REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Mohamed Khider – BISKRA

Faculté des Sciences Exactes, des Sciences de la Nature et de la Vie

Département d'informatique

N° d'ordre: SIOD 4/M2/2020

Mémoire

Présenté pour obtenir le diplôme de master académique en

Informatique

Parcours : Informatique décisionnelle et optimisation

La classification des récepteurs couplés aux protéines G (RCPG)

Par :
CHEBAANI BADRA

Soutenu le 05 septembre 2020, devant le jury composé de :

		Président
Belounnar Saliha	**MAA**	**Rapporteur**
		Examinateur

Année universitaire : 2019 / 2020

Remerciements

Je veux exprimer de remerciements mâ gratitude envers tous œux en qui, par leur présence, leur soutien, leur disponibilité, et leurs conseils j'ai trouvé courage afin d'accomplir ce projet.

Je commence par remercier Mme Belounnar Saliha qui m'a fait l'honneur d'être mon encadrant. Je la remercie profondément pour son encouragement continu et aussi d'être toujours là pour m'écouter, elle m'a aidé et me guider à retrouver le bon chemin par son sagesse et ses précieux, ce qui m'a donné la force et le courage d'accomplir ce projet.

Je remercie également le Président et les membres du Jury qui me font l'honneur d'accepter de juger mon travail.

Mes sincères remerciements à tous ceux qui ont participé de près ou de loin à la réalisation de ce modeste travail.

Dédicaces

Je dédie ce mémoire à mes chers parents pour leur patience
A ma chère mère et A mon cher père

Qui n'ont jamais cessé, de formuler des prières à mon égard, de me soutenir et de m'épauler pour que je puisse atteindre mes objectifs

A ma sœur et mes frères

Source d'espoir et de motivation

A Mon mari et Mes enfants Amin, Mayar,...

Que Dieu te protège pour moi

A tous mes amis

Particulièrement Responsable de progres Zehour Tedjini. et Elaiza Nouioua, Mimouna Benkaddour, Abir Bensouici, Meriem Tatai…

Ainsi que tous ceux qui m'ont soutenu dans l'accomplissement de cetteoeuvre, je tiens à leur exprimé toute ma gratitude pour leur sincère aide.

Résumé

La prédiction de la fonction des gènes et des protéines est un champ de recherche qui sert de pointe de départ aux analyses expérimentales pour élucider les activités biochimiques et les processus pour lesquels sont impliqués les protéines, elles jouent un rôle primordial dans la compréhension des protéines responsables des maladies, de développer les médicaments les plus efficaces et avoir une médecine préventive contre les fléaux éventuels.

Les méthodes de prédiction des fonctions à grandement évoluer au cours des temps notamment avec l'arrivée de la génomique, l'apparition de données obtenues à grande échelle, le développement des outils bioinformatique pour traiter le sujet de la classification ayant trait à la famille récepteur couplée aux protéines G (RCPG), choisir une méthode de sélection des caractéristiques composition en pseudo acides aminés (PseAAC) qui peut être déterminant car une méthode appropriée permet d'améliorer grandement les performances de la classification, en utilisant la méthode de classification supervisée machine à vecteurs de support (SVM) pour mettre en œuvre l'apprentissage., les taux de classification que nous avons trouvé montre que nos résultats sont compétitifs.

Mot clé : prédiction, famille récepteur couplé aux protéines G (RCPG), composition en pseudo acides aminés (PseAAC), SVM, Apprentissage.

Abstract

The prediction of the function of genes and proteins is a field of research that serves as a starting point for experimental analyzes to elucidate the biochemical activities and processes for which proteins are involved. They play a key role in understanding the proteinsres ponsible. diseases, develop the most effective drugs and have preventive medicine against possibles courges.

The methods of predicting functions to greatly evolve over time especially with the arrival of genomics, the appearance of large-scale data, the development of bioinformatics tools to address the subject of classification relating to the G-protein coupled receptor family (GPCR), choose a method for selecting characteristics pseudo amino acid composition (PseAAC) that can be decisive because an appropriate method can greatly improve the performance of the classification, using the classification method Supervised support vector machine (SVM) to implement learning., The classification rates we found shows that our results are competitive.

Key words: prediction, G protein-coupled receptor family (GPCR), pseudo amino acid composition (PseAAC), SVM, Learning.

ملخص

إن التنبؤ بوظيفة الجينات و البروتينات هو مجال من البحوث يعمل كنقطة انطلاق للتحليلات التجريبية لتوضيح الأنشطة و العمليات الكيميائية الحيوية التي تشارك فيها البروتينات, وهي تلعب دورًا رئيسيًا في فهم البروتينات المسؤولة. الأمراض, و تطوير الأدوية الأكثر فعالية و يكون الدواء الوقائي ضد الآفات المحتملة.

تتطور طرق التنبؤ بالوظائف إلي حد كبير بمرور الوقت خاصة مع وصول الجينوميات, و ظهور البيانات علي نطاق واسع , و تطوير أدوات المعلوماتية لمعالجة موضوع التصنيف المتعلق ب عائلة مستقبلات-G البروتين (GPCR) , اختر طريقة لاختيار خصائص تركيبة الأحماض الأمينة الزائفة (PseAAC) التي يمكن أن تكون حاسمة لأن الطريقة المناسبة يمكن أن تحسن أداء التصنيف بشكل كبير , باستخدام طريقة التصنيف آلة ناقلة الدعم الخاضعة للإشراف (SVM) لتنفيذ التعلم , توضح معدلات التصنيف التي وجدناها أن نتائجنا تنافسية.

الكلمات المفتاحية : التنبؤ, عائلة مستقبلات - G البروتينة (GPCR) , تركيبة الأحماض الأمينية الزائفة (PseAAC) , SVM ,التعلم.

La table des matières

Chapitre III : Conception

Chapitre IV : Implémentation

La table des Figures

LISTE DE TABLEAU

Introduction

Introduction générale

Dans le cadre de leurs travaux, les biologistes ont à collecter et à interpréter un grand nombre de données. Ces dernières années, ce nombre a augmenté de façon exponentielle grâce aux nouvelles technologies expérimentales complexes qui permettent une collecte de données beaucoup plus rapide que leur interprétation. Afin d'aider les biologistes dans la collecte, le stockage et l'analyse de ces données champs multi-disciplinaire impliquant la biologie, l'informatique, les mathématiques, les statistiques dont l'objectif est d'analyser les séquences biologiques et de prédire la structure et la fonction des macromolécules.

De plus en plus, la bioinformatique est développée dans un but d'application à l'agriculture, la pharmacologie, la médecine.

Discipline qui évolue en fonction des nouveaux problèmes posés par la biologie.

La bioinformatique: Discipline plus pragmatique. Développement d'outils pratiques pour l'analyse et l'organisation des données. Moins d'emphase sur l'exactitude ou l'efficacité de la méthode. Dédiée à des applications pratiques comme l'identification de protéines cible pour la conception de médicaments.

Elle représente aussi est une discipline récente qui propose et développe des modèles, des méthodes et des outils afin d'analyser l'information biologique (génomes, protéomes, etc.) Et produire de nouvelles connaissances. Pour mieux comprendre et résoudre les problèmes posés par la biologie.

Ce présent mémoire englobe cinq chapitres :

- **Premier chapitre:** est consacré aux concepts bioinformatique, plus spécifiquement, la protéomique qui découle de la génomique et dont les principaux fondements sont les protéines et leurs composantes, les acides aminés.
- **Deuxième chapitre :** exposition de la classification et les différentes méthodes de classification qui font partie des techniques de fouille de données.
- **Troisième chapitre :** est consacré à la conception de notre système présentant la conception globale et le détail des étapes de la conception.
- **Quatrième chapitre :** pour finir on a mis en apparence l'implémentation de notre application qui contient des captures d'écran et guide d'utilisation de l'application.

Le mémoire qui reprend le détail de l'introduction générale est le résultat de mes différents travaux ayant trait à la prédiction de fonctions des protéines.

Chapitre I : Bioinformatique Et Fonction Des Protéines

I.1 Introduction

Ces dernières années, une croissance massive de l'information biologique recueillie par les communautés scientifiques a vu le jour. L'affluence de ce type d'informations sous la forme de génomes, de séquences protéiques, de données d'expressions génétiques etc.…a conduit à la nécessité de concevoir des outils informatiques performants pour stocker, analyser et interpréter ces données, ce qui a engendré une science nouvelle appelée Bioinformatique.

Le terme bioinformatique signifie littéralement la science de l'informatique appliquée à la recherche biologique.

