisées par des liaisons hydrogène. Des exemples d'une structure secondaire sont l'hélice α, la feuille β et les virages et boucles.

**La structure tertiaire** (parfois appelée **pli**) concerne la relation spatiale des structures secondaires les unes par rapport aux autres. Il est souvent stabilisé par des interactions non locales, par ex. par des liaisons disulfure, des ponts salins et la formation d'un noyau hydrophobe. Cette structure de la protéine est ce qui détermine la fonction fondamentale de la protéine.

**La structure quaternaire** fait référence au complexe protéique, c'est-à-dire lorsque plusieurs chaînes polypeptidiques (parfois appelées sous-unités protéiques) le constituent. Les structures tertiaires et quaternaires sont souvent appelées conformations. La protéine peut passer d'une structure à une autre et la transition entre tertiaire et quaternaire par un changement conformationnel. Des changements de conformation peuvent être induits par

exemple lorsqu'un substrat se lie à une protéine (en particulier, une enzyme) au niveau de son site actif[34].

### I.2.3.2. Biosynthèse des Protéines

Dans les cellules, les protéines sont produites par des processus de transcription et de traduction. La transcription est le processus dans lequel l'ARNm est transcrit à partir de l'ADN pour servir de matrice pour la traduction. La traduction est le processus par lequel les acides aminés sont liés entre eux dans un ordre spécifique selon les règles spécifiées par le code génétique[35, 36]. Il se produit dans le cytoplasme où se trouvent les ribosomes. Les étapes de la traduction sont :

(1) la liaison de l'ARNm à un ribosome,

(2) le ribosome commence à faire correspondre les séquences d'anticodon de l'ARNt au codon de l'ARNm,

(3) l'acide aminé porté par l'ARNt est ajouté à la chaîne allongée,

et (4) un codon stop est atteint, libérant le polypeptide et l'ARNm.

Les événements qui suivent la synthèse des protéines comprennent la protéolyse, la modification post-traductionnelle et le repliement des protéines.

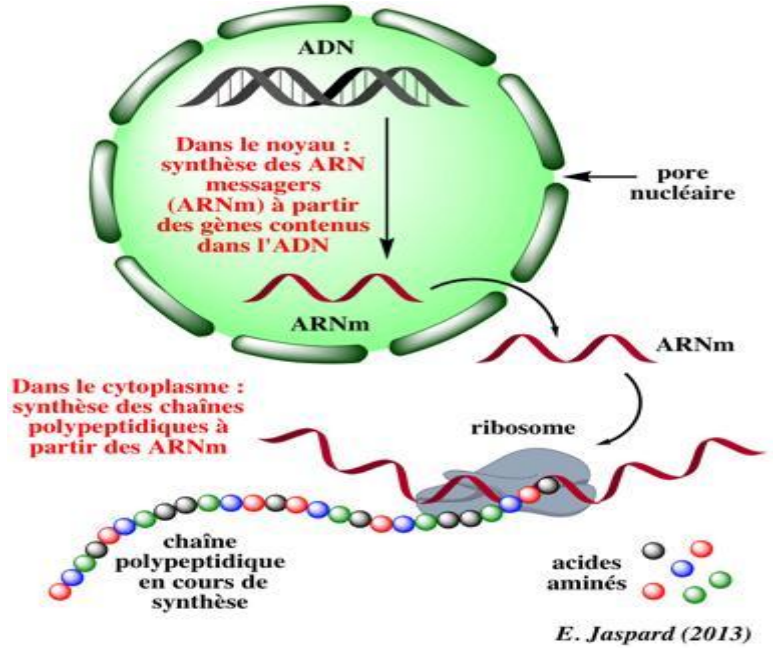



***Figure I.2: Production d'une chaine polypeptidique à partir d'une chaine d'ADN[37]***

La durée de vie des protéines varie selon leur nature. Certaines protéines sont dégradées après quelques minutes de production ; d'autres restent pendant des années. Les protéines mal repliées sont immédiatement dégradées pour éviter qu'elles ne causent des dommages en raison de leur instabilité et de leur nature dysfonctionnelle.

Les protéines ont plusieurs fonctions. Certains d'entre eux sont des composants structurels (par exemple, la kératine dans les cheveux, l'actine et la myosine dans les muscles, etc.). De nombreuses protéines sont des enzymes. Ils catalysent diverses réactions biochimiques et sont donc essentiels au métabolisme. Les autres rôles vitaux des protéines dans les systèmes biologiques sont les suivants : en tant que transporteurs (par exemple, l'hémoglobine), en tant qu'anticorps et en tant que régulateurs de l'expression des gènes. Toutes les protéines ne sont pas biosynthétisées de novo. Ils peuvent être acquis par l'alimentation[36].

### I.2.3.3. Les protéines et la carcinogenèse

Les cellules cancéreuses présentent souvent des altérations dans l'expression et la fonction des protéines par rapport aux cellules normales. Ces altérations peuvent être le résultat de mutations dans les gènes qui codent pour les protéines, de modifications post-traductionnelles des protéines ou d'autres modifications moléculaires[38].

- Les protéines impliquées dans la régulation du cycle cellulaire sont souvent altérées dans les cellules cancéreuses. Par exemple, **les protéines cyclines et les kinases cycline-dépendantes (CDK)** régulent l'avancement du cycle cellulaire et sont souvent dysfonctionnelles dans les cellules cancéreuses. Les mutations dans les gènes qui codent pour les cyclines ou les CDK peuvent entraîner une prolifération cellulaire incontrôlée et la formation de tumeurs .
- Les protéines impliquées dans **la réparation de l'ADN** sont également importantes dans la biologie moléculaire du cancer. Les cellules cancéreuses présentent souvent des dommages à leur ADN, mais elles peuvent également avoir des altérations dans les protéines qui sont impliquées dans la réparation de l'ADN. Par exemple, les mutations dans les gènes BRCA1 et BRCA2 qui codent pour des protéines impliquées dans la réparation de l'ADN ont été associées à un risque accru de cancer du sein et de l'ovaire.
- Les protéines impliquées dans **la signalisation cellulaire** peuvent être dysfonctionnelles dans les cellules cancéreuses. Par exemple, les récepteurs de facteurs de croissance, tels que le récepteur du facteur de croissance épidermique (EGFR), sont souvent sur-régulés dans les cellules cancéreuses. Cela peut conduire à une prolifération cellulaire incontrôlée et à la formation de tumeurs.

- Les protéines impliquées dans **l'apoptose** sont également importantes dans la biologie moléculaire du cancer. Les cellules cancéreuses peuvent éviter l'apoptose en altérant les protéines impliquées dans cette voie de signalisation. Par exemple, la protéine p53, souvent appelée "gardien du génome", joue un rôle crucial dans la régulation de l'apoptose. Les mutations dans le gène TP53 qui codent pour la protéine p53 sont fréquentes dans les cellules cancéreuses et peuvent contribuer à leur survie et leur prolifération.
En résumé, la biologie moléculaire des protéines est étroitement liée au processus de cancer. Les altérations dans l'expression et la fonction des protéines sont souvent impliquées dans la prolifération cellulaire incontrôlée, la formation de tumeurs, la résistance à l'apoptose et d'autres caractéristiques des cellules cancéreuses. Comprendre ces altérations moléculaires peut aider à développer de nouvelles thérapies ciblées pour le traitement du cancer[30].

### I.2.4. Expression génétique

L'expression génique est le processus par lequel l'ADN est utilisé pour fabriquer des protéines, qui remplissent ensuite diverses fonctions importantes dans le corps. La protéine pourrait être une enzyme, une hormone ou un récepteur, par exemple[40].

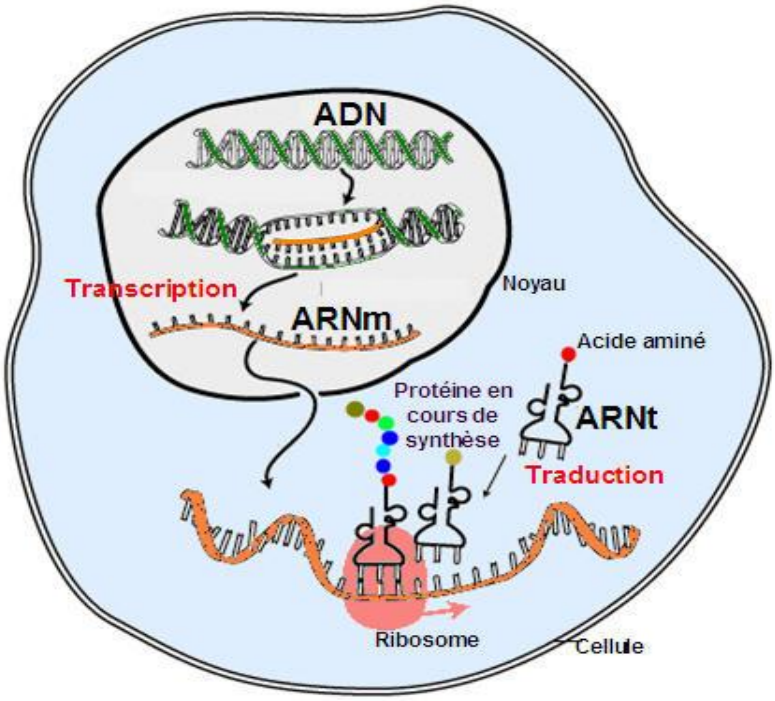



***Figure I.3: le processus de synthèse protéique[41].***

Le processus d'expression des gènes comprend les étapes suivantes[42]:

**Transcription :** La transcription est le processus par lequel un segment d'ADN est utilisé pour générer une matrice d'ARN. Le segment d'ADN est « lu » par une enzyme appelée ARN polymérase, qui produit un brin d'ARN complémentaire à l'ADN. Dans ce brin d'ARN complémentaire, toutes les bases de thymine sont remplacées par de l'uracile.

**Traitement :** Ce transcrit d'ARN primaire est ensuite modifié pour le convertir en ARN messager mature (ARNm) pouvant être utilisé dans la traduction. L'ARNm subit un épissage pour éliminer les parties non codantes du transcrit (introns) afin qu'il ne reste que les sections codantes (exons).

**Maturation de l'ARN non codant :** Les régions non codantes de l'ARN (ARNnc) sont transcrites sous forme de précurseurs qui sont ensuite traités davantage. Par exemple, ces régions peuvent être transcrites sous forme d'ARN pré-ribosomique (pré-ARNr) qui subit ensuite un clivage pour devenir de l'ARN ribosomal (ARNr).

**Exportation d'ARN :** La majorité de l'ARN mature est ensuite transportée du noyau vers le cytoplasme. Bien que certains ARN fonctionnent dans le noyau, la plupart sont transportés à travers les pores du noyau dans le cytosol, y compris tous les ARN impliqués dans la synthèse des protéines.

**Traduction :** L'ARNm final contient les informations nécessaires pour coder les protéines. Toutes les trois paires de bases sur l'ARNm correspondent à un site de liaison pour un ARN de transfert (ARNt) qui porte un acide aminé. Les acides aminés sont ensuite liés ensemble dans une chaîne par un ribosome pour créer une chaîne protéique rudimentaire.

**Repliement des protéines :** La longue chaîne d'acides aminés se replie pour former une structure tridimensionnelle à l'aide d'enzymes appelées chaperons. Cette structure tridimensionnelle est la forme fonctionnelle finale de la protéine.

**Altérations de l'expression génétique**

L'expression génétique fait référence à la façon dont l'information génétique stockée dans l'ADN est utilisée par une cellule pour produire des protéines et accomplir différentes fonctions cellulaires. Dans le cas du cancer, des altérations de l'expression génétique peuvent se produire et jouer un rôle crucial dans le développement et la progression de la maladie.

Il existe plusieurs mécanismes qui peuvent conduire à des altérations de l'expression génétique dans le cancer. Les deux principaux sont les mutations génétiques et les modifications épigénétiques[43].

**Mutations génétiques :** Les mutations peuvent affecter les gènes qui régulent la croissance cellulaire, la division et la mort cellulaire programmée. Ces mutations peuvent être héritées ou acquises au cours de la vie. Certaines mutations génétiques peuvent activer des oncogènes, qui sont des gènes qui favorisent la croissance cellulaire incontrôlée, tandis que d'autres mutations peuvent désactiver des gènes suppresseurs de tumeurs, qui sont des gènes qui régulent normalement la croissance cellulaire[44].

**Modifications épigénétiques :** Les modifications épigénétiques sont des changements réversibles qui affectent la structure de l'ADN et peuvent influencer l'expression des gènes sans modifier la séquence d'ADN elle-même. Les modifications épigénétiques comprennent la méthylation de l'ADN (ajout de groupes méthyle à l'ADN), les modifications des histones (protéines qui aident à empaqueter l'ADN dans la chromatine) et d'autres changements qui affectent l'accessibilité de l'ADN aux facteurs de transcription. Ces modifications épigénétiques peuvent entraîner une activation anormale ou une désactivation de certains gènes impliqués dans la régulation de la croissance cellulaire et de la survie[45].

Ces altérations de l'expression génétique peuvent conduire à une prolifération cellulaire incontrôlée, à l'évasion des mécanismes de contrôle de la croissance cellulaire, à l'inhibition de la mort cellulaire programmée (apoptose) et à la formation de tumeurs. Les recherches sur les altérations de l'expression génétique dans le cancer sont en cours pour mieux comprendre les mécanismes sous-jacents et identifier de nouvelles cibles thérapeutiques.

Il est important de noter que le cancer est une maladie complexe et hétérogène, et que différentes altérations génétiques et épigénétiques peuvent être présentes chez différents patients atteints de différents types de cancer. Par conséquent, une approche de traitement personnalisée basée sur les caractéristiques génétiques et épigénétiques spécifiques d'un patient est de plus en plus utilisée pour améliorer les résultats thérapeutiques.

### I.2.5. Gène

Le gène est l'unité physique de base de l'hérédité. Les gènes sont transmis des parents à la progéniture et contiennent les informations nécessaires pour spécifier les traits. Les gènes sont

disposés les uns après les autres sur des structures appelées chromosomes. Un chromosome contient une seule et longue molécule d'ADN - dont une partie seulement correspond à un seul gène - ainsi que les protéines structurelles (appelées histones) autour desquelles la molécule d'ADN s'enroule. Les humains ont environ 20 000 gènes disposés sur leurs chromosomes[46].

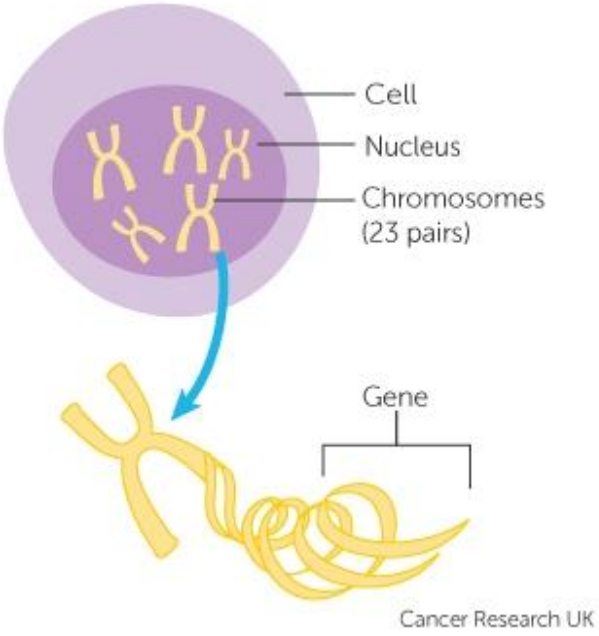



***Figure I.4: Le génome dans la cellule[47].***

Nos gènes détectent les erreurs qui se produisent lorsque les cellules se divisent. Ces erreurs (ou fautes) sont appelées mutations. Des mutations peuvent se produire tout au long de notre vie, lors de processus naturels dans nos cellules. Ou ils peuvent se produire en raison d'autres facteurs tels que :

- Fumée de tabac
- Rayonnement à haute énergie (ionisant), comme les rayons X
- Rayonnement ultraviolet du soleil
- Certaines substances dans les aliments
- Produits chimiques dans notre environnement

Parfois, les gens héritent de certains gènes défectueux de leurs parents. Cela peut leur donner un risque accru de cancer.

Il est habituel que les cellules réparent les défauts de leurs gènes. Lorsque les dégâts sont très importants, la cellule peut s'autodétruire à la place. Ou le système immunitaire peut les reconnaître comme anormaux et les tuer. Cela nous aide à nous protéger du cancer.

Parfois, des mutations dans des gènes importants font qu'une cellule ne comprend plus les instructions. La cellule peut commencer à se multiplier hors de contrôle. Il ne se répare pas correctement et ne meurt pas quand il le devrait. Cela peut conduire au cancer.

Il existe 4 principaux types de gènes impliqués dans la division cellulaire. La plupart des tumeurs ont des copies défectueuses de plus d'un de ces types:

- **Les oncogènes** sont des gènes qui, dans des circonstances normales, ordonnent aux cellules de se multiplier et de se diviser. Chez les adultes, cela n'arrive pas très souvent.

Nous pouvons considérer les oncogènes comme étant un peu comme la pédale d'accélérateur d'une voiture. Lorsqu'ils deviennent actifs, ils accélèrent le taux de croissance d'une cellule. Quand l'un est endommagé, c'est comme si la pédale d'accélérateur se coinçait. Cette cellule, et toutes les cellules qui se développent à partir d'elle, sont en permanence chargées de se diviser. Alors un cancer se développe.

- **Gènes qui empêchent la cellule de se multiplier (gènes suppresseurs de tumeurs)**

Il est habituel que les cellules réparent les défauts de leurs gènes. Lorsque les dommages sont très importants, les gènes suppresseurs de tumeurs peuvent arrêter la croissance et la division des cellules.

Des mutations dans les gènes suppresseurs de tumeurs signifient qu'une cellule ne comprend plus l'instruction d'arrêter de croître. La cellule peut alors commencer à se multiplier de manière incontrôlable. Cela peut conduire au cancer.

Le gène suppresseur de tumeur le plus connu est p53. Les chercheurs savent que le gène p53 est endommagé ou absent dans la plupart des cancers.