D'autre part l'informatique c'est la gestion et l'analyse de données en utilisant diverses techniques computationnelles de pointe, en d'autres termes la bioinformatique peut être décrite comme l'application de méthodes computationnelles pour la découverte de connaissances biologiques, elle représente une symbiose de plusieurs domaines différents de la science, notamment l'informatique, la biologie, les mathématiques et les statistiques.

Ce chapitre sera consacré sur une introduction dans le domaine de la bioinformatique ainsi nous aborderons les fondements biologiques de la bioinformatique tel que le génome, les protéines, les acides aminés et nous détaillons les classifications des protéines à base de leurs fonctions à la fin de ce chapitre on a donné une vue particulière sur la famille des RCPG s car notre étude se focalise sur cette famille.

I.2 Bioinformatique

I.2.1 Définition Bioinformatique

La bioinformatique est l'application des techniques de traitement de l'information à la gestion de données biologiques ou, en d'autres termes,

La bio-informatique est un champ de recherche multi-disciplinaire de la biotechnologie où travaillent de concert biologistes, médecins, informaticiens, mathématiciens, physiciens et bio-informaticiens, dans le but de résoudre un problème scientifique posé par la biologie.

Plus généralement, la bio-informatique est l'application de la statistique et de l'informatique à la science biologique. [1]

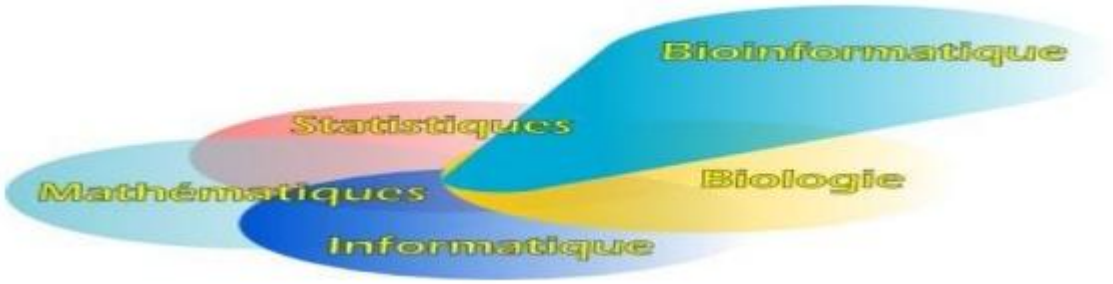


Figure I.1: L'interaction des disciplines construit en bioinformatique.

I.2.2 Objectifs de la bioinformatique

Le rôle de la bioinformatique est d'aider les biologistes dans la collecte et le traitement des données génomiques afin d'étudier la fonction des gènes et des protéines.

Un autre rôle important de la bioinformatique est d'aider les chercheurs des compagnies pharmaceutiques à élaborer des études détaillées des fonctions des protéines afin de faciliter la conception de médicaments.

Les objectifs de la bioinformatique peuvent se résumer comme suit :

- Collecter et stocker des informations dans des bases de données accessibles en ligne.
- Fournir des outils de comparaison de séquences protéiques et nucléotidiques.

  - Identifier une séquence en le comparant aux séquences d'une base de données.

  - Déterminer le degré de similitude entre deux séquences.

  - Repérer des motifs structuraux.

- Fournir des outils de traduction des séquences.

  - Simplifier les tâches de traduction.

  - Proposer plusieurs possibilités de protéines pour une même séquence.

  - Repérer les exons/introns.

- Fournir des outils de prédiction physiologique et fonctionnelle et de prédiction expérimentale.

La bioinformatique nous aide à visualiser les structures invisibles tels que les protéines et d'en apprendre davantage sur leur travail et leur fonction. Cela conduit à comprendre les questions essentielles de la vie: Comment les organismes fonctionnent-ils? Comment la vie s'est-elle développée? Comment peuvent se développer de nouveaux traitements contre des maladies. [2]

I.3 La génomique

La génomique est la science des génomes, elle étudie les séquences d'ADN des êtres vivants, un génome est formé de l'ensemble des informations génétiques contenues dans la cellule. [3]

I.3.1 Acide désoxyribonucléique (ADN)

L'acide désoxyribonucléique ou l'ADN est une macromolécule biologique présente dans toutes les cellules ainsi que chez de nombreux virus, L'ADN contient toute l'information génétique, appelée génome, permettant le développement, le fonctionnement et la reproduction des êtres vivants. Déterminant l'identité biologique permet la survie et l'évolution des organismes vivants en indiquant la façon, et l'endroit où les différents types de protéines sont créés.

Les molécules d'ADN des cellules vivantes sont formées de deux brins antiparallèles enroulés l'un autour de l'autre sous forme de double brin qui prend une structure hélicoïdale. [4]

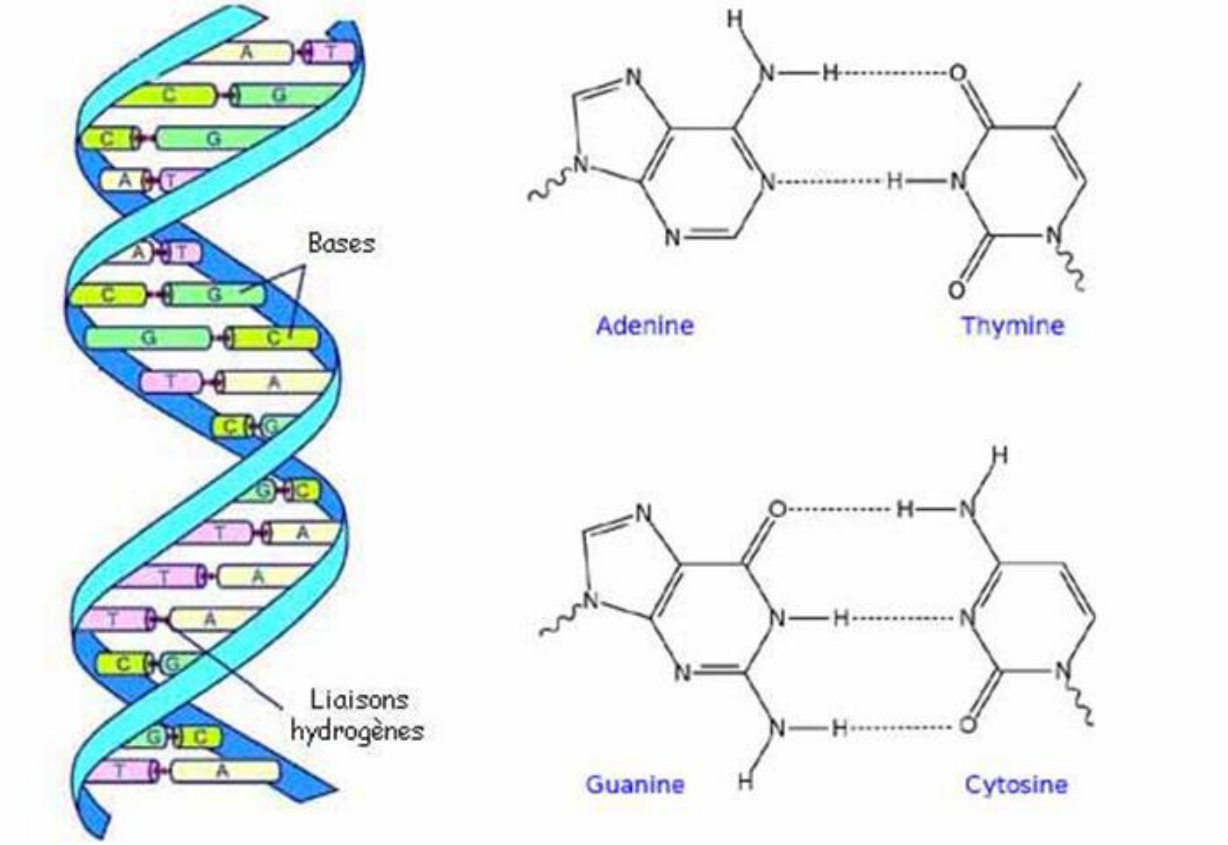


Figure I.2 : Structure d'une molécule d'ADN.

I.3.2 Gènes

Un gène, en génétique, est une unité de base d'hérédité, les gènes sont constitués d'AND, certains gènes agissent comme des instructions pour fabriquer des molécules appelées protéines. [5]

I.3.3 Génome

Le génome est l'ensemble du matériel génétique d'une espèce codé dans son acide désoxyribonucléique (ADN) à l'exception de certains virus dont le génome est constitué d'acide ribonucléique (ARN). Il contient en particulier tous les gènes codant des protéines ou correspondant à des ARN structurés. [6]

I.3.4 Acide ribonucléique (L'ARN)

L'acide ribonucléique (ARN) est une molécule biologique présente chez pratiquement tous les êtres vivants, et aussi L'ARN est très proche chimiquement de l'ADN et il est d'ailleurs en général synthétisé dans les cellules à partir d'une matrice d'ADN dont il est une copie. L'ARN est très proche chimiquement de l'ADN et il est d'ailleurs en général synthétisé dans les

cellules à partir d'une matrice d'ADN dont il est une copie. Les cellules utilisent en particulier l'ARN comme un support intermédiaire des gènes pour synthétiser les protéines dont elles ont besoin.

Chimiquement, l'ARN est un polymère linéaire constitué d'un enchaînement de nucléotides. On trouve quatre bases nucléiques dans l'ARN : l'adénine, la guanine, la cytosine et l'uracile.

L'ARN se trouve le plus souvent dans les cellules sous forme monocaténaire, c'est-à-dire de simple brin, tandis que l'ADN est présent sous forme de deux brins complémentaires formant une double-hélice. Enfin, les molécules d'ARN présentes dans les cellules sont plus courtes que l'ADN du génome, L'ARN est généralement généré par la transcription de l'ADN situé dans le noyau à l'aide d'une enzyme de la famille des ARN polymérases et quelque protéine, l'ARN transmet l'information génétique entre le noyau et les structures cellulaires qui seront chargées de la traduction de son contenu pour produire des protéines en trois étapes:

- **ARN messager (ARNm) :** il est formé par transcription d'un gène de l'ADN dont il est la copie. Son rôle consiste à transporter l'information génétique recueillie du noyau vers le cytoplasme où elle sera traduite en protéine.

- **ARN transfert (ARNt) :** Généralement sous forme d'un tréfile, l'ARN transfert est une courte chaine d'ARN responsable de la traduction des ribosomes d'ARN messager à des acides aminés. Un ARNt est un brin court qui a un anticodon sur sa boucle, et un acide aminé attaché à l'autre extrémité et qui sera transféré à la protéine en formation.

- **ARN ribosomique (ARNr) :** représente 80 % de l'ARN total d'une cellule, il a comme rôle la traduction du code génétique porté par l'ARN messager (ARNm). [7]

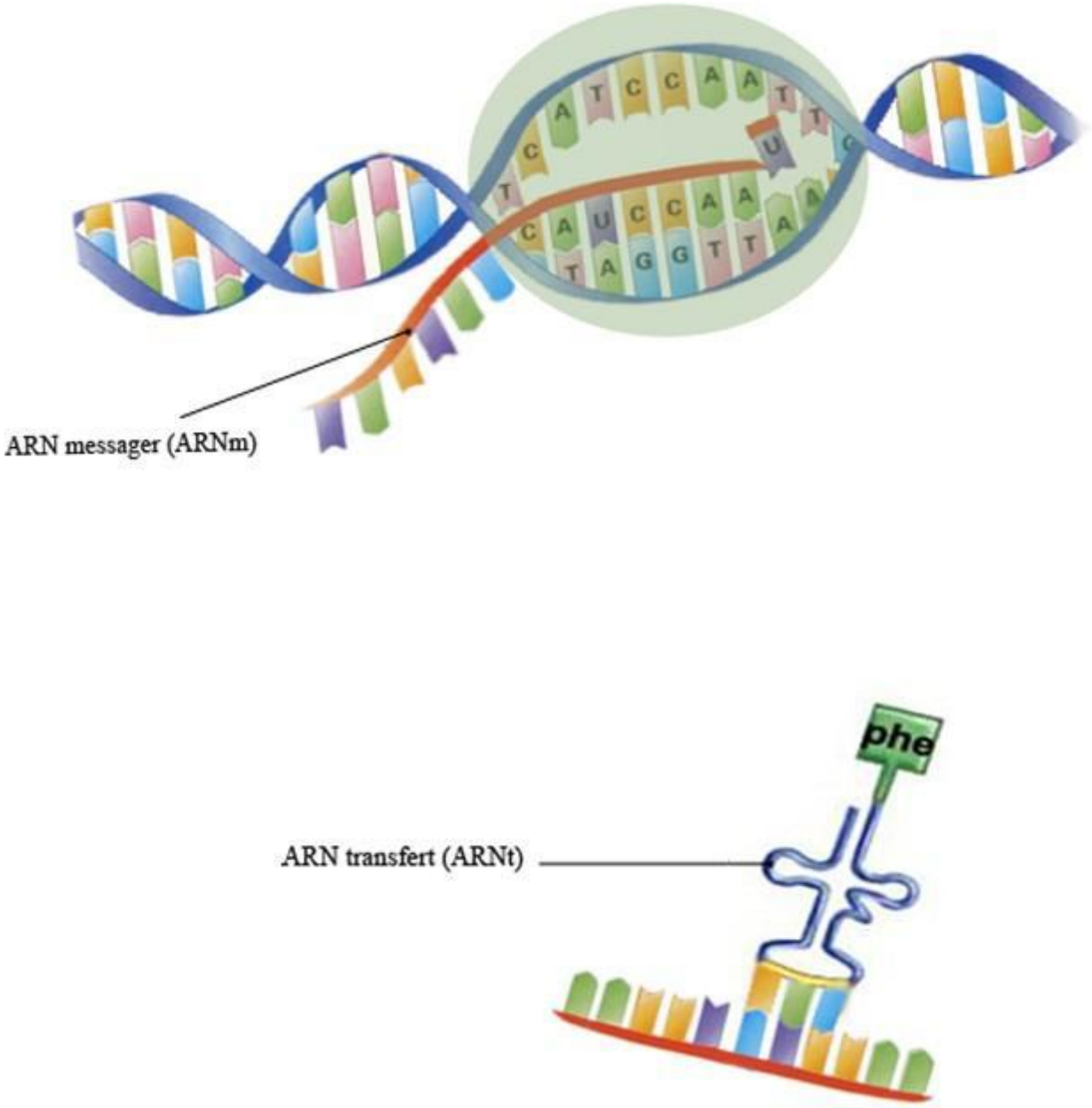


Figure I.3 : Les différents types d'ARN.

I.4 La protéomique

La protéomique désigne la science qui étudie les protéomes, c'est-à-dire l'ensemble des protéines d'une cellule, organite, tissu, organe ou organisme à un moment donné et sous des conditions données.

Dans la pratique, la protéomique s'attache à identifier les protéines extraites d'une culture cellulaire, d'un tissu ou d'un fluide biologique, leur localisation dans les compartiments cellulaires, leurs modifications post-traductionnelles ainsi que leur quantité. Elle peut également permettre de quantifier les variations de leur taux d'expression en fonction du temps, de leur

environnement, de leur état de développement, de leur état physiologique et pathologique, de l'espèce d'origine. Elle étudie aussi les interactions que les protéines ont avec d'autres protéines, avec l'ADN ou l'ARN ainsi que les fonctions de chaque protéine. [8]

I.4.1 Les protéines

Les protéines représentent l'une des classes moléculaires les plus importantes dans les organismes vivants, leurs fonctions incluent la catalyse de processus métaboliques sous la forme d'enzymes, elles jouent un rôle important dans la transmission du signal, les mécanismes de défense, de transport de molécules, elles sont utilisées comme matériaux de construction par exemple dans les cheveux (la protéine de la kératine), chaque protéine est une macromolécule produite par un organisme vivant, les protéines sont formées de chaînes d'acides aminés liées par des liaisons peptidiques.

Elles sont généralement représentées sous forme de séquences, elles se replient en structures tridimensionnelles plus ou moins stables. (voir Figure I.4), les protéines ont des tailles de plusieurs centaines d'acides aminés plus spécifiquement les petites chaînes sont appelées peptides les protéines étant des polypeptides pouvant être réunies par des ponts disulfures.