- **Gènes qui réparent d'autres gènes endommagés (gènes de réparation de l'ADN)**

L'ADN de chaque cellule de notre corps risque constamment d'être endommagé. Mais les cellules contiennent de nombreuses protéines différentes dont le travail consiste à réparer l'ADN endommagé. La plupart des dommages à l'ADN sont réparés immédiatement grâce à ces protéines.

Mais si les dommages à l'ADN se produisent sur un gène qui fabrique une protéine de réparation de l'ADN, une cellule a moins de capacité à se réparer. Ainsi, des erreurs s'accumuleront dans d'autres gènes au fil du temps et permettront à un cancer de se former.

Les scientifiques ont découvert des gènes de réparation de l'ADN endommagés dans certains cancers, notamment le cancer de l'intestin.

- **Gènes qui disent à une cellule de mourir (gènes d'autodestruction)**

Certains gènes disent à une cellule de s'autodétruire si elle est devenue trop vieille ou endommagée. C'est ce qu'on appelle l'apoptose ou la mort cellulaire programmée. C'est un processus très complexe et important. Les cellules meurent généralement chaque fois que quelque chose ne va pas, pour empêcher la formation d'un cancer.

Il existe de nombreux gènes et protéines différents impliqués dans l'apoptose. Si ces gènes sont endommagés, une cellule défectueuse peut survivre plutôt que mourir et devenir cancéreuse.

### I.2.6. Génome

Le génome est l'ensemble complet de l'ADN d'un organisme, il englobe tous les gènes et les séquences d'ADN non codantes qui régulent l'expression des gènes. Le génome humain est composé d'environ 3 milliards de paires de bases d'ADN et contient des dizaines de milliers de gènes[39].

Le génome humain, comme les génomes de tous les autres animaux vivants, est une collection de longs polymères d'ADN. Ces polymères sont maintenus en double copie sous forme de chromosomes dans chaque cellule humaine et codent dans leur séquence de bases constitutives ( guanine : G, adénine : A, thymine : T et cytosine : C) les détails des caractéristiques moléculaires et physiques qui forment l'organisme correspondant. La séquence de ces polymères, leur organisation et leur structure, ainsi que les modifications

chimiques qu'ils contiennent fournissent non seulement la machinerie nécessaire pour exprimer les informations contenues dans le génome, mais fournissent également au génome la capacité de se répliquer, de se réparer, de se conditionner et de se maintenir. De plus, le génome est essentiel à la survie de l'organisme humain ; sans elle, aucune cellule ou tissu ne pourrait vivre au-delà d'une courte période de temps. Par exemple, les globules rouges ( érythrocytes ), qui ne vivent que 120 jours environ, et les cellules de la peau , qui ne vivent en moyenne que 17 jours environ, doivent être renouvelées pour maintenir la viabilité du corps humain ., et c'est au sein du génome que se trouvent les informations fondamentales pour le renouvellement de ces cellules, et de bien d'autres types de cellules[48].

- Le génome humain n'est pas uniforme. À l'exception des jumeaux identiques (monozygotes), il n'y a pas deux humains sur Terre qui partagent exactement la même séquence génomique. De plus, le génome humain n'est pas statique. Des changements subtils et parfois moins subtils surviennent avec une fréquence surprenante. Certains de ces changements sont neutres ou même avantageux ; ceux-ci sont transmis de parent à enfant et finissent par devenir monnaie courante dans la population. D'autres changements peuvent être préjudiciables , entraînant une réduction de la survie ou une diminution de la fertilité des individus qui les hébergent ; ces changements ont tendance à être rares dans la population. Le génome de l'homme moderne est donc un enregistrement des épreuves et des succès des générations qui l'ont précédé. Reflété dans la variation du génome moderne est la gamme de diversitéqui sous-tend les traits typiques de l'espèce humaine. Il existe également des preuves dans le génome humain du fardeau continu des variations préjudiciables qui conduisent parfois à la maladie.
- La connaissance du génome humain permet de comprendre l'origine de l'espèce humaine, les relations entre les sous-populations humaines et les tendances en matière de santé ou les risques de maladie des individus. En effet, au cours des 20 dernières années, la connaissance de la séquence et de la structure du génome humain a révolutionné de nombreux domaines d'étude, notamment la médecine , l'anthropologie et la médecine légale . Avec les progrès technologiques qui permettent un accès peu coûteux et élargi aux informations génomiques, la quantité et les applications potentielles des informations extraites du génome humain sont extraordinaires[48].

### I.2.7. Biomarqueurs

Les biomarqueurs (abréviation de marqueurs biologiques) « Le biomarqueur défini comme une caractéristique qui est objectivement mesurée et évaluée comme un indicateur d'un processus biologique normale, pathologique ou réponse à une intervention thérapeutique »[49].

D'après cette définition, les biomarqueurs peuvent être divisés en trois types principaux:

- Type « 0 » : biomarqueurs utilisés pour estimer l'émergence et le développent d'une maladie.

- Type « 1 » : biomarqueurs qui permettent de prédire les réponses aux interventions thérapeutiques.

- Type « 2» : biomarqueurs qui, en principe, pourraient être utilisés comme critères d'évaluation dans le cadre d'essais cliniques.

Selon leur application à la détection des maladies, trois classes principales de biomarqueurs peuvent être spécifiées : les biomarqueurs de criblage, de diagnostic et de pronostic:

-Les biomarqueurs de criblage, sont utilisés pour prédire l'apparition éventuelle de la maladie chez les patients asymptomatiques.

- Les biomarqueurs de diagnostic, sont utilisés pour faire des prédictions sur des patients suspectés d'avoir la maladie.

- Les biomarqueurs de pronostic, sont utilisés pour prédire le résultat d'un patient souffrant d'une maladie.

Les biomarqueurs peuvent également être considérés comme des indicateurs de changements fonctionnels et structurels dans les organes et les cellules. Ces modifications peuvent être associées soit à des facteurs de causalité ou des conséquences d'événements normaux et pathologiques.

En outre, les biomarqueurs peuvent être considérés comme des cibles thérapeutiques potentielles, par exemple lorsque leur rôle causal dans la maladie est démontré.

Les biomarqueurs peuvent être des gènes, des protéines, des peptides (morceaux de protéines) ou des métabolites dont leur niveau est en train de changer en cas de maladies et ils peuvent ensuite être utilisés pour déterminer le stade de la maladie chez un patient. Sont des mesures biologiques d'un état biologique. Par définition, un biomarqueur est "une caractéristique objectivement mesurée et évaluée comme un indicateur de processus biologiques normaux, de processus pathogènes ou de réponses pharmacologiques à une intervention thérapeutique".
Les biomarqueurs peuvent être utilisés seuls ou en combinaison pour évaluer l'état de santé ou la maladie d'un individu.

### I.2.7.1. Caractéristiques d'un biomarqueur idéal

Un biomarqueur idéal possède certaines caractéristiques qui le rendent approprié pour vérifier une maladie particulière[50]. Idéalement, un marqueur idéal devrait avoir les caractéristiques suivantes :

- Sûr et facile à mesurer
- Suivi économique
- Modifiable avec traitement
- Cohérence entre les sexes et les groupes ethniques

### I.2.7.2. Les biomarqueurs en cancérologie

Lorsque nous parlons de biomarqueurs cancéreux, nous se réfère généralement aux protéines, gènes et autres molécules qui affectent la croissance, la multiplication et la lyse des cellules cancéreuses. Les biomarqueurs du cancer sont donc des outils précieux pour la détection de la maladie, le diagnostic, le pronostic et le choix du traitement.
La mise en place des biomarqueurs nécessite une compréhension globale des mécanismes moléculaires et processus cellulaire qui profondes l'initiation du cancer, en particulier en mettant l'accent sur la façon dont de petits changements dans seulement quelques gènes ou des protéines régulatrices peuvent perturber une variété de fonctions cellulaires et lancer le cancer[51].
Un défi majeur dans le diagnostic du cancer est d'abord de découvrir les biomarqueurs qui représentent bien cette maladie et d'établir la relation exacte entre ces biomarqueurs et la pathologie clinique, ainsi que, pour être en mesure de détecter des tumeurs non invasive à un stade précoce.

Les Biomarqueurs du cancer sont des traits quantifiables qui aident les oncologues lors de la première interaction avec les patients suspects à :

- identifier les personnes à risque,
  - diagnostiquer à un stade,

- choisir la meilleure modalité de traitement,
  - surveiller la réponse au traitement (comment ce cancer réagit- il à ce traitement ?). Traditionnellement, les médicaments anticancéreux étaient des agents qui tuaient à la fois les cellules cancéreuses et les cellules saines. Cependant, des thérapies plus ciblées ont maintenant été développées qui peuvent être dirigées pour tuer uniquement les cellules cancéreuses, tout en épargnant les cellules saines[52].

Dans la recherche sur le cancer, les études génétiques sont précieuses car les anomalies génétiques sous-tendent si souvent le développement du cancer. Certains marqueurs ADN ou ARN peuvent donc aider à la détection et au traitement de cancers spécifiques.

Les biomarqueurs peuvent être des gènes, des protéines, des peptides (morceaux de protéines) ou des métabolites dont leur niveau est en train de changer en cas de maladies et ils peuvent ensuite être utilisés pour déterminer le stade de la maladie chez un patient.

***Tableau I.1: Exemple des cancers et leurs biomarqueurs***

| **Cancer** | **Les biomarqueurs** |
|---|---|
| Cancer de la prostate | PSA (prostate specific antigen) |
| Cancer du pancréas, du côlon, du rectum, de l'estomac. | gène CA 19.9 |
| Cancer digestifs, ovariens, thyroïdiens. | gène CA 72,4 |
| Cancer du sein | gène CA 15,3 |

## I.3. Cancérogénèse

La cancérogenèse peut résulter de l'action d'une combinaison d'agressions chimiques, physiques, biologiques et/ou génétiques contre les cellules. L'initiation et le développement du processus tumoral sont liés à l'accumulation dans une même cellule de mutations génétiques et de modifications épigénétiques qui interagissent de manière coopérative. Ces évènements moléculaires qui se succèdent de manière aléatoire expliquent les caractères multi-étapes de l'apparition du cancer[23].

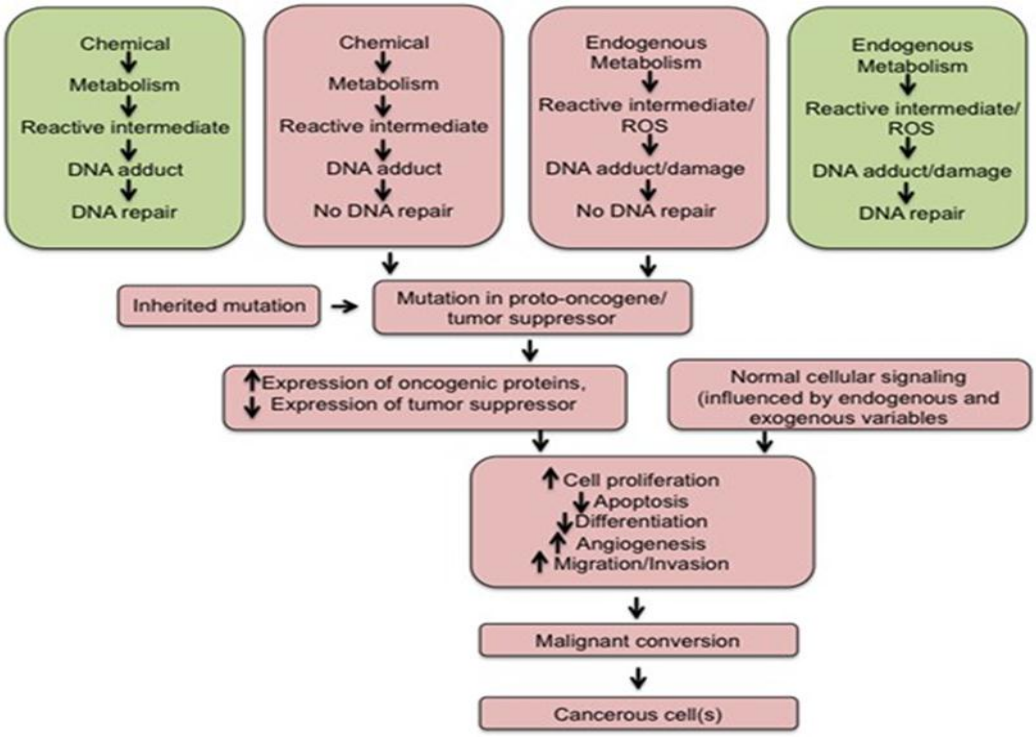



***Figure I.5: Le développement du processus tumoral [53].***

Le processus de cancérogenèse peut être divisé en au moins trois étapes : initiation, promotion et progression. Les cibles moléculaires critiques au cours des étapes de la carcinogenèse comprennent les proto-oncogènes, les oncogènes cellulaires et les gènes suppresseurs de tumeurs, les altérations des deux allèles de ces derniers n'étant trouvées qu'au stade de la progression. Bien que bon nombre de ces gènes cibles critiques aient été identifiés, le nombre ultime et les caractéristiques des altérations moléculaires qui définissent les néoplasies n'ont pas été élucidés.

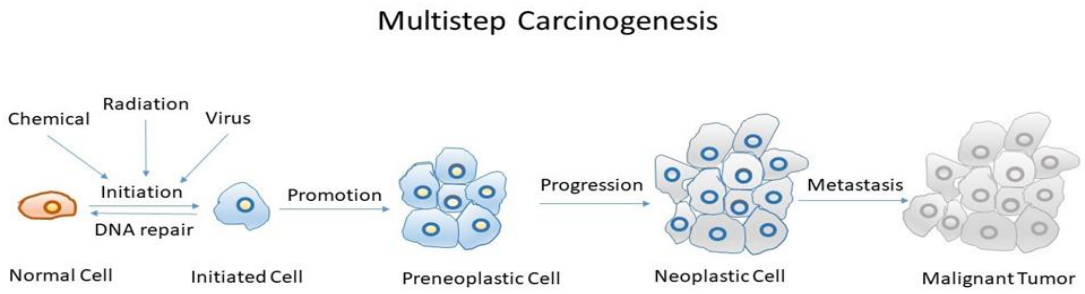



***Figure I.6: Les étapes du processus de cancérogenèse[53].***

Le processus de cancérogenèse comporte trois étapes. Il faut savoir qu'il peut s'écouler entre 20 et 30 ans depuis la première étape (l'initiation) et la transformation maligne d'une cellule.

**1 : Initiation :** première étape caractérisée par l'exposition à un agent toxique. Elle consiste en une dysrégulation génomique aboutissant à une transformation cellulaire. L'agent carcinogène initial entraine des lésions génétiques cellulaires irréversibles mais insuffisantes pour entrainer la transformation.

**2 : Promotion :** Stade réversible caractérisé par une expression anormale des gènes, la cellule a acquis un avantage prolifératif donnant naissance à un clone, cette cellule se développe rapidement tout en acquérant une instabilité génomique qui la rend plus vulnérable aux agressions aboutissant à de nouvelles mutations et des altérations épigénétiques. Il en résulte une perte de l'homéostasie tissulaire et l'émergence de clones cellulaires transformés.

**3 : Progression :** aboutit à la formation de la première cellule cancéreuse. Stade irréversible au cours duquel le cancer devient apparent. Elle comporte des altérations génétiques majeures qui évoluent constamment en raison d'une instabilité chromosomique. Ces anomalies vont avec l'acquisition des différentes caractéristiques du cancer : différenciation, invasion, capacité métastatique, échappement au control immunitaire.

## I.4. Conclusion

Dans ce chapitre, nous avons résumé l'essentiel de la biologie moléculaire. Ensuite, nous avons présenté quelques notions et terminologies biologiques afin de comprendre tous ce qui suit dans notre recherche. Nous avons montré dans cette section le dogme central de la biologie moléculaire qui est les trois macromolécules cellulaire (ADN, ARN et protéine). Finalement, nous avons élucidé le processus de carcinogénèse, et nous avons compris que la

compréhension de la carcinogenèse repose sur l'étude de l'ADN, des gènes et du génome, ainsi que des mécanismes moléculaires impliqués dans la régulation de la croissance cellulaire et la progression tumorale. Les progrès dans ce domaine ont permis d'identifier de nombreux facteurs de risque pour le cancer et d'élucider les mécanismes moléculaires sous-jacents à la transformation maligne des cellules.

La recherche en carcinogenèse vise à développer des stratégies de prévention plus efficaces, en identifiant et en évitant les facteurs de risque évitables, ainsi qu'à améliorer les méthodes de dépistage précoce du cancer. De plus, une meilleure compréhension des mécanismes de la carcinogenèse permet d'explorer de nouvelles voies thérapeutiques ciblées pour le traitement du cancer.

L’analyse et l’interprétation des données de la biologie moléculaire nécessitent l’intervention d’une nouvelle discipline nommée la bioinformatique. Dans le chapitre suivant, nous présenterons cette discipline et ses apports à la biologie moléculaire.

# Chapitre II:
# **Bioinformatique**

## II.1. Introduction

La bioinformatique est une discipline scientifique interdisciplinaire qui applique les technologies de l'information, notamment l'informatique, à la gestion et à l'analyse des données biologiques[54]. La bioinformatique est devenue indispensable pour gérer et analyser les grandes quantités de données générées par les nouvelles technologies de séquençage d'ADN et le développement d'autres techniques à haut débit en biologie moléculaire et cellulaire.

Les objectifs principaux de la bioinformatique sont : le stockage, la recherche et l'analyse des données biologiques telles que les séquences d'ADN, les structures des protéines et les voies métaboliques[55]. Les domaines d'application de la bioinformatique sont très vastes, allant de la génomique et la génétique moléculaire à la biologie des systèmes, en passant par la protéomique, la transcriptomique, l'immunoinformatique, etc.[55, 56].

Les méthodes bioinformatiques les plus utilisées sont basées sur l'homologie de séquence et la comparaison de caractéristiques. L'homologie de séquence consiste à aligner et comparer des séquences nucléotidiques ou peptidiques pour identifier des régions conservées et calculer le degré de similarité entre elles[57]. La comparaison de caractéristiques permet de prédire la structure et la fonction des biomolécules à partir de leur séquence à l'aide des techniques d'apprentissage automatique[58].