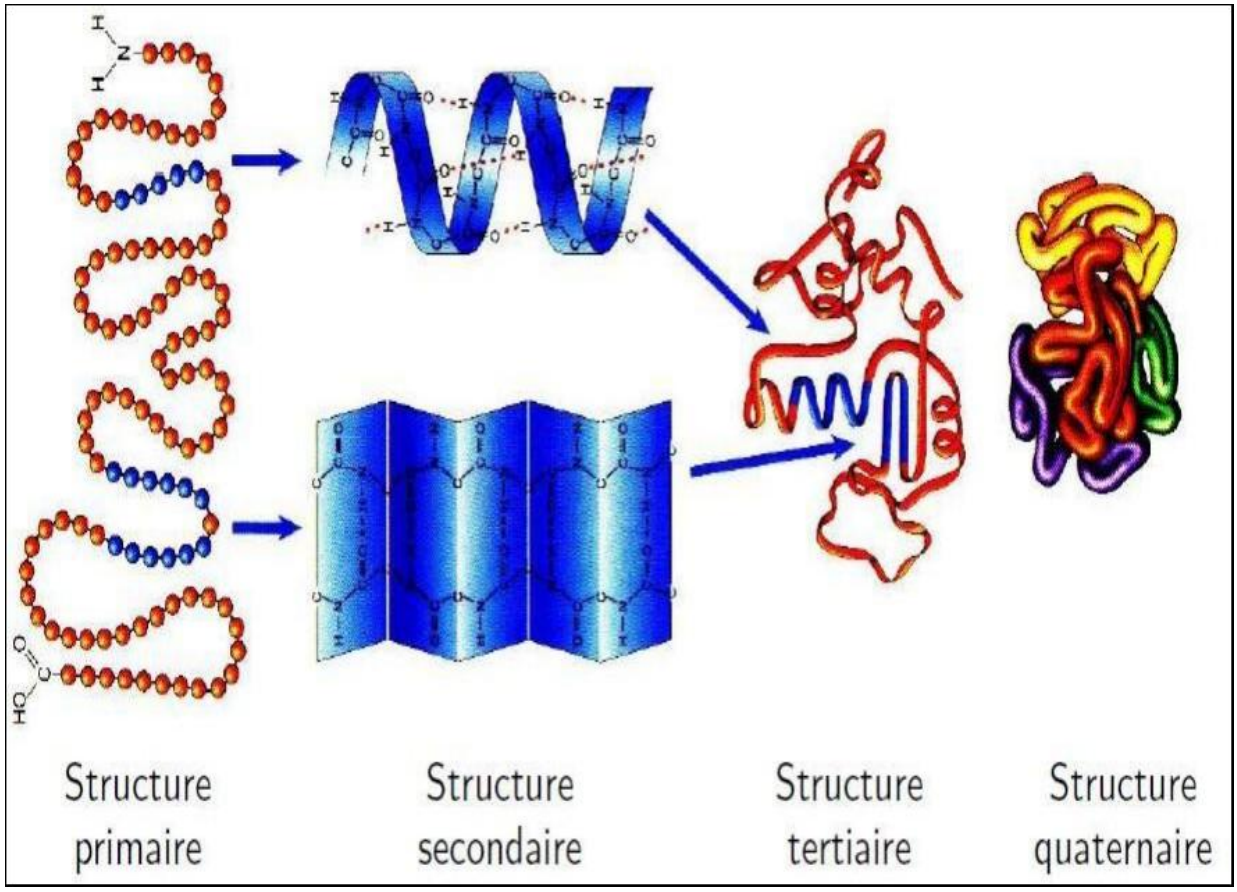


Figure I.4 : Les différentes structures d'une protéine.

Les protéines se répartissent en quatre classes générales sur la base de leur structure :

1. Structure primaire : La structure primaire décrit l'ordre unique dans lequel les acides aminés sont liés pour former une protéine.

2. Structure secondaire désigne l'enroulement ou le repliement d'une chaîne polypeptidique qui donne à la protéine sa forme tridimensionnelle. Il existe deux types de structures secondaires observées dans les protéines.

- Un type est la structure d'hélice alpha (α), cette structure ressemble à un ressort enroulé qui est fixée par liaison hydrogène dans la chaîne polypeptidique.
- Le second type de structure secondaire dans les protéines est la feuille plissée bêta (β).

Cette structure semble être pliée ou plissée qui est maintenue ensemble par liaison hydrogène entre des unités polypeptidiques de la chaîne pliée qui sont adjacentes les unes aux autres.

3. La structure tertiaire fait référence à la structure 3D complète de la chaîne polypeptidique d'une protéine. Il existe plusieurs types de liaisons et de forces qui retiennent une protéine dans sa structure tertiaire.

- Interactions hydrophobes.
- Liaison hydrogène.
- Une liaison ionique.
- Pont disulfure.

4. Structure quaternaire La structure quaternaire fait référence à la structure d'une macromolécule de protéine formée par des interactions entre plusieurs chaînes polypeptidiques, chaque chaîne polypeptidique est appelée sous-unité.

Les protéines à structure quaternaire peuvent être constituées de plusieurs sous-unités du même type, ils peuvent également être composés de différentes sous-unités, l'hémoglobine est un exemple de protéine à structure quaternaire. [9]

I.4.2 Les acides aminés

Un acide aminé est une petite molécule élémentaire des protéines.

Il existe 20 formes différentes synthétisées par voie ribosomale dans le monde du vivant. [10]

Code en une lettre	Abréviation	Nom
A	Ala.	Alanine
R	Arg.	Arginine
N	Asn.	Asparagine
D	Asp.	Acideaspartique
C	Cys.	Cystéine
Q	Gln.	Glutamine
E	Glu.	Acideglutamique
G	Gly.	Glycine
H	His.	Histidine
I	Ile.	Isoleucine
L	Len.	Leucine
K	Lys.	Lysine
M	Met.	Méthionine
F	Phe.	Phénylalanine
P	Pro.	Proline
S	Ser.	Sérine
T	Thr.	Thréonine
W	Trp.	Tryptophane
Y	Tyr.	Tyrosine
V	Val.	Valine

Tableau I.1 : Les vingt acides aminés natifs et leur code officiel.

La configuration générale des acides aminés naturels est caractérisée par un groupe amine et un groupe carboxyle autour d'un atome de carbone central Cα , la chaîne latérale respective de chaque acide aminé détermine les propriétés chimiques telles que l'hydrophobie, la polarité, ou l'acidité . (Figure I.5)

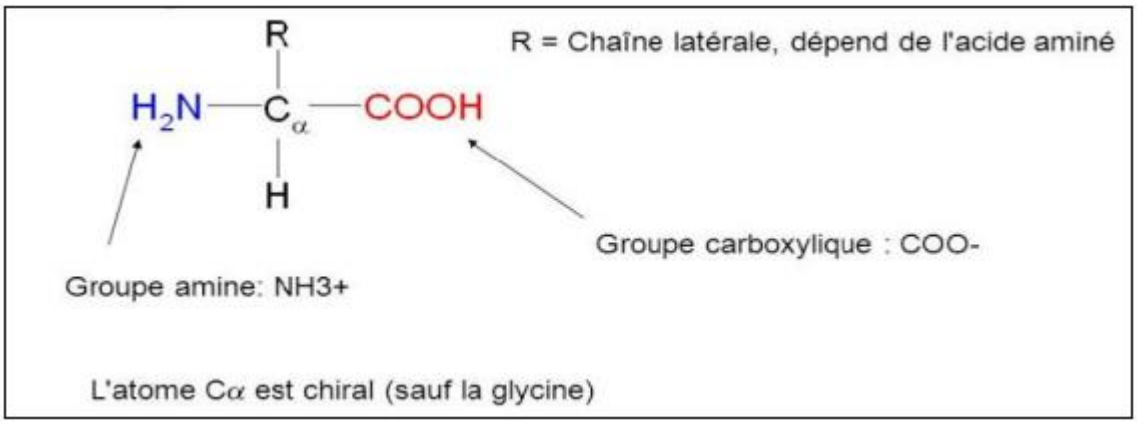


Figure I.5 : La formule générale d'un acide aminé.

Des liaisons peptidiques connectent des acides aminés individuels dans une chaîne polypeptidique, chaque acide aminé est lié via la liaison amide de l'acide de son groupe α carboxylique au groupe amine α de l'autre. [11] .par conséquent, ils ont des extrémités N- et C- terminales.

La structure primaire polypeptidique, c'est-à-dire, la séquence d'acides aminés de l'extrémité N- à l'extrémité C-terminale, peut contenir entre trois et plusieurs centaines d'acides aminés. Chaque acide aminé dans la chaîne polypeptidique est abrégé soit par un code à trois lettres ou une lettre. (Voir Tableau I.1). (Figure I.6)

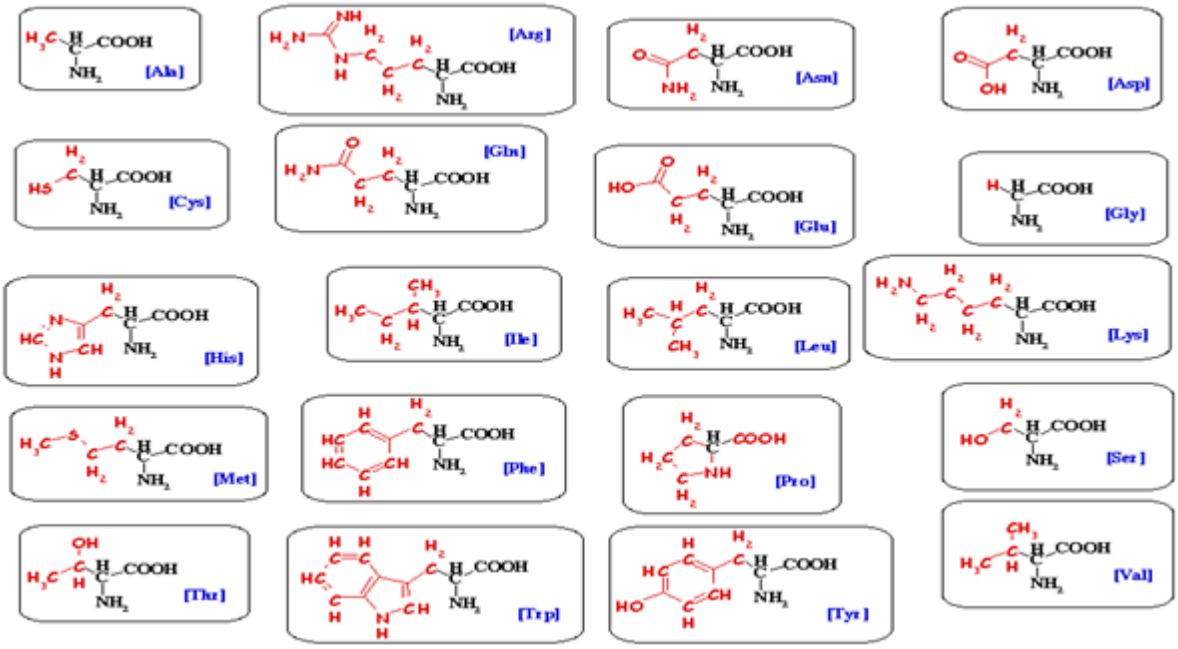


Figure I.6 : Les 20 formes différentes de L'acide aminé.

Les propriétés structurales et physico-chimiques de chaque acide aminé sont très variées Cependant en se basant sur la composition chimique on peut regrouper les acides aminés en 8 familles :

1.Les acides aminés aliphatiques dont le radical est une chaîne hydrogénons carbonée apolaire.

2. Les acides aminés hydroxylés qui portent un groupe alcool sont polaires mais non chargés et neutres.

3. Les acides aminés représentés uniquement par la proline, La chaîne latérale est repliée et établit une liaison covalente avec l'atome d'azote du groupement amine.

4. Les acides aminés soufrés qui comportent un atome de soufre dans la chaîne latérale l'un d'eux la cystéine est un thiol, deux molécules de cystéine peuvent établir une liaison covalente entre leurs atomes de soufre et établir une liaison supplémentaire dans la chaîne protéique.

5. Les acides di carboxyliques portent un groupement acide organique à l'extrémité de leur chaîne latérale sont donc polaires chargés négativement (à pH neutre) et également acides.

6. Les acides amidés il s'agit des versions amidées des acides aminés du groupe précédent, le groupement OH de l'acide carboxylique est remplacé par un groupement NH3.

7. Les acides diaminés La chaîne latérale porte un groupement aminé sont donc polaires chargés positivement (à pH neutre) et basiques.

8. Les acides aminés aromatiques comportent un cycle aromatique dans leur chaîne latérale, ces molécules sont non chargées et fortement polaires. [11]. (Figure I.7)

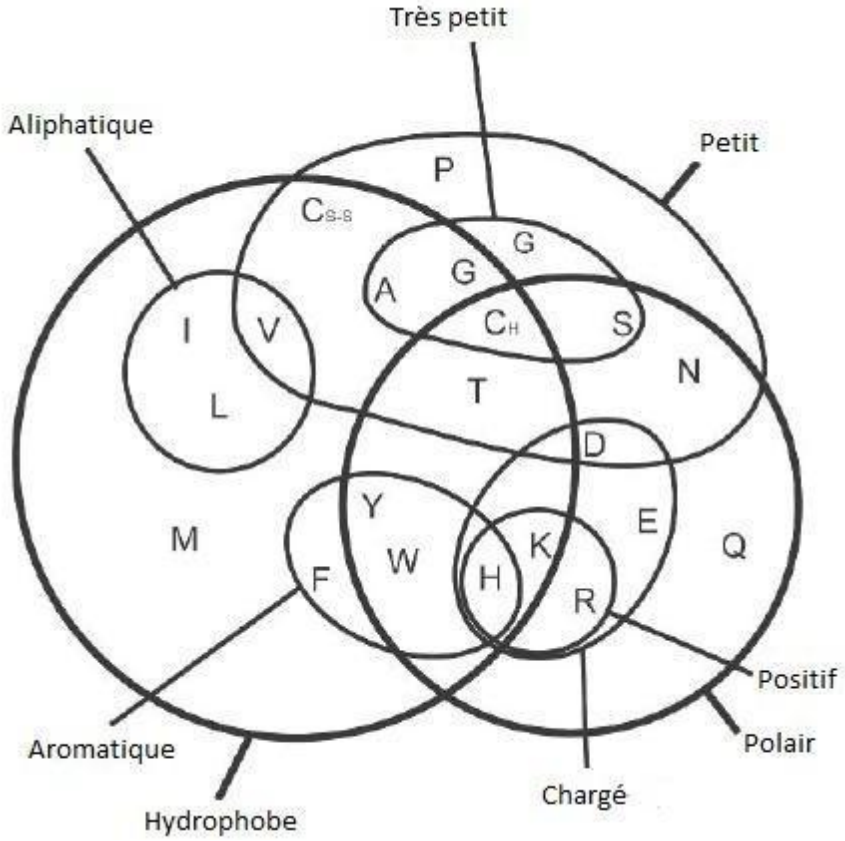


Figure I.7 : Diagramme de Venn des propriétés des acides aminés.

I.4.3 De la génomique vers la protéomique

Comprendre le fonctionnement d'une cellule vivante suppose celle des mécanismes moléculaires complexes qui sous-entendent les diverses activités cellulaires.

Tous les gènes d'un organisme ou son génome constituent une base de données statique et spécifique de l'être vivant à partir d'un génome unique, chaque type cellulaire d'un organisme exprimera un ensemble de protéines ou

protéome qui variera en fonction de l'environnement des cellules. La synthèse des protéines comprend deux étapes

- La transcription permet de copier l'ADN en ARN messager (ARNm) elle se déroule dans le noyau
- La traduction correspond au décodage de l'information portée par l'ARNm en polypeptides reliés en protéines

La génomique et la protéomique sont intrinsèquement « globales » dans le sens où des centaines si ce n'est des milliers de bases de données, de bases de connaissances, de programmes informatiques et de bibliothèques de documents sont disponibles via Internet et sont utilisés par des chercheurs et des développeurs à travers le monde dans le cadre de leurs travaux. [12]

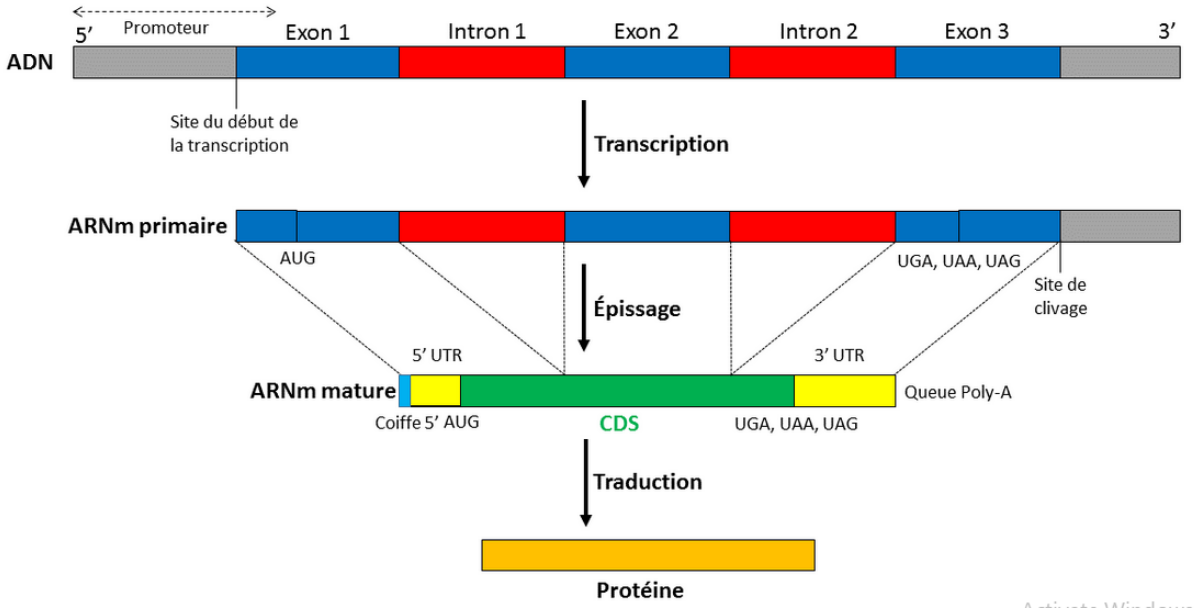


Figure I.8 : Représentation de la génomique vers la protéomique**.**

I.5 Prédiction des fonctions de protéines

La prédiction de la fonction des protéines est un champ de recherche à part entière et sert de point de départ aux analyses expérimentales pour élucider les activités biochimiques et les processus dans lesquels sont impliquées les protéines.