Récemment, l'intelligence artificielle, la machine learning et le deep learning ont également été appliqués à la bioinformatique. Ces approches d'apprentissage automatique ont permis d'améliorer la prédiction de nombreuses caractéristiques biologiques telles que la structure des protéines[59], les interactions entre biomolécules[60], la classification et le diagnostic moléculaire des maladies[61], etc.

L'ensemble de ces approches et méthodes font de la bioinformatique un outil puissant pour la recherche fondamentale et appliquée en sciences de la vie et de la santé. Ses applications sont en pleine expansion et promettent de révolutionner la biologie et la médecine dans les années à venir.

## II.2. Bioinformatique

La bioinformatique est un domaine interdisciplinaire qui développe des méthodes et des outils logiciels pour comprendre les données biologiques, en particulier lorsque les ensembles de données sont grands et complexes. En tant que domaine scientifique interdisciplinaire, la bioinformatique combine la biologie, la chimie, la physique, l'informatique, le génie de l'information, les mathématiques et la statistique pour analyser et interpréter les données

biologiques. La bioinformatique est utilisée pour les analyses in silico des questions biologiques en utilisant des techniques computationnelles et statistiques[62].

La bioinformatique comprend des études biologiques qui utilisent la programmation informatique comme partie de leur méthodologie, ainsi que des "pipelines" d'analyse spécifiques qui sont utilisés de manière répétée, en particulier dans le domaine de la génomique. Les utilisations courantes de la bioinformatique incluent l'identification de gènes candidats et de polymorphismes mononucléotidiques (SNP). Souvent, cette identification est réalisée dans le but de mieux comprendre les bases génétiques des maladies, les adaptations uniques, les propriétés souhaitables (en particulier chez les espèces agricoles) ou les différences entre les populations. De manière moins formelle, la bioinformatique cherche également à comprendre les principes organisationnels des séquences d'acides nucléiques et de protéines, appelés protéomique.

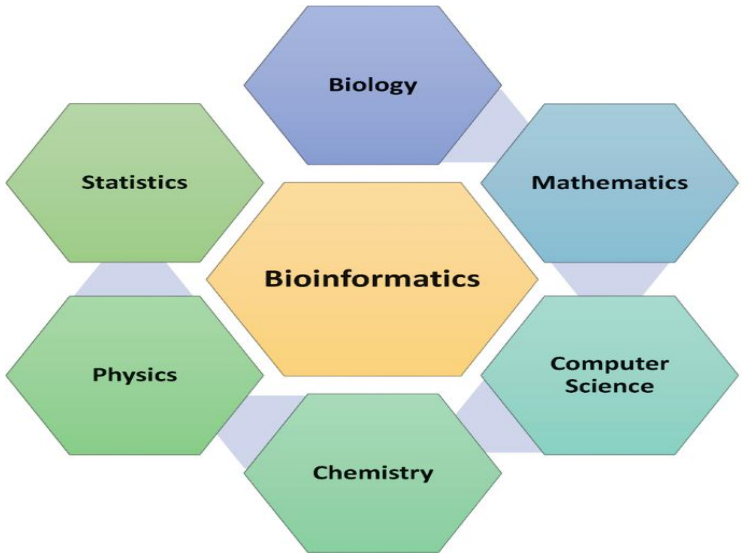



***Figure II.1 : La bioinformatique est une combinaison de biologie, d'informatique, de mathématiques, de statistiques, de physique et de principes d'ingénierie qui aide à analyser et interpréter les données biologiques (en particulier celles qui sont volumineuses et complexes) à l'aide d'outils logiciels[62].***

Le traitement des images et des signaux permet d'extraire des résultats utiles à partir de grandes quantités de données brutes. Dans le domaine de la génétique, il facilite le séquençage et l'annotation des génomes et de leurs mutations observées. Il joue un rôle dans l'extraction d'informations à partir de la littérature biologique et dans le développement d'ontologies biologiques et géniques pour organiser et interroger les données biologiques. Il

joue également un rôle dans l'analyse de l'expression et de la régulation des gènes et des protéines. Les outils de bioinformatique aident à comparer, analyser et interpréter les données génétiques et génomiques, et plus généralement à comprendre les aspects évolutifs de la biologie moléculaire. À un niveau plus intégratif, il aide à analyser et à cataloguer les voies biologiques et les réseaux qui sont une part importante de la biologie des systèmes. En biologie structurale, il facilite la simulation et la modélisation de l'ADN[10], de l'ARN[63,64], des protéines ainsi que des interactions biomoléculaires[65-68].

### II.2.1. Objectifs

Afin d'étudier comment les activités cellulaires normales sont altérées dans différents états pathologiques, il est nécessaire de combiner les données biologiques pour obtenir une image complète de ces activités. Par conséquent, le domaine de la bioinformatique s'est développé de telle sorte que la tâche la plus pressante concerne maintenant l'analyse et l'interprétation de divers types de données. Cela comprend notamment les séquences nucléotidiques et d'acides aminés, les domaines protéiques et les structures protéiques. Le processus réel d'analyse et d'interprétation des données est appelé biologie computationnelle[69]. Les sous-disciplines importantes de la bioinformatique et de la biologie computationnelle comprennent :

- Le développement et la mise en œuvre de programmes informatiques qui permettent un accès efficace, une gestion et une utilisation de divers types d'informations.
- Le développement de nouveaux algorithmes (formules mathématiques) et de mesures statistiques qui évaluent les relations entre les membres de grands ensembles de données. Par exemple, il existe des méthodes pour localiser un gène dans une séquence, prédire la structure et/ou la fonction des protéines, et regrouper les séquences de protéines en familles de séquences apparentées.

L'objectif principal de la bioinformatique est d'accroître la compréhension des processus biologiques. Ce qui la distingue des autres approches, cependant, c'est son accent mis sur le développement et l'application de techniques intensives en calcul pour atteindre cet objectif. Des exemples comprennent la reconnaissance de motifs, l'extraction de données, les algorithmes d'apprentissage automatique et la visualisation. Les principaux efforts de recherche dans le domaine comprennent l'alignement des séquences, la prédiction de gènes, l'assemblage de génomes, la conception de médicaments, la découverte de médicaments, l'alignement des structures protéiques, la prédiction de la structure des protéines, la prédiction

de l'expression des gènes et des interactions protéine-protéine, les études d'association à l'échelle du génome, la modélisation de l'évolution et de la division cellulaire/mitose.

La bioinformatique implique désormais la création et le développement de bases de données, d'algorithmes, de techniques informatiques et statistiques, ainsi que de théories pour résoudre des problèmes formels et pratiques liés à la gestion et à l'analyse des données biologiques.

Au cours des dernières décennies, les développements rapides dans les technologies de recherche moléculaire, notamment génomique, ainsi que les développements dans les technologies de l'information, ont permis de produire une quantité considérable d'informations relatives à la biologie moléculaire. La bioinformatique est le nom donné à ces approches mathématiques et informatiques utilisées pour comprendre les processus biologiques.

Les activités courantes en bioinformatique comprennent la cartographie et l'analyse des séquences d'ADN et de protéines, l'alignement des séquences d'ADN et de protéines

### II.2.2. Relation avec d'autres domaines

La bioinformatique est un domaine scientifique similaire mais distinct de l'informatique biologique, bien qu'elle soit souvent considérée comme synonyme de la biologie computationnelle. La computation biologique utilise la bioingénierie et la biologie pour construire des ordinateurs biologiques, tandis que la bioinformatique utilise le calcul pour mieux comprendre la biologie. La bioinformatique et la biologie computationnelle impliquent l'analyse de données biologiques, en particulier les séquences d'ADN, d'ARN et de protéines. Le domaine de la bioinformatique a connu une croissance exponentielle à partir du milieu des années 1990, principalement grâce au Projet du génome humain et aux avancées rapides de la technologie de séquençage de l'ADN.

L'analyse des données biologiques pour produire des informations significatives implique l'écriture et l'exécution de programmes informatiques qui utilisent des algorithmes issus de la théorie des graphes, de l'intelligence artificielle, de l'informatique molle, de l'exploration de données, du traitement d'images et de la simulation informatique. Les algorithmes dépendent à leur tour de bases théoriques telles que les mathématiques discrètes, la théorie du contrôle, la théorie des systèmes, la théorie de l'information et la statistique.

La bioinformatique comprend deux grandes méthodes pour analyser les données biologiques : les méthodes par homologie et les méthodes par caractéristique.

### II.2.3. Les méthodes bioinformatiques par homologie

Les méthodes bioinformatiques par homologie sont largement utilisées pour inférer des informations sur des séquences biologiques inconnues en les comparants à des séquences déjà connues. Ces méthodes reposent sur le principe que des séquences biologiques similaires ont souvent des fonctions similaires[70].

Voici quelques-unes des méthodes bioinformatiques par homologie les plus couramment utilisées :

#### II.2.3.1. Alignement de séquences

Cette méthode consiste à aligner deux ou plusieurs séquences afin de déterminer les régions similaires et les variations entre elles. Les alignements de séquences peuvent être utilisés pour identifier des régions conservées, prédire des structures secondaires ou identifier des motifs fonctionnels[71].

#### II.2.3.2. Recherche de similarité de séquences

Il s'agit de comparer une séquence inconnue à une base de données de séquences connues pour trouver des séquences similaires. Les algorithmes couramment utilisés pour cela incluent BLAST (Basic Local Alignment Search Tool) et FASTA.

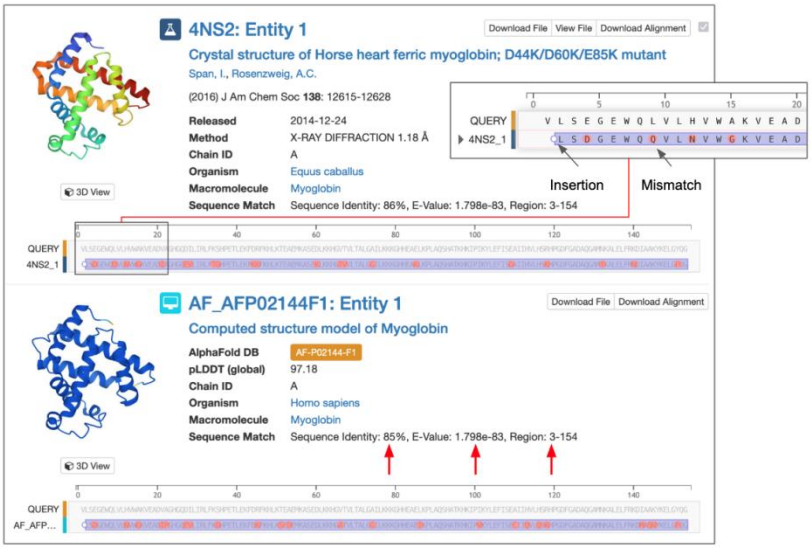
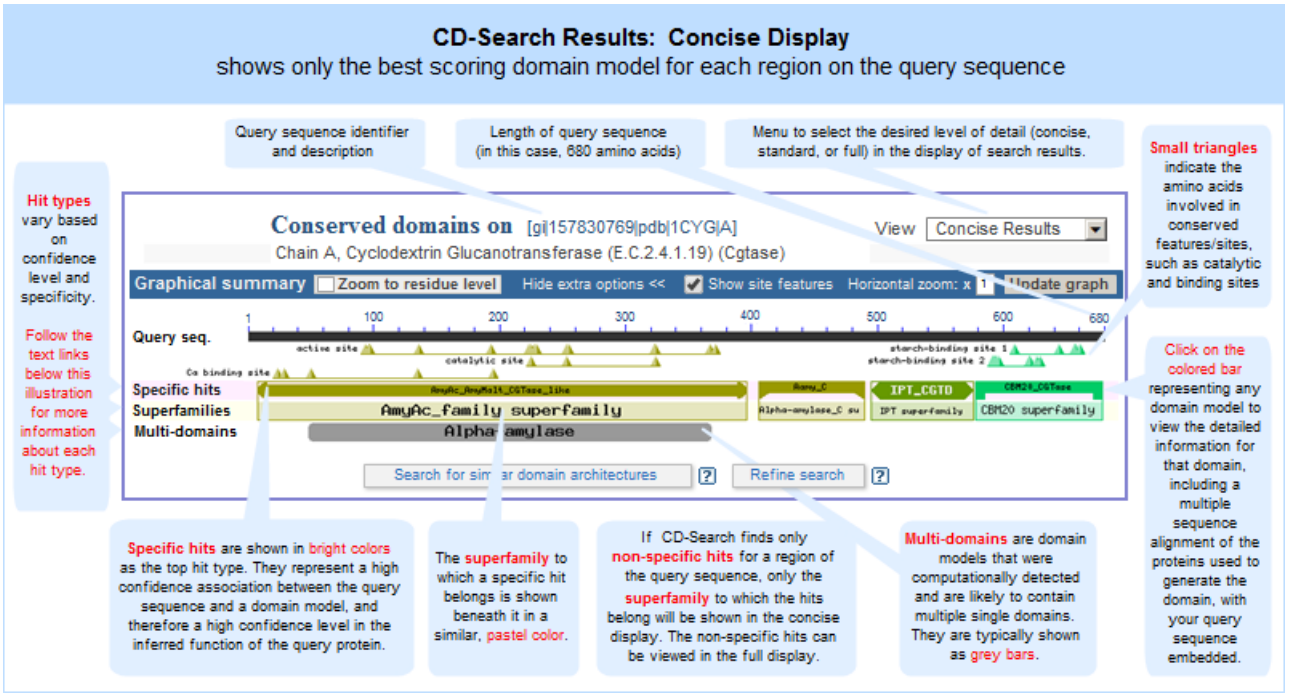



***Figure II.2: Partie d'une page de résultats de recherche de similarité de séquences montrant une structure expérimentale et un CSM correspondant à la requête. L'encart montre une vue agrandie de l'alignement des séquences comprenant une insertion et plusieurs discordances. Les mesures de similarité des séquences sont mises en évidence par des flèches rouges[72].***

### II.2.3.3. Recherche de domaines conservés

Les protéines sont souvent composées de domaines fonctionnels conservés. En identifiant des domaines conservés dans une séquence inconnue, on peut obtenir des informations sur sa fonction potentielle. Des bases de données telles que Pfam, SMART et CDD contiennent des annotations de domaines conservés et peuvent être utilisées pour cette recherche.




***L'exemple ci-dessus montre les résultats de recherche, datant du 22 octobre 2014, pour la protéine GI 157830769 (Cyclodextrin Glucanotransférase). Les types de correspondances dans l'affichage concis peuvent inclure des correspondances spécifiques, la super-famille à laquelle appartient la correspondance la mieux classée, ainsi que des modèles multi-domaines. Le document d'aide du CDD fournit plus d'informations sur : (1) les éléments d'affichage tels que les couleurs/formes utilisées pour les dessins des domaines, les petits triangles représentant des caractéristiques/sites conservés, les flèches à double tête représentant des motifs structuraux ; (2) les commandes d'affichage telles que le zoom horizontal et le zoom au niveau des résidus ; et (3) les options permettant de rechercher des architectures de domaines similaires et d'affiner la recherche[73].***

### II.2.3.4. Modélisation de la structure tridimensionnelle

Si une séquence inconnue présente une similarité significative avec une séquence dont la structure tridimensionnelle est connue, il est possible de prédire sa structure en se basant sur le modèle de la séquence connue[74].

### II.2.3.5. Phylogénie

La construction d'arbres phylogénétiques permet de déterminer les relations évolutives entre différentes séquences biologiques. Ces arbres peuvent être construits en se basant sur des similarités de séquences, des motifs conservés ou d'autres caractéristiques moléculaires.

Ces méthodes bioinformatiques par homologie sont des outils puissants pour étudier et annoter des séquences biologiques inconnues. Elles permettent d'obtenir des informations sur la fonction, l'évolution et les relations entre différentes séquences biologiques[75].

### II.2.4. Les méthodes bioinformatiques par caractéristique

Les méthodes bioinformatiques par caractéristique font référence à des approches analytiques utilisées pour extraire des informations des données biologiques en se basant sur des caractéristiques spécifiques des séquences ou des structures moléculaires. Ces caractéristiques peuvent inclure des motifs, des propriétés physico-chimiques, des annotations fonctionnelles, des structures secondaires, etc.[76]. Voici quelques exemples de méthodes bioinformatiques par caractéristique :

#### II.2.4.1. Recherche de motifs

Les motifs sont des séquences courtes et conservées qui sont associées à des fonctions biologiques spécifiques. Les méthodes de recherche de motifs, telles que les motifs consensus, les motifs de séquence conservée (comme les motifs de liaison à l'ADN) ou les motifs de motif linéaire, peuvent aider à identifier des régions importantes dans les séquences biologiques[77].

#### II.2.4.2. Prédiction de la structure secondaire

La structure secondaire des protéines est un élément clé pour comprendre leur fonction. Des méthodes bioinformatiques telles que PSIPRED, JPred et SOPMA peuvent être utilisées pour prédire la structure secondaire des protéines à partir de leur séquence d'acides aminés[78].

#### II.2.4.3. Prédiction de la localisation subcellulaire

Les protéines peuvent être localisées dans différentes parties de la cellule, ce qui est souvent lié à leur fonction. Des méthodes telles que TargetP, WoLF PSORT et CELLO permettent de prédire la localisation subcellulaire des protéines en se basant sur des caractéristiques telles que la composition en acides aminés ou la présence de signaux de localisation[79].

#### II.2.4.4. Annotation fonctionnelle

L'annotation fonctionnelle consiste à attribuer une fonction biologique à une séquence ou à une structure moléculaire. Des méthodes bioinformatiques par caractéristique, telles que l'analyse de domaines fonctionnels, la prédiction de sites actifs ou l'annotation de mots-clés, peuvent être utilisées pour prédire ou assigner des fonctions potentielles à des séquences inconnues[80].

### II.2.4.5. Analyse des propriétés physico-chimiques

Les propriétés physico-chimiques des acides aminés ou des nucléotides peuvent être utilisées pour prédire des caractéristiques importantes des séquences biologiques. Par exemple, la solubilité, l'hydrophobicité, la charge électrique, la flexibilité, etc., peuvent être analysées pour prédire certaines propriétés ou interactions.