Les méthodes de prédiction de fonction ont évoluées au cours du temps notamment avec l'arrivée de la génomique l'apparition de données obtenues à grande échelle et les développements des outils bioinformatique.

La prédiction de fonction repose sur les connaissances préexistantes, ces connaissances proviennent d'études expérimentales d'identification des fonctions, utilisation des approches de génétique, de biologie moléculaire, de biochimie et maintenant de bioinformatique.

Les approches développées avant l'apparition des données obtenues à grande échelle (génomes, protéomes, transcriptome, interactomes, etc...) sont basées sur les données de séquence et de structure.

Les approches basées sur la séquence nécessitent de classer les protéines en familles en utilisant par exemple des outils de détection de similarité comme BLAST.

L'approche la plus simple est la recherche de séquences similaires à cette protéine dont la fonction est connue dans les bases de données, les annotations de fonctions des protéines peuvent bénéficier de connaissances sur les processus protéique, l'utilisation de l'homologie de séquence seule n'est pas une bonne approche pour dériver cette connaissance l'alternative est d'utiliser les motifs qui sont très discriminants pour prédire les fonctions des protéines.

Par ailleurs il existe d'autres approches que nous énumérons tel que :

- L'approche basée sur l'interaction protéine- protéine.
- L'approche génomique basée sur le contexte. [13]

I.5.1 Classification des protéines suivant leurs fonctions

Le concept de fonction protéique est très sensible au contexte et pas très bien défini en fait ce concept agit généralement comme un terme générique pour tous les types d'activités dans lesquelles une protéine est impliquée qu'elles soient cellulaires, moléculaires ou physiologiques. Une telle catégorisation des types de fonctions qu'une protéine peut remplir. [14]

- **Fonction moléculaire.**
- **Fonction phénotypique.**
- **Fonction cellulaire**.

I.5.1.1 Les fonctions moléculaires

Les fonctions biochimiques assurées par une protéine telles que la liaison de ligand, la catalyse de réactions biochimiques et les changements de conformation. [16]

- **Liaison ou fixation :** Cette fonction fondamentale qui est à l'origine de toutes les autres fonctions biochimiques à la capacité de fixer une autre molécule ou macromolécule appelée alors ligand tandis que la protéine est qualifiée de récepteur.
- **Catalyse :** Certaines protéines sont des catalyseurs de réactions chimiques elles permettent une réaction rapide aux conditions de température et de pression conformes à la vie, ces protéines sont alors appelées enzymes.
- **Chaperon :** La fonction principale des protéines chaperons est de faciliter le repliement des protéines et de permettre leur désagrégation.
- **les protéines régulatrices :** qui modulent l'activité d'autres protéines.
- **les protéines de signalisation :** qui captent les signaux extérieurs et assurent leur transmission dans la cellule ou l'organisme.
- **les protéines motrices :** permettant aux cellules ou organismes de se mouvoir.
- **Communication :** Cette fonction consiste à modifier le comportement d'autres cellules en fonction de l'environnement.

I.5.1.2 Les fonctions phénotypiques

L'intégration des sous-systèmes physiologiques composés de diverses protéines remplissant leurs fonctions cellulaires, l'interaction de ce système intégré avec des stimuli environnementaux déterminant les propriétés phénotypiques et le comportement de l'organisme.

I.5.1.3 Les fonctions cellulaires

De nombreuses protéines s'associent pour réaliser des fonctions physiologiques complexes telles que le fonctionnement des voies métaboliques et la transduction du signal afin de maintenir le fonctionnement des divers composants de l'organisme.

- **les protéines des structures :** qui permettent à la cellule de maintenir son organisation dans l'espace.

- **les protéines de transport :** qui assurent le transfert des différentes molécules dans et en dehors des cellules. [15]

I.5.2 Les concepts et techniques de la bioinformatique

La tâche majeure de la bioinformatique est de permettre d'identifier les fonctions d'un gène ou d'une protéine à partir de données existantes puisque les données sont variées, incomplètes, bruyantes et couvrent une variété d'organismes, il y a un recours constant aux principes biologiques afin de filtrer les informations utiles Il y a différentes techniques qui conduisent à une meilleure compréhension de la fonction des gènes et des protéines telles que:

- **La construction évolutive d'arbre phylogénétique :** ces arbres sont souvent construits après comparaison de séquences appartenant à différents organismes d'une même espèce, les arbres regroupent les séquences selon leur degré de similitude, ils servent de guide pour le raisonnement sur la façon dont les séquences ont été transformées au courant de l'évolution par exemple ils déduisent l'homologie de la similitude et peuvent écarter des hypothèses erronées qui sont en contradiction avec le processus connu de l'évolution.
- **Détection de motifs dans les séquences :** il y a certaines parties de séquences de nucléotides et des séquences d'acides aminés qui doivent être détectées.

Il y a deux exemples principaux qui sont la recherche de gènes dans l'ADN et la détermination des sous-composants de séquences d'acides aminés (structure secondaire).

- **Déterminer des structures 3D à partir de séquences :** les problèmes en bioinformatique qui se rapportent aux structures tridimensionnelles impliquent des calculs difficiles à réaliser, la détermination de la forme d'ARN à partir de séquences nécessite des algorithmes de complexité cubique, l'inférence des formes de protéines à partir de séquences d'acides aminés reste à ce jour un problème non résolu.
- **Déduction de la régulation cellulaire :** la fonction d'un gène ou d'une protéine est mieux décrite par son rôle, les gènes interagissent les uns avec les autres les protéines peuvent également prévenir ou aider à la production d'autres protéines, les modèles disponibles de la régulation cellulaire peuvent être discrets ou continus il y a habituellement une distinction entre la simulation et la modélisation cellulaire.

- **Déterminer la fonction de protéine et les voies métaboliques :** c'est l'un des domaines les plus difficiles de la bioinformatique pour lequel il n'y a pas beaucoup de données disponibles, l'objectif ici est d'interpréter les annotations humaines pour la fonction des protéines et également de développer des bases de données représentant des graphiques qui peuvent être interrogés pour l'existence de nœuds (les réactions à préciser) et les chemins (en précisant les séquences de réactions).
- **Assembler les fragments d'ADN :** les fragments fournis par séquençage sont assemblées à l'aide d'ordinateurs, la partie la plus délicate de cet assemblage est que l'ADN à de nombreuses régions répétitives et les mêmes fragments peuvent appartenir à différentes régions, les algorithmes d'assemblage de l'ADN sont surtout utilisés par les grandes entreprises. [16]

I.5.3 Les bases de données biologiques

Le concept le plus important pour la bioinformatique appliquée est la collecte de données de séquence et son information biologique associée par exemple le projet de la prédiction de fonction des protéines génèrent quotidiennement une très grande quantité de données dans le monde entier pour utiliser ces données de façon appropriée un système de dépôt structuré de ces données est nécessaire mais les données devraient également être accessibles aux personnes intéressées.

De ce fait, des bases de données en ligne ont vu le jour, elles sont accessibles gratuitement à tous, ce qui représente un avantage considérable.

Les bases de données biologiques sont disponibles selon le type de données qu'elles contiennent : Deux catégories de bases de données biologiques peuvent être distinguées

- Les bases de données primaires qui contiennent des informations de séquences primaires (de nucléotides ou de protéines)
- Des informations d'annotation relatives à la fonction contrairement aux bases de données secondaires qui résument les résultats de l'analyse des bases de données de séquences de protéines primaire.

L'objectif de ces analyses est de tirer des caractéristiques communes pour les classes de Séquences qui peuvent à leur tour être utilisés pour la classification des séquences inconnues (annotation).

Les bases de données primaires contiennent différentes bases de données protéiques et nucléiques.