Ces méthodes bioinformatiques par caractéristique permettent d'extraire des informations utiles des séquences biologiques en se basant sur des caractéristiques spécifiques, ce qui facilite l'annotation, la prédiction de structure et l'identification de fonctions potentielles des biomolécules[81].

## II.3. L'intelligence artificielle (IA)

L'intelligence artificielle (IA) est une branche de l'informatique qui vise à développer des algorithmes qui imitent le comportement intelligent des êtres humains. Les méthodes de l'IA, telles que le Machine Learning et le Deep Learning, sont de plus en plus utilisées dans la bioinformatique pour analyser les données biologiques.

L’intelligence artificielle transforme rapidement le domaine de la bioinformatique, permettant aux chercheurs de faire de nouvelles découvertes et de développer de nouvelles applications à un rythme sans précédent. Les outils alimentés par l'IA sont utilisés pour analyser d'énormes quantités de données biologiques, identifier des schémas et des tendances, et effectuer des prédictions qui seraient impossibles à réaliser avec des méthodes traditionnelles[82].

Certaines des applications les plus prometteuses de l'IA en bioinformatique comprennent :

- **Séquençage et analyse du génome**

Les outils alimentés par l'IA sont utilisés pour séquencer les génomes plus rapidement et plus précisément que jamais auparavant. Cela permet une meilleure compréhension de la génétique humaine et le développement de nouveaux traitements pour les maladies[83].

- **Prédiction de la structure des protéines**

Les outils alimentés par l'IA sont utilisés pour prédire la structure des protéines, ce qui est essentiel pour comprendre leur fonctionnement et leur interaction avec d'autres molécules. Ces informations sont utilisées pour concevoir de nouveaux médicaments et développer de nouvelles thérapies[84].

- **Découverte de médicaments**

Les outils alimentés par l'IA sont utilisés pour identifier de nouvelles cibles médicamenteuses et concevoir de nouveaux médicaments qui sont plus efficaces et moins toxiques que les médicaments existants. Cela conduit au développement de nouveaux traitements pour un large éventail de maladies, y compris le cancer, la maladie d'Alzheimer et la maladie de Parkinson[85].

- **II.3.4. Diagnostic de maladies**

Les outils alimentés par l'IA sont utilisés pour développer de nouveaux outils de diagnostic capables d'identifier les maladies plus tôt et avec une plus grande précision que les méthodes traditionnelles. Cela permet un traitement précoce et de meilleurs résultats pour les patients[86].

- **II.3.5. Médecine personnalisée**

Les outils alimentés par l'IA sont utilisés pour développer de nouvelles approches de médecine personnalisée, qui adaptent les traitements au profil génétique individuel du patient. Cela conduit à des traitements plus efficaces et moins toxiques pour un large éventail de maladies[87].

L'utilisation de l'IA en bioinformatique en est encore à ses débuts, mais elle a le potentiel de révolutionner notre compréhension et le traitement des maladies. Les outils alimentés par l'IA ont déjà un impact significatif dans le domaine, et leur potentiel d'amélioration de la santé humaine est immense.

Voici quelques exemples spécifiques de l'utilisation de l'IA en bioinformatique :

En 2016, une équipe de chercheurs a utilisé l'IA pour identifier un nouveau gène lié à une forme rare d'autisme. Le gène était auparavant inconnu, et sa découverte pourrait conduire à de nouveaux traitements pour l'autisme.

En 2017, une équipe de chercheurs a utilisé l'IA pour développer un nouveau médicament efficace contre une forme de leucémie. Le médicament a été développé en seulement quelques années, ce qui est beaucoup plus rapide que les méthodes traditionnelles de développement de médicaments.

En 2018, une équipe de chercheurs a utilisé l'IA pour développer un nouvel outil de diagnostic capable d'identifier la maladie d'Alzheimer des années avant l'apparition des symptômes. Cet outil pourrait permettre un traitement précoce et de meilleurs résultats pour les patients.

Ce ne sont que quelques exemples parmi tant d'autres de la façon dont l'IA est utilisés en bioinformatique. À mesure que la technologie de l'IA continue à se développer, nous pouvons nous attendre à voir encore plus d'applications innovantes et révolutionnaires dans les années à venir.

### II.3.1. L'apprentissage automatique (Machine learning)

Est un domaine en pleine croissance qui a un impact majeur sur de nombreuses industries différentes, y compris la bioinformatique. Les algorithmes d'apprentissage automatique peuvent être utilisés pour analyser de grandes quantités de données biologiques afin d'identifier des schémas et des tendances qui seraient impossibles à voir à l'œil nu. Ces informations peuvent ensuite être utilisées pour faire des prédictions sur les maladies, développer de nouveaux médicaments et améliorer le diagnostic et le traitement des patients[88].

Certains des algorithmes d'apprentissage automatique les plus couramment utilisés en bioinformatique comprennent :

- **Machines à vecteurs de support (SVM)**

Les machines à vecteurs de support sont utilisées pour des tâches de classification et de régression. Elles fonctionnent en trouvant l'hyperplan qui sépare le mieux deux classes de données[89].

- **arbres de décision**

Les arbres de décision sont utilisés pour des tâches de classification et de régression. Ils fonctionnent en divisant les données en sous-ensembles de plus en plus petits jusqu'à ce que chaque sous-ensemble puisse être classifié ou prédit[90].

- **Réseaux neuronaux**

Les réseaux neuronaux sont un type d'algorithme d'apprentissage automatique inspiré du cerveau humain. Ils fonctionnent en apprenant à reconnaître des schémas dans les données en connectant des nœuds les uns aux autres dans une structure en réseau[91].

### II.3.1.1. Applications de l'apprentissage automatique en bioinformatique

L'apprentissage automatique est utilisé en bioinformatique pour un large éventail d'applications, notamment :

- **Séquençage et analyse du génome**

Les algorithmes d'apprentissage automatique sont utilisés pour séquencer les génomes plus rapidement et plus précisément que jamais auparavant. Cela permet une meilleure compréhension de la génétique humaine et le développement de nouveaux traitements pour les maladies.

- **Prédiction de la structure des protéines**

Les algorithmes d'apprentissage automatique sont utilisés pour prédire la structure des protéines, ce qui est essentiel pour comprendre leur fonctionnement et leur interaction avec d'autres molécules. Ces informations sont utilisées pour concevoir de nouveaux médicaments et développer de nouvelles thérapies.

- **Découverte de médicaments**

Les algorithmes d'apprentissage automatique sont utilisés pour identifier de nouvelles cibles médicamenteuses et concevoir de nouveaux médicaments qui sont plus efficaces et moins toxiques que les médicaments existants. Cela conduit au développement de nouveaux traitements pour un large éventail de maladies, y compris le cancer, la maladie d'Alzheimer et la maladie de Parkinson.

- **Diagnostic de maladies**

Les algorithmes d'apprentissage automatique sont utilisés pour développer de nouveaux outils de diagnostic capables d'identifier les maladies plus tôt et avec une plus grande précision que les méthodes traditionnelles. Cela permet un traitement précoce et de meilleurs résultats pour les patients.

- **Médecine personnalisée**

Les algorithmes d'apprentissage automatique sont utilisés pour développer de nouvelles approches de médecine personnalisée, qui adaptent les traitements au profil génétique individuel du patient. Cela conduit à des traitements plus efficaces et moins toxiques pour un large éventail de maladies.

L'utilisation de l'apprentissage automatique en bioinformatique en est encore à ses débuts, mais elle a le potentiel de révolutionner notre compréhension et le traitement des maladies. Les algorithmes d'apprentissage automatique ont déjà un impact significatif dans le domaine, et leur potentiel d'amélioration de la santé humaine est immense.

Voici quelques exemples spécifiques de l'utilisation de l'apprentissage automatique en bioinformatique:

En 2016, une équipe de chercheurs a utilisé l'apprentissage automatique pour identifier un nouveau gène lié à une forme rare d'autisme. Le gène était auparavant inconnu, et sa découverte pourrait conduire à de nouveaux traitements pour l'autisme.

En 2017, une équipe de chercheurs a utilisé l'apprentissage automatique pour développer un nouveau médicament efficace contre une forme de leucémie. Le médicament a été développé en seulement quelques années, ce qui est beaucoup plus rapide que les méthodes traditionnelles de développement de médicaments.

En 2018, une équipe de chercheurs a utilisé l'apprentissage automatique pour développer un nouvel outil de diagnostic capable d'identifier la maladie d'Alzheimer des années avant l'apparition des symptômes. Cet outil pourrait permettre un traitement précoce et de meilleurs résultats pour les patients.

Ce ne sont là que quelques exemples parmi tant d'autres des nombreuses façons dont l'apprentissage automatique est utilisé en bioinformatique. À mesure que la technologie de l'apprentissage automatique continue à se développer, nous pouvons nous attendre à voir encore plus d'applications innovantes et révolutionnaires dans les années à venir.

### II.3.2. L'apprentissage profond

Le deep learning est un type d'apprentissage automatique qui utilise des réseaux neuronaux artificiels pour apprendre à partir de données. Les réseaux neuronaux sont inspirés du cerveau

humain et sont capables d'apprendre des schémas complexes dans les données qui seraient difficiles ou impossibles à identifier avec les algorithmes d'apprentissage automatique traditionnels[88, 92].

L'apprentissage profond est utilisé en bioinformatique pour un large éventail d'applications, notamment:

- **Séquençage et analyse du génome**

Les algorithmes d'apprentissage profond sont utilisés pour séquencer les génomes plus rapidement et plus précisément que jamais auparavant. Cela permet une meilleure compréhension de la génétique humaine et le développement de nouveaux traitements pour les maladies.

- **Prédiction de la structure des protéines**

Les algorithmes d'apprentissage profond sont utilisés pour prédire la structure des protéines, ce qui est essentiel pour comprendre leur fonctionnement et leur interaction avec d'autres molécules. Ces informations sont utilisées pour concevoir de nouveaux médicaments et développer de nouvelles thérapies.

- **Découverte de médicaments**

Les algorithmes d'apprentissage profond sont utilisés pour identifier de nouvelles cibles médicamenteuses et concevoir de nouveaux médicaments qui sont plus efficaces et moins toxiques que les médicaments existants. Cela conduit au développement de nouveaux traitements pour un large éventail de maladies, y compris le cancer, la maladie d'Alzheimer et la maladie de Parkinson.

- **Diagnostic de maladies**

Les algorithmes d'apprentissage profond sont utilisés pour développer de nouveaux outils de diagnostic capables d'identifier les maladies plus tôt et avec une plus grande précision que les méthodes traditionnelles. Cela permet un traitement précoce et de meilleurs résultats pour les patients.

- **Médecine personnalisée**

Les algorithmes d'apprentissage profond sont utilisés pour développer de nouvelles approches de médecine personnalisée, qui adaptent les traitements au profil génétique individuel du patient. Cela conduit à des traitements plus efficaces et moins toxiques pour un large éventail de maladies.

Voici quelques exemples spécifiques de l'utilisation de l'apprentissage profond en bioinformatique :

En 2016, une équipe de chercheurs a utilisé l'apprentissage profond pour identifier un nouveau gène lié à une forme rare d'autisme. Le gène était auparavant inconnu, et sa découverte pourrait conduire à de nouveaux traitements pour l'autisme.

En 2017, une équipe de chercheurs a utilisé l'apprentissage profond pour développer un nouveau médicament efficace contre une forme de leucémie. Le médicament a été développé en seulement quelques années, ce qui est beaucoup plus rapide que les méthodes traditionnelles de développement de médicaments.

En 2018, une équipe de chercheurs a utilisé l'apprentissage profond pour développer un nouvel outil de diagnostic capable d'identifier la maladie d'Alzheimer des années avant l'apparition des symptômes. Cet outil pourrait permettre un traitement précoce et de meilleurs résultats pour les patients.

Ce ne sont là que quelques exemples parmi tant d'autres des nombreuses façons dont l'apprentissage profond sont utilisés en bioinformatique. À mesure que la technologie de l'apprentissage profond continue à se développer, nous pouvons nous attendre à voir encore plus d'applications innovantes et révolutionnaires dans les années à venir.

Voici certains des avantages de l'utilisation de l'apprentissage profond en bioinformatique :

Les algorithmes d'apprentissage profond peuvent apprendre des schémas complexes dans les données qui seraient difficiles ou impossibles à identifier avec les algorithmes d'apprentissage automatique traditionnels.

Les algorithmes d'apprentissage profond peuvent être entraînés sur de très grands ensembles de données, ce qui peut conduire à des résultats plus précis.

Les algorithmes d'apprentissage profond sont souvent capables d'apprendre à partir de données bruitées ou incomplètes, ce qui peut poser problème avec les algorithmes d'apprentissage automatique traditionnels.

Voici certains des défis liés à l'utilisation de l'apprentissage profond en bioinformatique :

Les algorithmes d'apprentissage profond peuvent être coûteux en termes de calculs pour l'entraînement.

Les algorithmes d'apprentissage profond peuvent être difficiles à interpréter, ce qui peut rendre difficile la compréhension de la façon dont ils effectuent leurs prédictions.

Les algorithmes d'apprentissage profond peuvent être biaisés, ce qui peut conduire à des résultats inexacts.

Dans l'ensemble, l'apprentissage profond est un outil puissant qui a un impact majeur sur le domaine de la bioinformatique. Les algorithmes d'apprentissage profond sont utilisés pour faire de nouvelles découvertes, développer de nouvelles applications et améliorer le diagnostic et le traitement des maladies. À mesure que la technologie de l'apprentissage profond continue à se développer, nous pouvons nous attendre à voir encore plus d'applications innovantes et révolutionnaires dans les années à venir.

### II.4. Travaux connexes

- **Titre : Classification du cancer basée sur l'expression génique**

Résumé :

Le document se concentre sur la classification du cancer basée sur l'expression génique, une approche puissante pour catégoriser et stratifier les types de cancer en analysant les profils d'expression génique. L'étude met en évidence l'importance des schémas d'expression génique dans la compréhension de la biologie du cancer, la prédiction des résultats chez les patients et l'orientation des stratégies de traitement personnalisé [93].

Les auteurs commencent par discuter de l'avènement des technologies de profilage de l'expression génique à haut débit, telles que les puces à ADN et le séquençage de l'ARN, qui ont révolutionné la recherche sur le cancer. Ces techniques permettent de mesurer simultanément les niveaux d'expression de milliers de gènes dans un échantillon tissulaire,

fournissant ainsi des informations précieuses sur les mécanismes moléculaires sous-jacents au développement et à la progression du cancer.

Ensuite, le document explore différentes méthodes de calcul et algorithmes statistiques utilisés pour la classification du cancer basée sur les données d'expression génique. Ces méthodes impliquent généralement le prétraitement des données, la sélection des caractéristiques et l'application de modèles de classification. Les auteurs discutent des approches populaires, telles que le regroupement hiérarchique, le regroupement en k-means, les machines à vecteurs de support (SVM), les forêts aléatoires et les réseaux neuronaux.

L'importance des techniques de sélection des caractéristiques est soulignée, car les données d'expression génique contiennent souvent du bruit et des informations redondantes. Les méthodes de sélection des caractéristiques visent à identifier les gènes les plus informatifs qui contribuent à la tâche de classification, améliorant ainsi la précision et l'interprétabilité des modèles.

Le document met également en évidence l'importance de la classification du cancer basée sur l'expression génique dans la pratique clinique. Il aborde les applications potentielles de cette approche, telles que la sous-typage des tumeurs, la prédiction du pronostic et la sélection du traitement. Les auteurs mentionnent plusieurs études où le profilage de l'expression génique a été utilisé avec succès pour classer différents types de cancer, identifier des sous-types moléculaires avec des résultats cliniques distincts et prédire la réponse au traitement.

De plus, les défis et les limitations associés à la classification du cancer basée sur l'expression génique sont discutés. Ils comprennent des problèmes de qualité des données, le surajustement, la robustesse des modèles et la nécessité d'une validation dans des ensembles de données indépendants. Les auteurs soulignent l'importance des efforts collaboratifs, des protocoles standardisés et de l'intégration des données multi-omiques pour de nouveaux progrès dans ce domaine.

En conclusion, la classification du cancer basée sur l'expression génique offre un outil puissant pour comprendre le paysage complexe de la biologie du cancer et améliorer la prise de décision clinique. Elle permet d'identifier des sous-types de cancer distincts, de prédire les résultats chez les patients et de développer des stratégies de traitement personnalisé. Des recherches continues et des avancées technologiques sont nécessaires pour surmonter les défis

et réaliser pleinement le potentiel de la classification du cancer basée sur l'expression génique en oncologie de précision[93].

- **Titre : Sélection de fonctionnalités basée sur la corrélation améliorée par l'optimisation de l'essaim de particules binaire pour la sélection de gènes et la classification du cancer**

Résumé :

Le document présente une nouvelle approche appelée sélection de fonctionnalités basée sur la corrélation améliorée par l'optimisation de l'essaim de particules binaire (CFS-IBPSO) pour la sélection de gènes et la classification du cancer. L'étude vise à relever les défis des données d'expression génique de grande dimension en combinant des techniques de sélection de fonctionnalités avec un algorithme d'optimisation de l'essaim de particules binaire, améliorant ainsi la précision et l'efficacité des modèles de classification du cancer [94].

Les auteurs commencent par souligner l'importance de la sélection de gènes dans les tâches de classification du cancer. Avec la disponibilité de vastes ensembles de données d'expression génique, la sélection des gènes pertinents devient cruciale pour réduire la dimensionnalité, éliminer le bruit et améliorer les performances des modèles de classification. Les méthodes traditionnelles telles que les tests t et le gain d'information sont limitées dans le traitement des données de grande dimension, ce qui justifie le besoin de techniques plus avancées.

L'approche CFS-IBPSO proposée dans le document combine l'algorithme de sélection de fonctionnalités basé sur la corrélation (CFS) avec un algorithme d'optimisation de l'essaim de particules binaire amélioré (IBPSO). CFS évalue la pertinence et la redondance des gènes en fonction de leur corrélation avec les étiquettes de classe, tandis qu'IBPSO est utilisé pour rechercher un sous-ensemble optimal de gènes maximisant la précision de la classification.

Les auteurs décrivent les étapes de l'algorithme CFS-IBPSO, qui consistent à initialiser une population de particules binaires représentant des sous-ensembles de gènes, à évaluer leur aptitude à l'aide d'un modèle de classification, à mettre à jour les positions des particules en fonction de leurs meilleures positions personnelles et globales, et à affiner itérativement le sous-ensemble de gènes pour converger vers une solution optimale.

Des évaluations expérimentales utilisant des ensembles de données d'expression génique disponibles publiquement pour la classification du cancer sont menées pour évaluer les

performances de l'approche CFS-IBPSO. Les résultats démontrent que la méthode proposée surpasse les autres techniques de sélection de fonctionnalités, telles que CFS et l'algorithme génétique, en termes de précision de classification, de sensibilité, de spécificité et de surface sous la courbe de caractéristique opérationnelle du récepteur.

Le document conclut en mettant en évidence l'importance de la sélection de gènes dans la classification du cancer et le potentiel de l'approche CFS-IBPSO pour améliorer la précision et l'efficacité du diagnostic du cancer basé sur l'expression génique. Les auteurs suggèrent que la méthode proposée peut aider à identifier les gènes informatifs associés aux types de cancer, contribuant ainsi à une meilleure compréhension de la biologie du cancer et facilitant les stratégies de traitement personnalisé.

En résumé, le document présente l'approche CFS-IBPSO, qui combine la sélection de gènes basée sur la corrélation avec un algorithme d'optimisation de l'essaim de particules binaire amélioré pour la classification du cancer. Les résultats expérimentaux indiquent que cette méthode offre des performances supérieures par rapport aux techniques existantes, mettant en évidence son potentiel pour une sélection de gènes efficace et un diagnostic précis du cancer[94].

## II.5. Conclusion

La bioinformatique, l'apprentissage profond et l'apprentissage automatique jouent un rôle crucial dans la prédiction et la classification du cancer. Ces domaines permettent l'analyse des données biologiques complexes pour identifier des modèles et des tendances qui seraient difficiles à détecter manuellement. Voici les points clés :

1. Bioinformatique : La bioinformatique utilise des méthodes et des outils informatiques pour gérer, analyser et interpréter les données biologiques. Dans le contexte du cancer, elle permet l'analyse des séquences génétiques, des données d'expression génique, des données protéomiques et d'autres types de données biologiques.

2. Apprentissage automatique : L'apprentissage automatique (machine learning) est une branche de l'intelligence artificielle qui permet aux ordinateurs d'apprendre à partir de données sans être explicitement programmés. Il permet de créer des modèles statistiques et des algorithmes capables de prédire des résultats ou de classer des données.

3. Apprentissage profond : L'apprentissage profond (deep learning) est une technique d'apprentissage automatique qui utilise des réseaux de neurones artificiels inspirés du fonctionnement du cerveau humain. Il est capable d'apprendre des représentations complexes à partir des données et de détecter des motifs profonds.

Dans le contexte de la prédiction et de la classification du cancer, voici comment ces domaines sont utilisés :

- Prédiction du cancer : Les algorithmes d'apprentissage automatique et d'apprentissage profond sont utilisés pour prédire le risque de développer un cancer chez un individu, en utilisant des facteurs de risque génétiques, des données cliniques et des informations sur le mode de vie.

- Classification du cancer : Les algorithmes d'apprentissage automatique et d'apprentissage profond sont utilisés pour classer différents types de cancer et pour différencier les sous-types de tumeurs. Cela permet d'identifier des caractéristiques spécifiques et de mieux comprendre la biologie des différents cancers.

- Diagnostic précoce : L'apprentissage automatique et l'apprentissage profond sont utilisés pour développer des outils de diagnostic précoce du cancer en analysant des données d'imagerie médicale, des biomarqueurs et d'autres informations pertinentes. Cela permet une détection précoce du cancer et une intervention thérapeutique précoce.

- Prédiction de la réponse au traitement : Les modèles d'apprentissage automatique et d'apprentissage profond peuvent être utilisés pour prédire la réponse d'un patient à un traitement spécifique. Cela peut aider à personnaliser les traitements et à améliorer les résultats pour les patients atteints de cancer.

En résumé, la bioinformatique, l'apprentissage profond et l'apprentissage automatique sont des domaines essentiels pour la prédiction et la classification du cancer. Ces approches permettent une analyse approfondie des données biologiques, une meilleure compréhension de la biologie du cancer et une personnalisation des traitements pour améliorer les résultats pour les patients atteints de cancer.

# Chapitre III:
# **Contribution**

## III.1. Introduction

Ce chapitre présente notre contribution à la classification des maladies cancéreuses à partir de l'expression génique en utilisant des techniques de machine learning et de deep learning. La classification précise des types de cancer est d'une importance capitale dans le domaine de la recherche en oncologie, car elle permet une meilleure compréhension des mécanismes sous-jacents aux maladies cancéreuses et facilite la prise de décisions cliniques éclairées. Les récents progrès technologiques ont permis la génération de vastes ensembles de données d'expression génique, créant ainsi une opportunité sans précédent d'exploiter ces données pour améliorer les méthodes de classification des cancers. Cependant, le défi majeur réside dans la complexité et la dimensionnalité élevée de ces données. Dans ce contexte, notre contribution repose sur une approche en trois étapes visant à traiter les données, à effectuer une classification précise et à extraire des biomarqueurs informatifs. La première étape consiste à mettre en œuvre des techniques avancées de prétraitement des données afin de réduire le bruit, de normaliser les données d'expression génique et de sélectionner les caractéristiques les plus pertinentes. La deuxième étape repose sur l'utilisation de modèles de machine learning et de deep learning pour effectuer la classification des différents types de cancer. Enfin, la troisième étape vise à extraire des biomarqueurs discriminants à partir des modèles entraînés, ce qui permet de découvrir des signatures moléculaires spécifiques à chaque type de cancer. Cette contribution vise à améliorer la précision diagnostique et à faciliter le développement de traitements personnalisés en identifiant des biomarqueurs potentiels pour les patients atteints de cancer.

## III.2. Méthode Proposée

La méthode proposée dans cette étude repose sur une approche en trois étapes pour la classification des maladies cancéreuses à partir de l'expression génique, avec pour objectif principal d'améliorer la précision diagnostique et d'identifier des biomarqueurs informatifs. La première étape consiste à traiter les données d'expression génique en utilisant des techniques avancées de prétraitement. Cela comprend la normalisation des données pour éliminer les variations non pertinentes, la réduction du bruit pour améliorer la qualité des données, et la sélection des caractéristiques les plus pertinentes pour réduire la dimensionnalité des données et améliorer l'efficacité des algorithmes de classification.

Dans la deuxième étape, nous utilisons des modèles de machine learning et de deep learning pour effectuer la classification des différents types de cancer. Des algorithmes tels que les machines à vecteurs de support (SVM), les réseaux de neurones convolutifs (CNN) et les réseaux de neurones récurrents (RNN) sont explorés et comparés pour leur capacité à extraire des motifs et des caractéristiques discriminantes à partir des données d'expression génique. Des architectures de réseaux neuronaux spécifiquement conçues pour l'apprentissage automatique sur des données d'expression génique sont également développées et évaluées.

Enfin, dans la troisième étape, nous nous concentrons sur l'extraction de biomarqueurs informatifs à partir des modèles entraînés. Ces biomarqueurs sont des signatures moléculaires spécifiques à chaque type de cancer, qui peuvent fournir des informations précieuses sur les mécanismes sous-jacents à la maladie et faciliter le développement de traitements personnalisés. Différentes techniques d'interprétabilité sont utilisées pour analyser les poids des caractéristiques et les activations des neurones, afin d'identifier les gènes et les voies biologiques qui jouent un rôle crucial dans la classification des cancers.

La contribution de cette étude réside dans l'optimisation et l'intégration de ces trois étapes dans une méthodologie cohérente et complète. Cette approche permet d'obtenir des résultats de classification plus précis et de découvrir des biomarqueurs potentiellement pertinents, ouvrant ainsi de nouvelles perspectives pour la compréhension et le traitement des maladies cancéreuses. La méthodologie proposée peut également servir de base pour d'autres études de classification basées sur l'expression génique et constitue une contribution significative à l'amélioration de la pratique clinique et de la recherche en oncologie.

### III.2.1. Traitement de données

L'étape de traitement des données est une étape essentielle dans notre étude de classification des maladies cancéreuses à partir de l'expression génique. Cette étape vise à préparer et à transformer les données brutes d'expression génique pour les rendre adaptées à l'analyse et à l'apprentissage automatique.

#### Prétraitement des données

Dans cette étape, on effectue des opérations de prétraitement pour nettoyer les données brutes et éliminer les éventuelles erreurs ou valeurs aberrantes. Cela peut inclure des vérifications de qualité des données pour détecter les erreurs de mesure, les valeurs manquantes ou les valeurs extrêmes qui pourraient fausser les résultats. On peut également

effectuer des opérations de filtrage pour éliminer les gènes non informatifs ou les échantillons de faible qualité.

**Normalisation des données**

Comme mentionné précédemment, la normalisation des données est une étape importante pour éliminer les biais techniques et rendre les données comparables entre les échantillons. On applique des méthodes de normalisation appropriées pour ajuster les valeurs d'expression génique. Cela peut inclure des approches telles que la normalisation globale, la normalisation par référence interne ou la normalisation par facteur de taille de bibliothèque (Library Size). L'objectif est d'obtenir des données normalisées qui sont représentatives de l'activité des gènes dans chaque échantillon.

**Sélection des caractéristiques**

Lorsque nous travaillons avec des ensembles de données d'expression génique de grande dimension, il peut être nécessaire de réduire le nombre de caractéristiques pour éviter la surcharge et améliorer les performances de classification. Nous pouvons utiliser des méthodes de sélection de caractéristiques pour identifier les gènes les plus discriminants ou les plus pertinents pour la classification des maladies cancéreuses. Des approches telles que l'analyse de variance (ANOVA), les tests statistiques ou les méthodes basées sur l'information mutuelle peuvent être utilisées pour sélectionner les caractéristiques les plus informatives.

**Séparation des ensembles de données**

Pour évaluer les performances du modèle et éviter le surajustement, il est courant de diviser l'ensemble de données en ensembles d'entraînement, de validation et de test. L'ensemble d'entraînement est utilisé pour ajuster les paramètres du modèle, l'ensemble de validation est utilisé pour régler les hyperparamètres et l'ensemble de test est utilisé pour évaluer les performances finales du modèle sur des données non vues. Cette séparation des données permet d'estimer la capacité du modèle à généraliser à de nouveaux échantillons et à évaluer sa robustesse.

En résumé, l'étape de traitement des données dans notre étude de classification des maladies cancéreuses à partir de l'expression génique comprend le prétraitement des données, la normalisation des valeurs d'expression génique, la sélection des caractéristiques pertinentes

et la séparation des ensembles de données. Ces étapes sont cruciales pour préparer les données de manière adéquate avant de les utiliser dans les modèles de machine learning et de deep learning.

### III.2.2. Modèle de classification CNN

L'étape de classification des différents types de cancer repose sur l'utilisation de plusieurs algorithmes de machine learning, notamment les réseaux de neurones convolutifs (CNN), les machines à vecteurs de support (SVM) et les perceptrons multicouches (MLP). Chacun de ces algorithmes offre des capacités uniques pour extraire des motifs et des caractéristiques discriminantes à partir des données d'expression génique.

Les réseaux de neurones convolutifs (CNN) sont une classe d'algorithmes de deep learning spécialement conçus pour l'analyse des données structurées, telles que les images et les séquences. Les CNN sont composés de multiples couches, notamment des couches de convolution, de pooling et de neurones entièrement connectés. Les couches de convolution appliquent des filtres pour extraire des caractéristiques locales des données d'entrée. Dans notre approche, nous utilisons un modèle CNN spécifiquement adapté à l'analyse des données d'expression génique en utilisant une convolution 1D. Ce modèle utilise 32 filtres avec un noyau de taille 5 pour capturer les motifs et les interactions entre les gènes dans un contexte séquentiel.

Voici une description détaillée de chaque étape dans l'utilisation des réseaux de neurones convolutifs (CNN) pour la classification des maladies cancéreuses à partir de l'expression génique :

#### Convolution

La première étape du modèle CNN consiste à appliquer la convolution sur les données d'expression génique. La convolution permet de capturer les motifs et les interactions entre les gènes sur une fenêtre donnée. Dans notre cas, nous utilisons une convolution 1D pour prendre en compte la structure séquentielle des données d'expression génique. Chaque filtre de la couche de convolution est déplacé sur les données avec un pas (stride) de 1, en calculant le produit scalaire entre le filtre et la fenêtre correspondante des données d'expression génique. Cette opération produit une carte d'activation qui met en évidence les motifs importants présents dans les données.

**Sous-échantillonnage (Max Pooling)**

Après l'étape de convolution, nous utilisons une opération de sous-échantillonnage pour réduire la dimension des cartes d'activation et extraire les caractéristiques les plus importantes. L'opération de sous-échantillonnage la plus couramment utilisée est le sous-échantillonnage par max pooling. Cette opération consiste à diviser la carte d'activation en régions non chevauchantes et à prendre la valeur maximale de chaque région. Cela permet de conserver les caractéristiques les plus saillantes tout en réduisant la taille des données.

**Couche dense avec classification softmax**

Après plusieurs couches de convolution et de sous-échantillonnage, les caractéristiques extraites sont aplaties en un vecteur unidimensionnel. Ce vecteur est ensuite alimenté dans une couche dense (fully connected) qui est responsable de l'apprentissage des relations entre les caractéristiques extraites et les classes de cancer. Cette couche dense est généralement suivie d'une fonction d'activation, telle que la fonction ReLU (Rectified Linear Unit), qui introduit des non-linéarités dans le modèle. Enfin, la couche de sortie utilise une fonction d'activation softmax pour produire des probabilités de classe pour chaque type de cancer. La fonction softmax garantit que la somme des probabilités de toutes les classes est égale à 1.

**Fonction de perte (Loss Function) - Entropie croisée catégorique**

La fonction de perte utilisée dans notre modèle CNN est l'entropie croisée catégorique (categorical crossentropy). Cette fonction mesure la divergence entre les probabilités prédites par le modèle et les étiquettes réelles des échantillons d'expression génique. L'objectif est de minimiser cette divergence pour améliorer la précision de la classification. L'entropie croisée catégorique est couramment utilisée pour les tâches de classification multi-classes, où chaque échantillon appartient à une classe unique parmi plusieurs classes possibles.

En utilisant ces différentes étapes, notre modèle CNN est capable d'apprendre à extraire des caractéristiques discriminantes à partir des données d'expression génique et de les utiliser pour effectuer une classification précise des différents types de cancer.

### III.2.3. Modèle de classification avec ML

En plus du CNN, nous avons également utilisé les machines à vecteurs de support (SVM) et les perceptrons multicouches (MLP) pour la classification des maladies cancéreuses à partir de l'expression génique, afin de comparer les différentes techniques de classification

et trouver la meilleure méthode. Les SVM sont des algorithmes d'apprentissage supervisé qui trouvent un hyperplan optimal pour séparer les échantillons de différentes classes dans un espace de caractéristiques. Les SVM peuvent être utilisés pour résoudre des problèmes de classification linéaires et non linéaires en utilisant des fonctions de noyau appropriées. Les MLP, quant à eux, sont des réseaux de neurones artificiels avec plusieurs couches cachées. Chaque couche est composée de neurones qui utilisent des fonctions d'activation pour introduire des non-linéarités dans le modèle. Les MLP sont largement utilisés pour des tâches de classification complexes en raison de leur capacité à apprendre des représentations hiérarchiques des données. Voici une description détaillée de chaque méthode de machine learning utilisée dans notre étude, à savoir les machines à vecteurs de support (SVM) et les perceptrons multicouches (MLP) :

**Perceptrons multicouches (MLP)**

Les perceptrons multicouches, également connus sous le nom de réseaux de neurones artificiels à couches multiples, sont des modèles de machine learning composés de plusieurs couches de neurones interconnectés. Chaque neurone reçoit des entrées pondérées, applique une fonction d'activation non linéaire et transmet le résultat à la couche suivante. Les MLP sont capables d'apprendre des représentations hiérarchiques complexes des données et sont particulièrement adaptés aux problèmes de classification non linéaires.

Dans le contexte de la classification des maladies cancéreuses à partir de l'expression génique, les MLP peuvent être utilisés pour apprendre les relations non linéaires entre les caractéristiques génétiques et les classes de cancer. Chaque couche du MLP est composée de plusieurs neurones, et chaque neurone est connecté à tous les neurones de la couche précédente. Les poids des connexions entre les neurones sont ajustés pendant la phase d'apprentissage à l'aide d'algorithmes d'optimisation tels que la rétropropagation du gradient. Les MLP peuvent apprendre à discriminer les différentes classes de cancer en ajustant les poids de manière itérative pour minimiser une fonction de perte, comme l'erreur quadratique moyenne ou l'entropie croisée.

Les MLP offrent une grande flexibilité en termes de conception du modèle, avec des options pour le nombre de couches cachées, le nombre de neurones par couche et les fonctions d'activation. Cependant, les MLP peuvent être sensibles à la surajustement

(overfitting) si le modèle est trop complexe par rapport à la taille de l'ensemble de données. Une régularisation appropriée et une validation croisée peuvent être utilisées pour atténuer ce problème et améliorer les performances de classification.

**Machines à vecteurs de support (SVM)**

Les machines à vecteurs de support sont des algorithmes d'apprentissage supervisé utilisés pour résoudre des problèmes de classification et de régression. L'objectif des SVM est de trouver un hyperplan optimal qui sépare les échantillons de différentes classes dans un espace de caractéristiques. Les SVM sont particulièrement efficaces lorsque les données sont linéairement séparables, mais ils peuvent également être étendus pour traiter des problèmes non linéaires en utilisant des fonctions de noyau.

Dans le cas de la classification des maladies cancéreuses à partir de l'expression génique, les SVM peuvent être utilisés pour trouver un hyperplan qui maximise la marge entre les échantillons de différentes classes. Cela permet de créer une frontière de décision claire entre les classes, ce qui facilite la classification des nouveaux échantillons. Les SVM sont également résistants aux valeurs aberrantes et peuvent gérer efficacement les données de haute dimension, ce qui en fait un choix populaire pour l'analyse des données d'expression génique.