Il y a quatre grandes bases de données nucléiques très connues à travers le monde :

1. la base de données GenBank du centre national d'information technologique américain.

2. (National Center of Biotechnology Information – NCBI), qui contient plus de 76 million de séquences nucléiques.

3. La base de données EMBL (European Molecular Biology Laboratory) de l'institut bioinformatique européen (European Bioinfomatics Institute – EBI).

4. La base de données DDBJ (DNA Data Bank of Japan) du centre pour l'information biologique du Japon (Center for Information Biology – CIB). Ces trois organismes : NCBI, EBI et CIB comprennent des bases de données nucléiques internationales et synchronisent leurs bases de données toutes les24h. [17]

I.6 Les récepteurs couplés aux protéines g

Parmi les grandes familles des protéines il est important de mettre en relief les RCPG qui sont retrouvées dans tous types de tissus pour cela notre travail sera axée essentiellement sur les récepteurs couplés aux protéines G constitue l'une des familles la plus importante et la plus étudiée, les RCPG partagent tous une unité fonctionnelle commune (communication cellulaire) qui forme sept domaines transmembranaires en forme d'hélices cependant les RCPG comportent également des domaines fonctionnels variés notamment au niveau de leur extrémité extracellulaire (extrémité N-terminale) et intracellulaire (extrémité C-terminale), ces récepteurs sont retrouvés dans tous types de tissus et sont impliqués dans la plupart des processus physiologiques et physiopathologiques ce qui en fait une cible de choix pour la découverte de médicaments.

En effet à l'heure actuelle on estime qu'environ 30% des médicaments commercialisés ciblent des RCPG cependant ces médicaments ne ciblent qu‟une trentaine de RCPG ce qui laisse un potentiel important de découverte de nouveaux médicaments agissant sur les autres RCPG. [18]

I.6.1 Définition

Récepteur couplé à la protéine G (RCPG) également appelé récepteur à sept transmembranes ou récepteur heptahélique, protéine située dans la membrane cellulaire qui se lie aux substances extracellulaires et transmet les signaux de ces substances à une molécule intracellulaire appelée protéine G (protéine de liaison à la guanine), les RCPG se trouvent dans les membranes cellulaires d'un large éventail d'organismes y compris les mammifères, les plantes, les microorganismes et les invertébrés. Il existe de nombreux types de RCPG- environ 1 000 types sont codés par le génome humain seul en tant que groupe, ils répondent à un large éventail de substances notamment la lumière, les hormones, les amines, les neurotransmetteurs et les lipides, certains exemples de RCPG incluent les récepteurs bêta-adrénergiques qui se lient à l'épinéphrine, les récepteurs des prostaglandines E2 qui lient des substances inflammatoires appelées prostaglandines et la rhodopsine qui contient un produit chimique photo réactif appelé rétinal qui répond aux signaux lumineux reçus par les cellules en tige de l'oeil.

L'existence de RCPG a été démontrée dans les années 1970 par le médecin et biologiste moléculaire américain Robert J. Lefkowitz. Lefkowitzqui a partagé le prix Nobel de chimie 2012 avec son collègue Brian K. Kobilka, qui a contribué à élucider la structure et les fonctions du RCPG. [19]

I.6.2 Classification RCPG

GPCR classé en cinq familles selon le système GRAFS: Metabotrope-glutamate, Rhodopsin-like, AMP cyclique, Frizzled/smoothened, fungalpheromone et Secretin-like. Dans ce chapitre nous ne détaillerons que la classification de Kolakowski, du fait de son utilisation prédominante.

Celle-ci attribue une lettre à chaque famille de RCPG, A pour Rhodopsin-like, B pour Secretin-like, C pour Metabotrope-glutamate/phéromone, D pour fungalpheromone, E pour AMP cyclique, et enfin F pour Frizzled/smoothened.

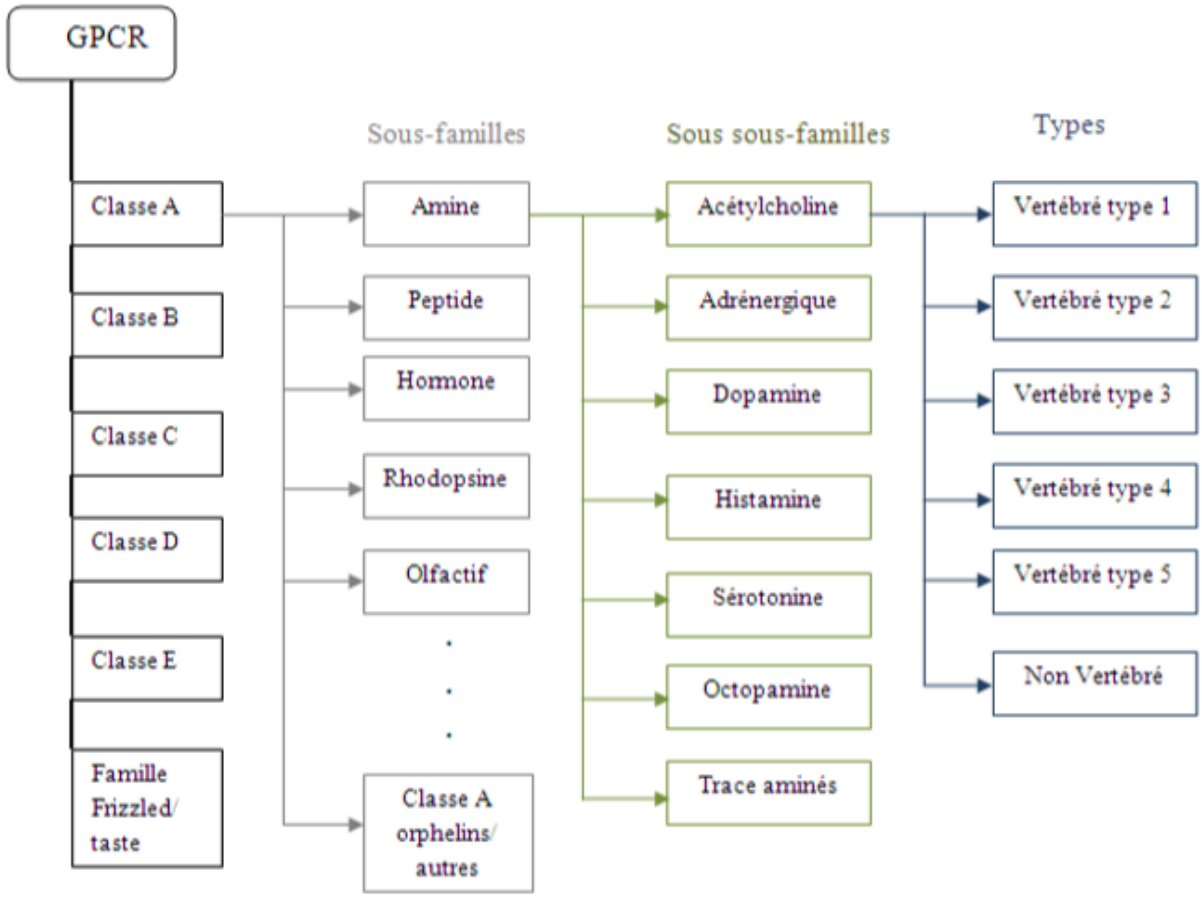


Figure I.9 : Les familles de récepteurs couplés aux protéines G.

La famille A : « Rhodopsin-like »

La famille A : est la mieux caractérisée et celle-ci comporte le plus grand nombre de récepteurs avec plus de 80% des RCPG elle inclut la rhodopsine et le récepteur 2β-adrénergique, cette classe de récepteurs lie particulièrement les hormones, les polypeptides de petite et grande taille, les amines biogènes, les photons ou encore les substances apparentées aux lipides. [20]

La famille B : « Secretin-Like »

La famille B : Cette famille des RCPG ne présente qu'une faible homologie avec la classe A puisque les membres de ces classes partagent seulement 12% d"homologie de séquence, cette famille regroupe des récepteurs structurellement proches du récepteur de la sécrétine. [21]

La famille C : « Metabotrope-glutamate/pheromone »

La famille C : Cette famille des RCPG regroupe les récepteurs métabotropiques du glutamate, les récepteurs au calcium, les récepteurs GABAB, les récepteurs du goût, les RAIG (retinoic acid-inducible orphan GPCR) et les récepteurs aux phéromones.

Les autres familles : D, E et F

La famille D : Cette famille est constituée de récepteurs aux phéromones principalement exprimée dans l'organe voméronasal, avec la protéine G de type Gi2, les récepteurs aux phéromones ou OR (olfactory receptor) comportent une soixantaine de gènes chez les insectes.[22]

La famille E : Cette famille est formée par les quatre récepteurs à l'AMPc (cAR 1–4) et d‟autres récepteurs « cAMP receptor-like » caractérisés chez Dictyostelium discoideum (moisissures).