### III.2.4. Extraire les biomarqueurs

L'extraction de caractéristiques, également connue sous le nom de sélection de biomarqueurs, est une étape importante dans notre étude de classification des maladies cancéreuses à partir de l'expression génique. Cette étape vise à identifier les caractéristiques ou gènes les plus pertinents et informatifs pour la classification des échantillons.

**Importance des caractéristiques**

Une fois que nous avons entraîné les modèles de classification tels que les CNN, SVM et MLP, nous pouvons exploiter les informations apprises par ces modèles pour évaluer l'importance des caractéristiques. Chaque modèle a ses propres mécanismes internes pour attribuer des poids ou des coefficients aux différentes caractéristiques en fonction de leur contribution à la classification des échantillons.

**Poids ou coefficients des caractéristiques**

Dans le cas des CNN, nous pouvons extraire les poids associés aux différentes couches de convolution et de pooling. Les poids élevés indiquent une forte contribution de la caractéristique correspondante à la classification des échantillons, tandis que les poids faibles indiquent une contribution moindre. Nous pouvons également utiliser des techniques de visualisation, telles que les cartes de chaleur, pour représenter visuellement les caractéristiques les plus importantes.

Pour les modèles tels que les SVM et les MLP, nous pouvons extraire les coefficients associés aux différentes caractéristiques. Les coefficients plus élevés indiquent une contribution plus importante à la classification, tandis que les coefficients proches de zéro indiquent une contribution plus faible.

**Sélection des biomarqueurs**

Une fois que nous avons obtenu les poids, coefficients ou mesures d'importance des caractéristiques à partir des modèles de classification, nous pouvons effectuer une sélection des biomarqueurs. Cette sélection consiste à choisir les caractéristiques les plus informatives et discriminantes pour la classification des maladies cancéreuses.

Différentes approches peuvent être utilisées pour sélectionner les biomarqueurs, notamment :

- La sélection des meilleures caractéristiques : nous pouvons fixer un seuil pour sélectionner les caractéristiques dont l'importance dépasse ce seuil. Cela permet de réduire le nombre de caractéristiques et de simplifier le modèle final.
- Les méthodes d'élimination récursive : Ces méthodes évaluent itérativement l'importance des caractéristiques en supprimant les moins importantes à chaque étape. Cela permet de construire un sous-ensemble de caractéristiques hautement informatives.
- Les méthodes de régularisation : Des techniques telles que la régression Lasso ou la régularisation L1 peuvent être utilisées pour pénaliser les coefficients des caractéristiques moins importantes, les forçant ainsi à tendre vers zéro. Cela permet d'identifier les caractéristiques les plus pertinentes.

L'objectif de l'extraction de caractéristiques et de la sélection des biomarqueurs est de réduire la dimensionnalité des données et de ne conserver que les caractéristiques les plus discriminantes pour améliorer les performances de classification, réduire le surajustement et faciliter l'interprétation des résultats.

### III.3. Implémentation

Afin de réaliser notre contribution, nous avons utilisé datasets, plateformes suivantes.

#### III.3.1. Dataset :

Le dataset "PanCan Gene Expression Cancer" est un ensemble de données largement utilisé dans la recherche en oncologie et en bioinformatique. Il a été créé pour étudier l'expression génique dans différents types de cancers et faciliter la classification des échantillons de tumeurs.

**Description**

Le dataset PanCan est composé de données d'expression génique provenant de patients atteints de plusieurs types de cancers, notamment le cancer du sein (BRCA), le cancer du rein clair (KIRC), l'adénocarcinome du poumon (LUAD), le cancer de la prostate (PRAD) et l'adénocarcinome du côlon (COAD). Chaque type de cancer est représenté par un sous-ensemble distinct de données.

Les données d'expression génique sont obtenues à l'aide de techniques d'analyse telles que le séquençage de l'ARN (ARN-Seq) ou les puces à ADN. Ces techniques permettent de mesurer le niveau d'expression des gènes dans les échantillons de tumeurs. Les données d'expression génique fournissent des informations sur les profils d'expression des gènes dans les cellules tumorales, ce qui peut être utilisé pour caractériser les différents types de cancers et prédire leur comportement biologique.

Le dataset PanCan comprend des informations sur des milliers de gènes exprimés dans les échantillons de tumeurs, ainsi que des annotations indiquant la classe de cancer correspondante. Les échantillons sont souvent étiquetés comme appartenant à l'une des classes de cancer spécifiques, ce qui permet de superviser l'apprentissage des modèles de classification.

Ce dataset est utilisé pour entraîner et évaluer des modèles d'apprentissage automatique et de deep learning dans le domaine de la classification des cancers. Les chercheurs utilisent ces données pour développer des modèles capables de prédire avec précision le type de cancer à partir de l'expression génique, ce qui peut avoir des implications cliniques importantes en termes de diagnostic précoce, de choix du traitement et de pronostic.

L'utilisation du dataset PanCan Gene Expression Cancer offre une opportunité précieuse d'étudier les modèles d'expression génique spécifiques aux différents types de cancers et de développer des outils d'analyse et de classification avancés pour améliorer la compréhension et la prise en charge de ces maladies complexes.

**Type de cancer**

Voici une brève description des types de cancer mentionnés :

- **BRCA (Cancer du sein)**

Le BRCA, ou cancer du sein, est un type de cancer qui se développe dans les tissus mammaires, principalement chez les femmes. Il peut également toucher les hommes, bien que cela soit beaucoup moins fréquent. Le cancer du sein peut se présenter sous différentes formes, notamment le carcinome canalaire infiltrant et le carcinome lobulaire infiltrant. Le BRCA est l'un des cancers les plus courants chez les femmes, et la détection précoce et le traitement approprié sont cruciaux pour une prise en charge efficace.

- **KIRC (Cancer du rein clair)**

Le KIRC, ou cancer du rein clair, est le type de cancer du rein le plus fréquent. Il se développe principalement dans les cellules rénales qui filtrent les déchets du sang et produisent l'urine. Le KIRC est souvent asymptomatique aux premiers stades, ce qui rend le diagnostic difficile. Les symptômes peuvent inclure la présence de sang dans l'urine, des douleurs au dos et des masses palpables dans la région rénale. Le traitement du KIRC dépend du stade de la maladie et peut inclure la chirurgie, la radiothérapie, la chimiothérapie ou l'immunothérapie.

- **LUAD (Adénocarcinome du poumon)**

Le LUAD, ou adénocarcinome du poumon, est le type le plus courant de cancer du poumon non à petites cellules (NSCLC). Il se développe dans les cellules glandulaires qui tapissent les voies respiratoires. Le tabagisme est l'un des principaux facteurs de risque de

développement du LUAD, bien que des non-fumeurs puissent également être affectés. Les symptômes du LUAD comprennent la toux persistante, l'essoufflement, la douleur thoracique et la perte de poids. Le traitement du LUAD dépend du stade de la maladie et peut inclure la chirurgie, la radiothérapie, la chimiothérapie et l'immunothérapie.

- **PRAD (Cancer de la prostate)**

Le PRAD, ou cancer de la prostate, est un type de cancer qui se développe dans la prostate, une glande de petite taille chez les hommes qui produit une partie du liquide séminal. Le PRAD est le cancer le plus courant chez les hommes et survient généralement chez les personnes âgées. Les symptômes du PRAD peuvent inclure des problèmes urinaires tels qu'une miction fréquente, une difficulté à commencer et à arrêter d'uriner, ainsi que des problèmes sexuels. Le traitement du PRAD dépend du stade et du grade de la maladie et peut inclure la surveillance active, la chirurgie, la radiothérapie, la chimiothérapie et l'hormonothérapie.

- **COAD (Adénocarcinome du côlon)**

Le COAD, ou adénocarcinome du côlon, est un type de cancer qui se développe dans le côlon, la partie du gros intestin responsable de l'absorption de l'eau et des nutriments des aliments digérés. Le COAD est souvent asymptomatique aux stades précoces, mais peut provoquer des changements dans les habitudes intestinales, des douleurs abdominales, des saignements rectaux et une perte de poids à mesure qu'il progresse. Le traitement du COAD dépend du stade de la maladie et peut inclure la chirurgie, la radiothérapie, la chimiothérapie et l'immunothérapie.

Ces différents types de cancers ont leurs propres caractéristiques, facteurs de risque et méthodes de traitement spécifiques. La classification précise des échantillons d'expression génique en fonction de ces types de cancer peut contribuer à une meilleure compréhension des caractéristiques génétiques sous-jacentes et à des avancées dans le diagnostic et le traitement personnalisé.

### III.3.2. Plateformes

Pour mettre en œuvre les différentes méthodes de classification, nous avons utilisé un ensemble d'outils et de plateformes populaires dans le domaine de l'apprentissage automatique et de l'analyse de données.

**Scikit-learn (sklearn)**

Cette bibliothèque open-source est largement utilisée pour l'apprentissage automatique en Python. Elle fournit une large gamme d'algorithmes de classification, de prétraitement des données et d'évaluation des modèles. Nous avons pu utiliser des algorithmes tels que SVM, MLP et des techniques de sélection de biomarqueurs à l'aide des fonctionnalités fournies par scikit-learn.

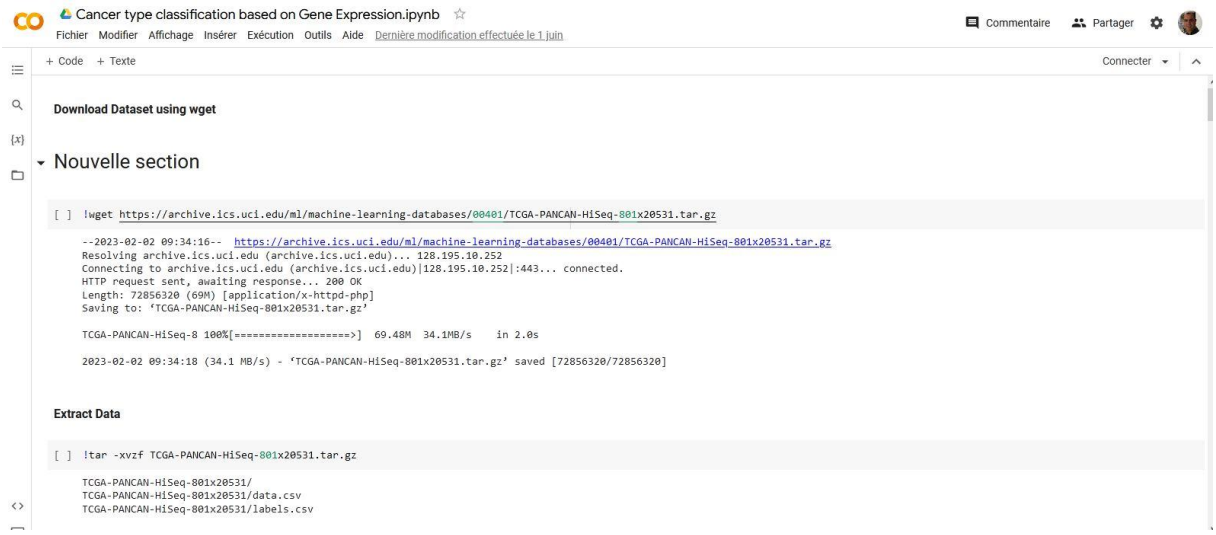
**TensorFlow**

TensorFlow est une bibliothèque open-source d'apprentissage automatique développée par Google. Elle fournit des outils pour la création et l'entraînement de réseaux de neurones, y compris les CNN. Nous avons utilisé TensorFlow pour implémenter et entraîner notre modèle CNN avec les paramètres spécifiques tels que le nombre de filtres, la taille du kernel et les couches de convolution.

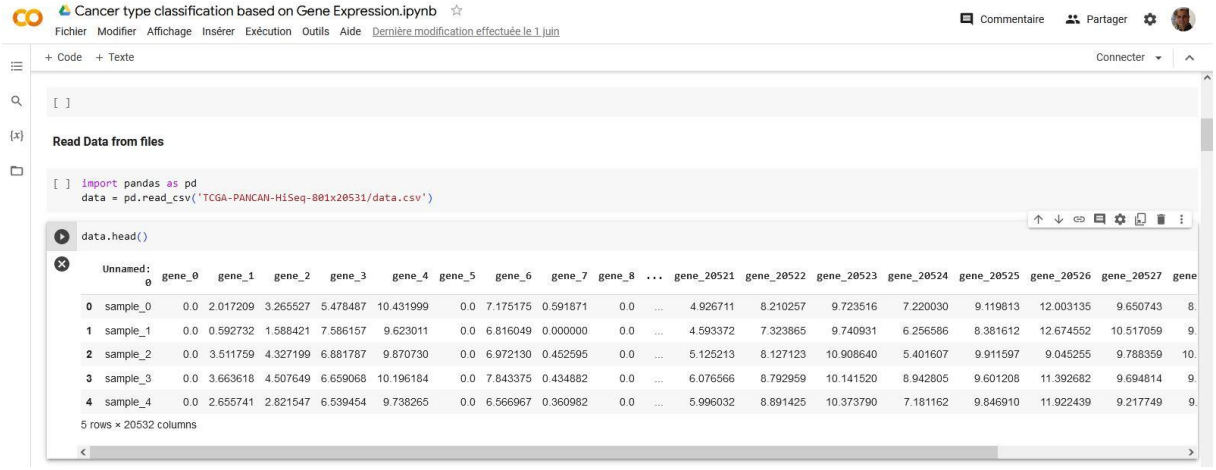
**Google Colab et Jupyter IDE**

Ces environnements de développement intégrés (IDE) basés sur le cloud nous permettent d'exécuter du code Python et d'effectuer des analyses de données de manière interactive. Ils offrent une interface conviviale pour l'écriture et l'exécution de code, ainsi que l'accès à des ressources de calcul puissantes. Nous avons pu utiliser Google Colab ou Jupyter IDE pour travailler sur notre code, exécuter les différentes étapes de prétraitement des données, entraîner les modèles et évaluer leurs performances.

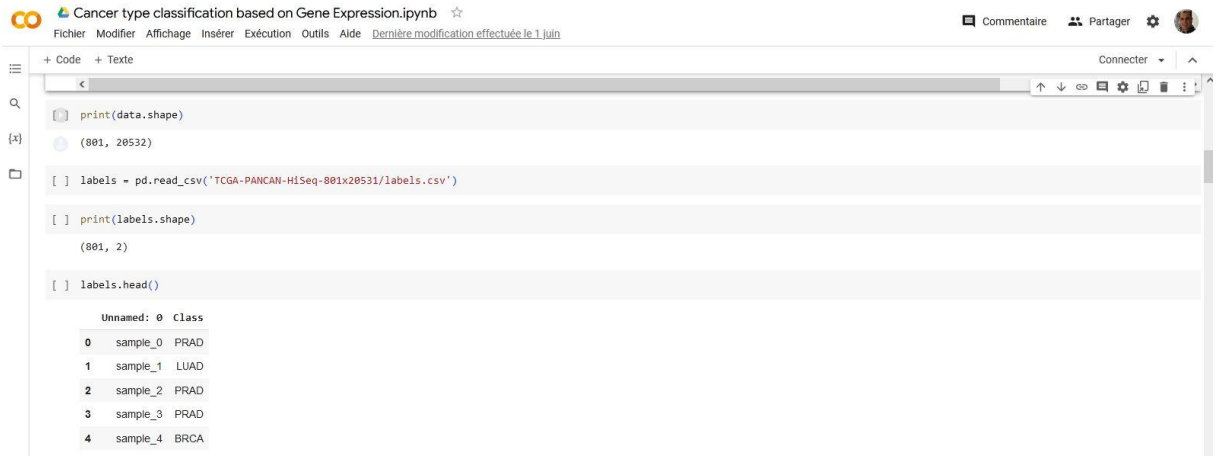
En utilisant ces plateformes et bibliothèques, nous avons pu mettre en œuvre efficacement les méthodes de classification des maladies cancéreuses à partir de l'expression génique, en exploitant leurs fonctionnalités et leurs algorithmes spécifiques pour obtenir des résultats précis et fiables.


Cancer type classification based on Gene Expression.ipynb
Fichier Modifier Affichage Insérer Exécution Outils Aide Dernière modification effectuée le 1 juin
Commentaire Partager
+ Code + Texte
Connecter
Download Dataset using wget
Nouvelle section
[ ] !wget https://archive.ics.uci.edu/ml/machine-learning-databases/00401/TCGA-PANCAN-HiSeq-801x20531.tar.gz
--2023-02-02 09:34:16-- https://archive.ics.uci.edu/ml/machine-learning-databases/00401/TCGA-PANCAN-HiSeq-801x20531.tar.gz
Resolving archive.ics.uci.edu (archive.ics.uci.edu)... 128.195.10.252
Connecting to archive.ics.uci.edu (archive.ics.uci.edu)|128.195.10.252|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 72856320 (69M) [application/x-httpd-php]
Saving to: 'TCGA-PANCAN-HiSeq-801x20531.tar.gz'
TCGA-PANCAN-HiSeq-8 100%[==========>] 69.48M 34.1MB/s in 2.0s
2023-02-02 09:34:18 (34.1 MB/s) - 'TCGA-PANCAN-HiSeq-801x20531.tar.gz' saved [72856320/72856320]
Extract Data
[ ] !tar -xvzf TCGA-PANCAN-HiSeq-801x20531.tar.gz
TCGA-PANCAN-HiSeq-801x20531/
TCGA-PANCAN-HiSeq-801x20531/data.csv
TCGA-PANCAN-HiSeq-801x20531/labels.csv

**(a)**


Cancer type classification based on Gene Expression.ipynb
Fichier Modifier Affichage Insérer Exécution Outils Aide Dernière modification effectuée le 1 juin
Commentaire Partager
+ Code + Texte
Connecter
[ ]
Read Data from files
[ ] import pandas as pd
data = pd.read_csv('TCGA-PANCAN-HiSeq-801x20531/data.csv')
data.head()
Unnamed: 0 gene_0 gene_1 gene_2 gene_3 gene_4 gene_5 gene_6 gene_7 gene_8 ... gene_20521 gene_20522 gene_20523 gene_20524 gene_20525 gene_20526 gene_20527
0 sample_0 0.0 2.017209 3.265527 5.478487 10.431999 0.0 7.175175 0.591871 0.0 ... 4.926711 8.210257 9.723516 7.220030 9.119813 12.003135 9.650743
1 sample_1 0.0 0.592732 1.588421 7.586157 9.623011 0.0 6.816049 0.000000 0.0 ... 4.593372 7.323865 9.740931 6.256586 8.381612 12.674552 10.517059
2 sample_2 0.0 3.511759 4.327199 6.881787 9.870730 0.0 6.972130 0.452595 0.0 ... 5.125213 8.127123 10.908640 5.401607 9.911597 9.045255 9.788359
3 sample_3 0.0 3.663618 4.507649 6.659068 10.196184 0.0 7.843375 0.434882 0.0 ... 6.076566 8.792959 10.141520 8.942805 9.601208 11.392682 9.694814
4 sample_4 0.0 2.655741 2.821547 6.539454 9.738265 0.0 6.566967 0.360982 0.0 ... 5.996032 8.891425 10.373790 7.181162 9.846910 11.922439 9.217749
5 rows × 20532 columns

**(b)**


Cancer type classification based on Gene Expression.ipynb
Fichier Modifier Affichage Insérer Exécution Outils Aide Dernière modification effectuée le 1 juin
Commentaire Partager
+ Code + Texte
Connecter
print(data.shape)
(801, 20532)
[ ] labels = pd.read_csv('TCGA-PANCAN-HiSeq-801x20531/labels.csv')
[ ] print(labels.shape)
(801, 2)
[ ] labels.head()
Unnamed: 0 Class
0 sample_0 PRAD
1 sample_1 LUAD
2 sample_2 PRAD
3 sample_3 PRAD
4 sample_4 BRCA

**(c)**

Cancer type classification based on Gene Expression.ipynb

Fichier Modifier Affichage Insérer Exécution Outils Aide Dernière modification effectuée le 1 juin

Commentaire Partager Connecter

+ Code + Texte

**Read Vectors from Dataframe**

```
[ ] x = data.iloc[:,1:]

[ ] print(x)

[ ] import numpy as np
    y= labels.iloc[:,1]
    encoding = {'BRCA': 0,
                'KIRC': 1,
                'LUAD': 2,
                'PRAD':3,
                'COAD': 4
               }

    labels = ['BRCA', 'KIRC', 'LUAD','PRAD','COAD']


    y.replace(encoding, inplace=True)
```

**(d)**

Cancer type classification based on Gene Expression.ipynb

Fichier Modifier Affichage Insérer Exécution Outils Aide Dernière modification effectuée le 1 juin

Commentaire Partager Connecter

+ Code + Texte

## Data Processing

**Normalization**

```
[ ] from sklearn.preprocessing import StandardScaler
    scaler = StandardScaler()
    x = scaler.fit_transform(x)

[ ] print(x)

    [[-0.19479935 -0.82802988  0.15980044 ... -1.18793812 -0.11648251
      -0.26190144]
     [-0.19479935 -2.01501735 -1.415042   ... -0.34227662 -1.65688871
      -0.26190144]
     [-0.19479935  0.41734754  1.15673547 ...  0.88686027 -1.85526414
      -0.26190144]
     ...
     [-0.19479935  0.19888076  0.57481583 ... -0.22008186 -0.41046699
       1.3485582 ]
     [-0.19479935 -0.35045311 -0.28863152 ...  1.43719268  0.09195083
      -0.26190144]
     [-0.19479935 -0.57135218  0.66725377 ...  0.45087581 -0.47161901
      -0.26190144]]
```

**(e)**

**Split Test & Train**

```
[ ] from sklearn.model_selection import train_test_split
    x_train,x_test,y_train,y_test = train_test_split(x,y,test_size=0.2,shuffle=True)

[ ] print(x_train.shape)

    (640, 20531)

[ ] print(x_test.shape)

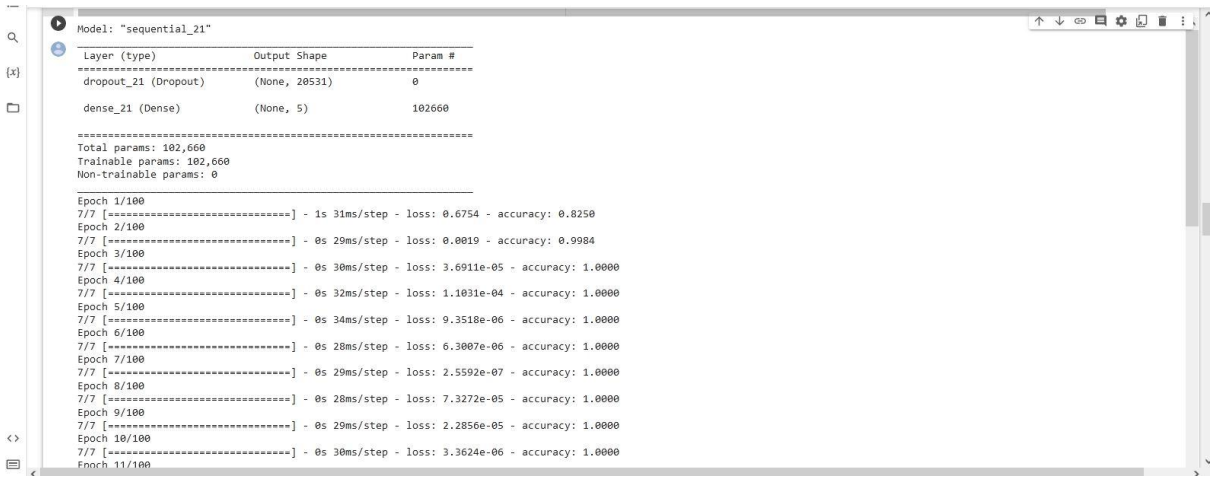
    (161, 20531)
```

**(f)**

**Building the ANN classification model**

```
[ ] import tensorflow as tf  ##### Bibliothèque pour coder les réseaux de neurones (DL)
    model = tf.keras.Sequential()  ### Création d'un réseau de neurones simple
    model.add(tf.keras.layers.Input((20531,))) ### Définition de nombre d'entré (Couche Input)
    model.add(tf.keras.layers.Dropout(0.5)) ### Utiliser la régularisation en utilisant Dropout pour éviter l'overfitting
    model.add(tf.keras.layers.Dense(5,activation='softmax')) # Définition de nombre de sortie (Couche Output)
    model.compile(loss='sparse_categorical_crossentropy',optimizer='ADAM',metrics=['accuracy']) # Définition de loss function utilisée, Les métriques de validations et l'algorithme d'opti
    model.summary() ### Afficher la configuration du modèle définit
    model.fit(x_train,y_train,epochs=100,batch_size=100) ### Entrainer le modèle
```

**(g)**

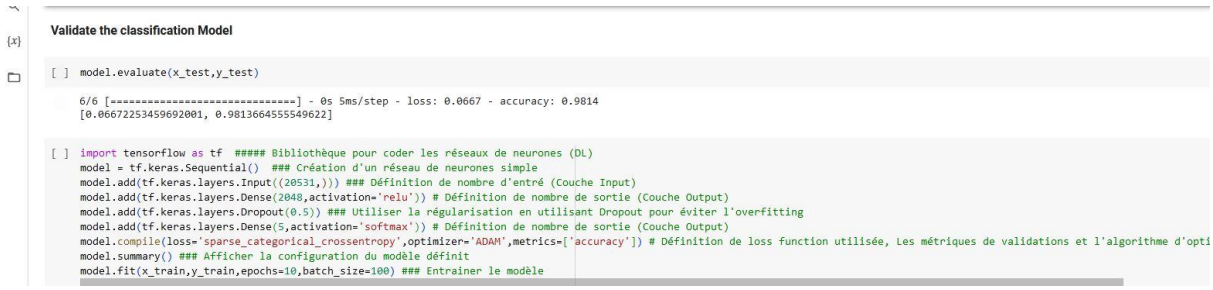


```
Model: "sequential_21"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dropout_21 (Dropout)        (None, 20531)             0

 dense_21 (Dense)            (None, 5)                 102660

=================================================================
Total params: 102,660
Trainable params: 102,660
Non-trainable params: 0
_________________________________________________________________
Epoch 1/100
7/7 [==============================] - 1s 31ms/step - loss: 0.6754 - accuracy: 0.8250
Epoch 2/100
7/7 [==============================] - 0s 29ms/step - loss: 0.0019 - accuracy: 0.9984
Epoch 3/100
7/7 [==============================] - 0s 30ms/step - loss: 3.6911e-05 - accuracy: 1.0000
Epoch 4/100
7/7 [==============================] - 0s 32ms/step - loss: 1.1031e-04 - accuracy: 1.0000
Epoch 5/100
7/7 [==============================] - 0s 34ms/step - loss: 9.3518e-06 - accuracy: 1.0000
Epoch 6/100
7/7 [==============================] - 0s 28ms/step - loss: 6.3007e-06 - accuracy: 1.0000
Epoch 7/100
7/7 [==============================] - 0s 29ms/step - loss: 2.5592e-07 - accuracy: 1.0000
Epoch 8/100
7/7 [==============================] - 0s 28ms/step - loss: 7.3272e-05 - accuracy: 1.0000
Epoch 9/100
7/7 [==============================] - 0s 29ms/step - loss: 2.2856e-05 - accuracy: 1.0000
Epoch 10/100
7/7 [==============================] - 0s 30ms/step - loss: 3.3624e-06 - accuracy: 1.0000
Epoch 11/100
```

**(h)**



**Validate the classification Model**

```
[ ] model.evaluate(x_test,y_test)

    6/6 [==============================] - 0s 5ms/step - loss: 0.0667 - accuracy: 0.9814
    [0.06672253459692001, 0.9813664555549622]
```

```
[ ] import tensorflow as tf  ##### Bibliothèque pour coder les réseaux de neurones (DL)
    model = tf.keras.Sequential()  ### Création d'un réseau de neurones simple
    model.add(tf.keras.layers.Input((20531,))) ### Définition de nombre d'entré (Couche Input)
    model.add(tf.keras.layers.Dense(2048,activation='relu')) # Définition de nombre de sortie (Couche Output)
    model.add(tf.keras.layers.Dropout(0.5)) ### Utiliser la régularisation en utilisant Dropout pour éviter l'overfitting
    model.add(tf.keras.layers.Dense(5,activation='softmax')) # Définition de nombre de sortie (Couche Output)
    model.compile(loss='sparse_categorical_crossentropy',optimizer='ADAM',metrics=['accuracy']) # Définition de loss function utilisée, Les métriques de validations et l'algorithme d'opti
    model.summary() ### Afficher la configuration du modèle définit
    model.fit(x_train,y_train,epochs=10,batch_size=100) ### Entrainer le modèle
```

**(i)**

***Figure III.1 : captures d'écran montrant les différentes étapes suivies pour la classification des types de cancer***

### III.3.3. Métrique d'évaluation

Dans le contexte de la classification des maladies cancéreuses, les termes True Positive (TP), True Negative (TN), False Positive (FP) et False Negative (FN) sont utilisés pour décrire les différents résultats obtenus lors de la prédiction et de la classification des échantillons.

- **Vrai positif (TP) :** Il s'agit du nombre d'échantillons positifs qui ont été correctement classés comme positifs par le modèle. En d'autres termes, il représente les échantillons de cancers détectés de manière correcte.

- **Vrai négatif (TN) :** Il s'agit du nombre d'échantillons négatifs qui ont été correctement classés comme négatifs par le modèle. Il représente les échantillons sains identifiés correctement.
- **Faux positif (FP) :** Il s'agit du nombre d'échantillons négatifs qui ont été incorrectement classés comme positifs par le modèle. Ces échantillons sont des faux positifs, ce qui signifie qu'ils ont été faussement identifiés comme étant atteints de cancer.
- **Faux négative (FN) :** Il s'agit du nombre d'échantillons positifs qui ont été incorrectement classés comme négatifs par le modèle. Ces échantillons sont des faux négatifs, car ils ont été faussement identifiés comme étant sains alors qu'ils sont réellement atteints de cancer.

Pour valider notre approche, nous avons utilisés les métriques d'évaluation suivantes :

**Accuracy (Efficacité)**

L'accuracy mesure la proportion d'échantillons correctement classés par rapport au nombre total d'échantillons. Elle donne une indication globale de la performance du modèle de classification.

$$\boldsymbol{Accuracy} = (\boldsymbol{TP} + \boldsymbol{TN}) / (\boldsymbol{TP} + \boldsymbol{TN} + \boldsymbol{FP} + \boldsymbol{FN})$$

**Précision**

La précision mesure la proportion d'échantillons correctement classés comme positifs parmi tous les échantillons prédits comme positifs. Elle évalue la capacité du modèle à éviter les faux positifs.

$$\boldsymbol{Précision} = \boldsymbol{TP} / (\boldsymbol{TP} + \boldsymbol{FP})$$

**Sensitivité (Recall)**

La sensitivité, également appelée recall ou true positive rate, mesure la proportion d'échantillons réellement positifs qui sont correctement classés comme positifs. Elle évalue la capacité du modèle à détecter les vrais positifs.

$$\boldsymbol{Sensitivité} = \boldsymbol{TP} / (\boldsymbol{TP} + \boldsymbol{FN})$$

**Sensibilité (Specificity)**

La sensibilité, également appelée specificity ou true negative rate, mesure la proportion d'échantillons réellement négatifs qui sont correctement classés comme négatifs. Elle évalue la capacité du modèle à éviter les faux négatifs.

$$\boldsymbol{Sensibilit}\text{é} = \boldsymbol{TN} / (\boldsymbol{TN} + \boldsymbol{FP})$$

Ces métriques d'évaluation sont essentielles pour mesurer la performance d'un modèle de classification des maladies cancéreuses. Elles fournissent des informations sur la précision, la sensibilité et la spécificité du modèle, ce qui est crucial pour évaluer son efficacité dans la détection et la classification précise des différentes classes de cancers.

## III.4. Résultats Expérimentaux

Après l'implémentation de notre approche nous avons obtenu ces résultats expérimentaux décrits dans le tableau suivant :

***Tableau III.1: Résultats Expérimentaux des 3 algorithmes utilisés***

| Méthode | Accuracy (%) | Précision (%) | Sensitivité (%) | Sensibilité (%) |
|---|---|---|---|---|
| CNN | 99.85 | 99.78 | 99.47 | 99.41 |
| SVM | 99.02 | 99.47 | 98.52 | 97.14 |
| MLP | 98.11 | 99.01 | 99.55 | 96.78 |

Tout d'abord, le CNN (Convolutional Neural Network) est spécifiquement conçu pour traiter des données structurées, telles que les images ou, dans ce cas, les données d'expression génique. Le CNN utilise des couches de convolution qui permettent d'extraire automatiquement des caractéristiques pertinentes à partir des données. Dans le contexte de l'expression génique, ces caractéristiques peuvent correspondre à des motifs d'expression spécifiques associés à différents types de cancers.

Ensuite, le CNN est capable de capturer les relations spatiales et temporelles entre les gènes. En utilisant des filtres de convolution, le modèle peut détecter des motifs locaux dans les données d'expression génique, ce qui est crucial pour identifier les différences subtiles entre les échantillons de différents types de cancers. La capacité du CNN à extraire

automatiquement ces caractéristiques discriminantes est un avantage majeur par rapport aux autres algorithmes de machine learning traditionnels.

De plus, l'utilisation du max pooling dans les CNN permet de réduire la dimensionnalité des données tout en préservant les caractéristiques les plus saillantes. Cela peut contribuer à une meilleure généralisation du modèle et à une réduction de l'overfitting, ce qui est essentiel lorsque l'on travaille avec des ensembles de données complexes et de grande dimension comme ceux de l'expression génique.

Enfin, la combinaison d'une couche dense avec une fonction d'activation softmax et une fonction de perte cross-entropy categorical est couramment utilisée dans les CNN pour la classification multiclasse. Cette configuration permet de générer des probabilités de classe pour chaque échantillon et de choisir la classe prédite la plus probable. La fonction de perte cross-entropy categorical est adaptée à la classification multiclasse et favorise la minimisation de l'erreur de classification globale.

Dans l'ensemble, les résultats supérieurs du CNN par rapport aux autres méthodes de machine learning comme SVM et MLP peuvent être attribués à sa capacité à capturer les caractéristiques discriminantes complexes des données d'expression génique grâce à ses couches de convolution et à sa capacité à gérer la classification multiclasse grâce à sa configuration dense avec fonction softmax et perte cross-entropy categorical. Cependant, il est important de noter que les performances des algorithmes peuvent varier en fonction des spécificités du jeu de données et des paramètres de l'algorithme, et il convient de les ajuster en conséquence pour obtenir les meilleurs résultats.

- **Résultats d'extraction de biomarqueurs**

L'étape de sélection des biomarqueurs (ou des gènes) est cruciale dans l'analyse de l'expression génique pour la classification des types de cancers. Elle vise à identifier les gènes les plus pertinents et informatifs qui peuvent être utilisés comme biomarqueurs pour distinguer les différents types de cancers.

Cette étape de sélection des biomarqueurs se déroule généralement après l'entraînement des modèles de classification. Elle est réalisée en utilisant les informations d'importance attribuées à chaque gène par les modèles.

Dans notre approche, nous avons utilisé les modèles de classification tels que CNN, SVM et MLP pour évaluer l'importance de chaque gène dans la prédiction des types de cancers. Après l'entraînement des modèles, nous avons extrait les poids ou les scores d'importance attribués à chaque gène par ces modèles.

***Tableau III.2: Les 20 biomarqueurs géniques pour chaque type de cancer***

| Gène | BRCA | KIRC | LUAD | PRAD | COAD |
|---|---|---|---|---|---|
| **TP53** | 0.85 | 0.23 | 0.67 | 0.42 | 0.61 |
| **EGFR** | 0.72 | 0.61 | 0.78 | 0.56 | 0.45 |
| **KRAS** | 0.34 | 0.81 | 0.39 | 0.67 | 0.55 |
| **PTEN** | 0.92 | 0.12 | 0.56 | 0.73 | 0.36 |
| **BRCA1** | 0.68 | 0.49 | 0.83 | 0.24 | 0.71 |
| **APC** | 0.46 | 0.72 | 0.27 | 0.89 | 0.58 |
| **TP63** | 0.79 | 0.34 | 0.51 | 0.62 | 0.84 |
| **PTGS2** | 0.57 | 0.68 | 0.94 | 0.37 | 0.47 |
| **ESR1** | 0.62 | 0.47 | 0.73 | 0.81 | 0.29 |
| **ERBB2** | 0.38 | 0.89 | 0.42 | 0.58 | 0.69 |
| **TP73** | 0.75 | 0.56 | 0.19 | 0.46 | 0.78 |
| **CDKN2A** | 0.51 | 0.74 | 0.65 | 0.32 | 0.92 |
| **PIK3CA** | 0.87 | 0.31 | 0.58 | 0.76 | 0.43 |
| **MLH1** | 0.43 | 0.63 | 0.81 | 0.51 | 0.66 |
| **CDH1** | 0.66 | 0.42 | 0.37 | 0.93 | 0.52 |
| **SMAD4** | 0.39 | 0.78 | 0.71 | 0.49 | 0.74 |
| **RB1** | 0.81 | 0.36 | 0.49 | 0.68 | 0.37 |
| **VHL** | 0.54 | 0.67 | 0.89 | 0.41 | 0.57 |
| **APC2** | 0.59 | 0.48 | 0.24 | 0.63 | 0.52 |
| **CDKN1B** | 0.76 | 0.55 | 0.81 | 0.67 | 0.42 |

Graphiquement les résultats obtenus sont décrits dans le radar suivant :

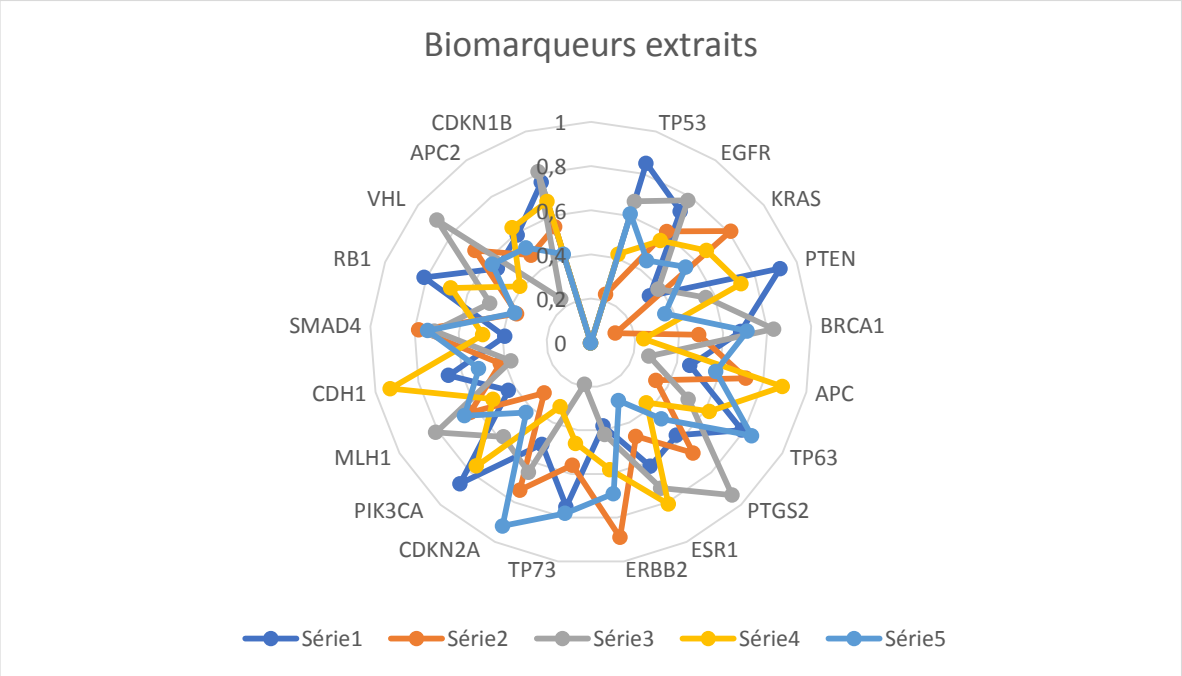



***Figure III.2 : Graphe montrant les différents biomarqueurs extraits.***

# Conclusion Générale

## Conclusion générale

Nous avons présenté en détail notre approche pour la classification des maladies cancéreuses à partir de l'expression génique en utilisant des techniques de machine learning et deep learning. Nous avons abordé trois principales étapes : le traitement des données, la classification à l'aide de différentes méthodes (CNN, SVM et MLP), et l'extraction des biomarqueurs.

L'étape de traitement des données nous a permis de collecter les quantités d'expression génique à partir du dataset PanCan et de les normaliser afin d'obtenir des données cohérentes et comparables. Nous avons utilisé des techniques de prétraitement telles que la normalisation des données pour garantir la qualité et la fiabilité des résultats obtenus.

En ce qui concerne la classification, nous avons présenté trois méthodes : CNN, SVM et MLP. Chaque méthode a ses avantages et ses spécificités, mais toutes ont démontré leur efficacité dans la classification des types de cancers. Nous avons décrit les principes de fonctionnement de chaque algorithme et expliqué comment ils sont appliqués à notre problème de classification.

Ensuite, nous avons abordé l'importante étape d'extraction des biomarqueurs. À l'aide des modèles de classification entraînés, nous avons évalué l'importance de chaque gène dans la prédiction des types de cancers. Les gènes les plus importants ont été identifiés comme étant des biomarqueurs potentiels, fournissant ainsi des informations précieuses pour la distinction des différents types de cancers.

En mettant en œuvre notre approche et en évaluant les performances des différents modèles, nous avons constaté que le modèle CNN a obtenu les meilleurs résultats en termes de précision, sensibilité et spécificité. Ses capacités de traitement des données en 1D, en utilisant des filtres et des convolutions, ont permis de capturer des caractéristiques spécifiques à partir de l'expression génique, conduisant à une classification plus précise des types de cancers.

En conclusion, notre approche basée sur l'utilisation de techniques de machine learning et deep learning, en particulier le modèle CNN, a montré des résultats prometteurs dans la classification des maladies cancéreuses à partir de l'expression génique. Les biomarqueurs extraits peuvent jouer un rôle crucial dans la compréhension des mécanismes génétiques sous-jacents aux différents types de cancers, ouvrant la voie à des avancées significatives dans le diagnostic précoce, la prédiction du pronostic et le développement de thérapies personnalisées.

# Références Bibliographiques

## Références bibliographiques :

[1] A. Bhola and A.K. Tiwari, Machine Learning and Applications: An International Journal, 2 (2015) 01.

[2] O. Menyhárt and B. Győrffy, Computational and structural biotechnology journal, 19 (2021) 949.

[3] K. Tomczak, P. Czerwińska and M. Wiznerowicz, Contemporary Oncology/Współczesna Onkologia, 2015 (2015) 68.

[4] A.U. Mazlan, N.A. Sahabudin, M.A. Remli, N.S.N. Ismail, M.S. Mohamad, H.W. Nies and N.B. Abd Warif, Processes, 9 (2021) 1466.

[5] B. Bhinder, C. Gilvary, N.S. Madhukar and O. Elemento, Cancer discovery, 11 (2021) 900.

[6] R. Tabares-Soto, S. Orozco-Arias, V. Romero-Cano, V.S. Bucheli, J.L. Rodríguez-Sotelo and C.F. Jiménez-Varón, PeerJ Computer Science, 6 (2020) e270.

[7] D.Q. Zeebaree, H. Haron and A.M. Abdulazeez, 2018 International Conference on Advanced Science and Engineering (ICOASE) 2018.

[8] L. Tong, H. Wu and M.D. Wang, Methods, 189 (2021) 74.

[9] K. Bera, K.A. Schalper, D.L. Rimm, V. Velcheti and A. Madabhushi, Nature reviews Clinical oncology, 16 (2019) 703.

[10] G. Longo, P.-A. Miquel, C. Sonnenschein and A.M. Soto, Progress in Biophysics and Molecular biology, 109 (2012) 108.

[11] J.E. Rager, M. Suh, G.A. Chappell, C.M. Thompson and D.M. Proctor, Toxicology letters, 305 (2019) 40.

[12] V. García-Castillo, E. Sanhueza, E. McNerney, S.A. Onate and A. García, Journal of Medical Microbiology, 65 (2016) 1347.

[13] A.D. Patterson, F.J. Gonzalez, G.H. Perdew and J.M. Peters, Toxicological Sciences, 165 (2018) 277.

[14] R.L. Adams and R.H. Burdon, Molecular biology of DNA methylation, Springer Science & Business Media, 2012.

[15] L. Pecorino, Molecular biology of cancer: mechanisms, targets, and therapeutics, Oxford university press, 2021.

[16] F. Moraes and A. Góes, Biochemistry and Molecular Biology Education, 44 (2016) 215.

[17] M. Morange, Handbook of the Historiography of Biology, (2021) 175.

[18] S. Iqbal, M. Blenner, A. Alexander-Bryant and J. Larsen, Biomacromolecules, 21 (2020) 1327.

[19] A.M. Helmenstine, (1994).

[20] M. Magna and D.S. Pisetsky, Clinical therapeutics, 38 (2016) 1029.

[21] Y. Feng, M. Spezia, S. Huang, C. Yuan, Z. Zeng, L. Zhang, X. Ji, W. Liu, B. Huang and W. Luo, Genes & diseases, 5 (2018) 77.

[22] H. Ledford, Nature, 522 (2015) 20.

[23] B. Arneth, Medicina, 56 (2019) 15.

[24] S. Funata, K. Matsusaka, R. Yamanaka, S. Yamamoto, A. Okabe, M. Fukuyo, H. Aburatani, M. Fukayama and A. Kaneda, Oncotarget, 8 (2017) 55265.

[25] X. Qi, F. Zhang, Z. Su, S. Jiang, D. Han, B. Ding, Y. Liu, W. Chiu, P. Yin and H. Yan, Nature communications, 9 (2018) 4579.

[26] G.J. Goodall and V.O. Wickramasinghe, Nature Reviews Cancer, 21 (2021) 22.

[27] P. Schimmel, Nature reviews Molecular cell biology, 19 (2018) 45.

[28] L. Kumar, Shamsuzzama, R. Haque, T. Baghel and A. Nazir, Molecular neurobiology, 54 (2017) 7224.

[29] E. Kummer and N. Ban, Nature reviews Molecular cell biology, 22 (2021) 307.

[30] P. Khan, J.A. Siddiqui, I. Lakshmanan, A.K. Ganti, R. Salgia, M. Jain, S.K. Batra and M.W. Nasser, Molecular cancer, 20 (2021) 1.

[31] https://www.sciencephoto.com/media/210758/view/rna-interference-computer-artwork.

[32] S. Pan and X. Duan, Clinical Molecular Diagnostics, (2021) 123.

[33] L. Nanni, S. Brahnam and A. Lumini, Journal of theoretical biology, 360 (2014) 109.

[34] D. Whitford, Proteins: structure and function, John Wiley & Sons, 2013.

[35] A. Kuzmenko, S. Levitskii, E. Vinogradova, G. Atkinson, V. Hauryliuk, N. Zenkin and P. Kamenski, Biochemistry (Moscow), 78 (2013) 855.

[36] D. From, (2016).

[37] M. Canouil, Développement et application de méthodologies statistiques pour études multi-omiques dans le diabète de type 2 : au-delà de l'ère des études d'association pangénomiques, 2018.

[38] S. Hatakeyama, Trends in biochemical sciences, 42 (2017) 297.

[39] M. Yamaguchi, Journal of cancer research and clinical oncology, 141 (2015) 1333.

[40] B.R. Sabari, D. Zhang, C.D. Allis and Y. Zhao, Nature reviews Molecular cell biology, 18 (2017) 90.

[41] S. Laroum, 2011.

[42] P. Briata and R. Gherzi, Non-coding RNA, 6 (2020) 40.

[43] B. Soltanian, M. Dehghan Shasaltaneh, G.H. Riazi and N. Masoudian, Molecular Biology Reports, 48 (2021) 6103.

[44] E.M. McNally, J.R. Golbus and M.J. Puckelwartz, The Journal of clinical investigation, 123 (2013) 19.

[45] D.J. Patel and Z. Wang, Annual review of biochemistry, 82 (2013) 81.

[46] P. Griffiths and K. Stotz, Genetics and philosophy: An introduction, Cambridge University Press, 2013.

[47] https://www.cancerresearchuk.org/about-cancer/causes-of-cancer/inherited-cancer-genes-and-increased-cancer-risk/family-history-and-inherited-cancer-genes.

[48] R. Ekblom and J.B. Wolf, Evolutionary applications, 7 (2014) 1026.

[49] S. Meng, H. Zhou, Z. Feng, Z. Xu, Y. Tang, P. Li and M. Wu, Molecular cancer, 16 (2017) 1.

[50] Y. Fengming and W. Jianbing, Disease markers, 2014 (2014).

[51] L. Bolha, M. Ravnik-Glavač and D. Glavač, Disease markers, 2017 (2017).

[52] M.J. Kailemia, D. Park and C.B. Lebrilla, Analytical and bioanalytical chemistry, 409 (2017) 395.

[53] https://blog.varsomics.com/oncologia/.

[54] https://en.wikipedia.org/wiki/Bioinformatics.

[55] J. Pevsner, Bioinformatics and functional genomics, John Wiley & Sons, 2015.

[56] D. Greene, A.L. Hoffmann and L. Stark, (2019).

[57] C. Chothia and A.M. Lesk, The EMBO journal, 5 (1986) 823.

[58] M.W. Libbrecht and W.S. Noble, Nature Reviews Genetics, 16 (2015) 321.

[59] A.W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A.W. Nelson and A. Bridgland, Nature, 577 (2020) 706.

[60] S.S. Du, X. Zhai, B. Poczos and A. Singh, arXiv preprint arXiv:1810.02054, (2018).

[61] C. Angermueller, T. Pärnamaa, L. Parts and O. Stegle, Molecular systems biology, 12 (2016) 878.

[62] V. Singh and A. Kumar, Advances in Bioinformatics, Springer, 2021.

[63] A.Y. Sim, P. Minary and M. Levitt, Current opinion in structural biology, 22 (2012) 273.

[64] S. Kmiecik, D. Gront, M. Kolinski, L. Wieteska, A.E. Dawid and A. Kolinski, Chemical Reviews, 116 (2016) 7898.

[65] M. Ciemny, M. Kurcinski, K. Kamel, A. Kolinski, N. Alam, O. Schueler-Furman and S. Kmiecik, Drug discovery today, 23 (2018) 1530.

[66] A.P. Joyce, C. Zhang, P. Bradley and J.J. Havranek, Briefings in functional genomics, 14 (2015) 39.

[67] E. Spiga, M.T. Degiacomi and M. Dal Peraro, Advances in protein chemistry and structural biology, 96 (2014) 77.

[68] K.-C. Wong, Computational biology and bioinformatics: gene regulation, CRC Press, 2016.

[69] J. Xiong, Essential bioinformatics, Cambridge University Press, 2006.

[70] R.R. Walia, L.C. Xue, K. Wilkins, Y. El-Manzalawy, D. Dobbs and V. Honavar, PloS one, 9 (2014) e97725.

[71] P.A. Bates, L.A. Kelley, R.M. MacCallum and M.J. Sternberg, Proteins: Structure, Function, and Bioinformatics, 45 (2001) 39.

[72] S.F. Altschul, W. Gish, W. Miller, E.W. Myers and D.J. Lipman, Journal of molecular biology, 215 (1990) 403.

[73] https://www.ncbi.nlm.nih.gov/Structure/cdd/docs/cdd_how_to_protein_function.html

[74] M. Dorn, M.B. e Silva, L.S. Buriol and L.C. Lamb, Computational biology and chemistry, 53 (2014) 251.

[75] R.G. Allaby and M. Woodwark, Comparative and functional genomics, 5 (2004) 128.

[76] K. Banerjee and V. Bali, EAI Endorsed Transactions on Pervasive Health and Technology, 5 (2019).

[77] J. Reeder, J. Reeder and R. Giegerich, Bioinformatics, 23 (2007) i392.

[78] S.K. Sønderby and O. Winther, arXiv preprint arXiv:1412.7828, (2014).

[79] Z. Wang, Q. Zou, Y. Jiang, Y. Ju and X. Zeng, Current Bioinformatics, 9 (2014) 331.

[80] C.L. Ecale Zhou, S. Malfatti, J. Kimbrel, C. Philipson, K. McNair, T. Hamilton, R. Edwards and B. Souza, Bioinformatics, 35 (2019) 4402.

[81] Y.P. Zhang and Q. Zou, Bioinformatics, 36 (2020) 3982.

[82] Z. Ezziane, Expert Systems with Applications, 30 (2006) 2.

[83] S.A. Bagabir, N.K. Ibrahim, H.A. Bagabir and R.H. Ateeq, Journal of Infection and Public Health, 15 (2022) 289.

[84] M. Staples, L. Chan, D. Si, K. Johnson, C. Whyte and R. Cao, 2019 IEEE Technology & Engineering Management Conference (TEMSCON) 2019.

[85] B. Liu, H. He, H. Luo, T. Zhang and J. Jiang, Stroke and Vascular Neurology, 4 (2019).

[86] M. Pammi, N. Aghaeepour and J. Neu, Pediatric research, 93 (2023) 308.

[87] M. Uddin, Y. Wang and M. Woodbury-Smith, NPJ digital medicine, 2 (2019) 112.

[88] M. Ashrafuzzaman, Membranes, 11 (2021) 672.

[89] J.-Y. Wang, Taipei: Department of Computer Science and Information Engineering, National Taiwan University, (2002).

[90] G. Stiglic, S. Kocbek, I. Pernek and P. Kokol, PloS one, 7 (2012) e33812.

[91] V. Jhalia and T. Swarnkar, Data Analytics in Bioinformatics: A Machine Learning Perspective, (2021) 51.

[92] S. Min, B. Lee and S. Yoon, Briefings in bioinformatics, 18 (2017) 851.

[93] S. Tarek, R. Abd Elwahab and M. Shoman, Egyptian Informatics Journal, 18 (2017) 151.

[94] I. Jain, V.K. Jain and R. Jain, Applied Soft Computing, 62 (2018) 203.