La famille F : Cette famille de récepteurs rassemble des récepteurs homologues aux protéines « frizzled » et « smoothened », les récepteurs de type « frizzled » jouent un rôle important dans le développement embryonnaire et plus particulièrement dans le contrôle de la prolifération et la polarité cellulaire.

I.6.3 Structure des RCPGs

Bien que les récepteurs couplés aux protéines G soient très nombreux et variés ils possèdent toute une topologie commune (**Figure I.10**). En effet une analyse détaillée des séquences de RCPG a révélé l‟existence d‟une structure membranaire analogue : sept hélices transmembranaires et hydrophobes particulièrement bien conservées et composées chacune d'environ 25 à 35 acides aminés.

Ces domaines transmembranaires (souvent noté TM 1 à 7) sont reliés entre eux par trois boucles intracellulaires (I1, I2 et I3) et trois boucles extracellulaires (nommées E1, E2 et E3) de taille variables, l'extrémité N-terminal est localisée du côté extracellulaire tandis que l'extrémité C-terminal est située dans la région intracellulaire, deux cystéines sont également conservées dans la plupart des RCPG pouvant former un pont disulfure entre la première et la deuxième boucle extracellulaire, certaines extrémités carboxy-terminales possèdent un site possible de palmitoylation qui peut former un ancrage lipidique dans la membrane (on parle d'hélice 8, H8) et créer une quatrième boucle intracellulaire. [23]

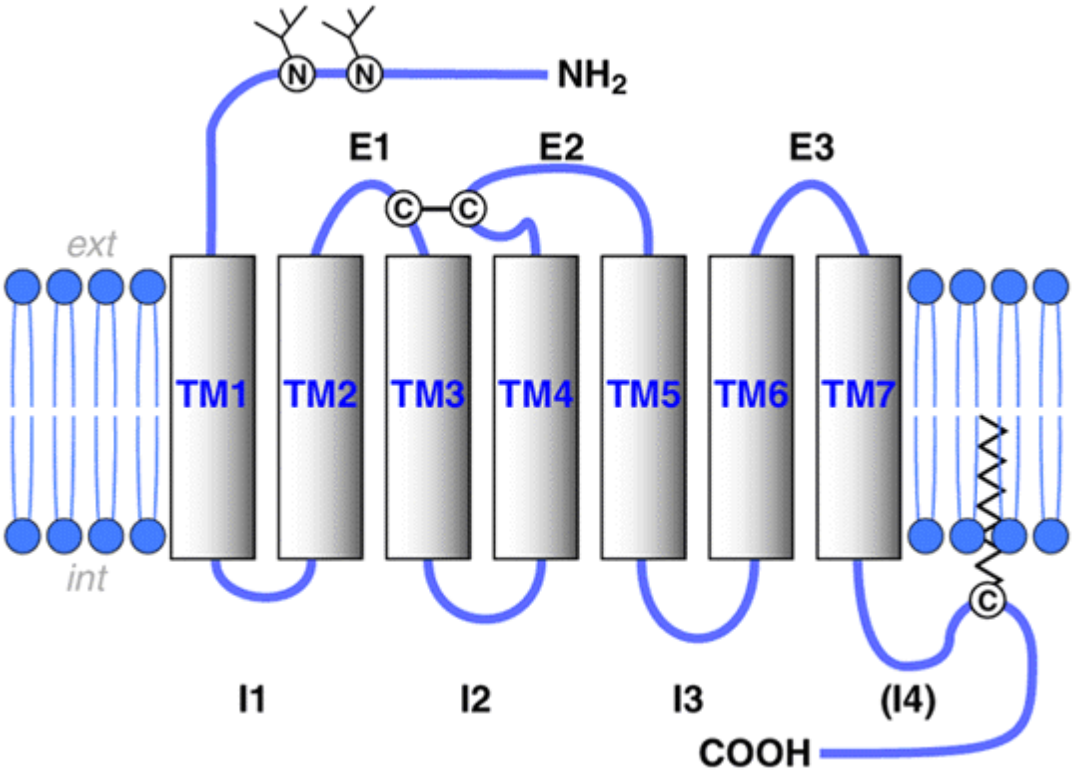


Figure I.10 : Structure de la majorité des récepteurs couplés aux protéines G.

I.6.4 Protéines G

Les protéines G sont des protéines hétérotrimériques composées de sous-unités α, β et γ. Lorsque la protéine G est activée par le récepteur elle se dissocie en deux parties indépendantes :

La sous-unité et le complexe qui vont l'un et l'autre interagir avec des protéines effectrices intracellulaires afin de moduler leur fonction. [24]

I.6.5 Transduction du signal par les protéines G

Une fois le ligand lié à son récepteur, le RCPG engage une cascade d'activation protéique à l'intérieur de la cellule qui aboutit le plus souvent à la transcription de gènes, le récepteur activé subit des changements conformationnels qui par extension active la protéine G couplé au récepteur, la protéine G activée peut alors initier la transduction du signal intracellulaire induite par le récepteur. [24]

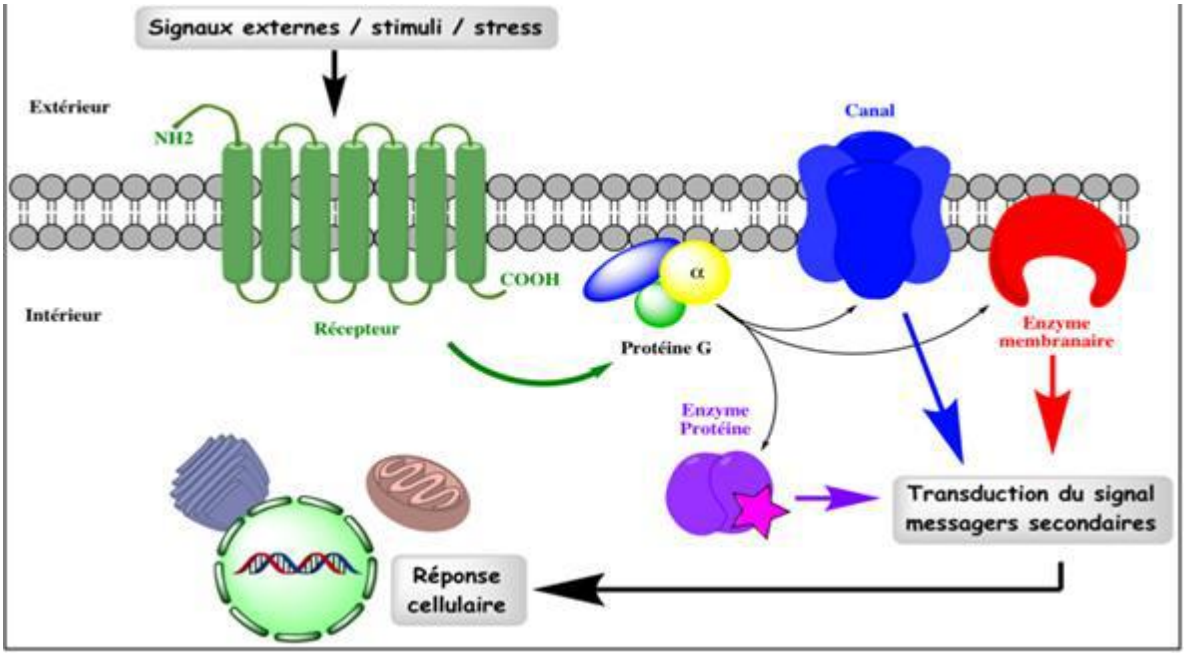


Figure I.11 : Transmission des signaux par les récepteurs couplés aux protéines G.

I.6.6 La base de données RCPGs

RCPGdb contient des données, des diagrammes et des outils Web pour les récepteurs couplés aux protéines G (RCPG), Les utilisateurs peuvent parcourir toutes les structures cristallines de RCPG et les plus grandes collections de mutants de récepteurs, des diagrammes peuvent être produits et téléchargés pour illustrer les résidus de récepteurs (diagrammes de tracé de serpent et de boîtes à hélices) et les relations (arbres phylogénétiques).

Les séquences basées sur la structure de référence (cristalline) tiennent compte des renflements et des constrictions de l'hélice, affichent des statistiques sur la conservation des acides aminés et se sont vu attribuer une numérotation résiduelle générique pour les résidus équivalents dans différents récepteurs.

La base de données GPCRDB qui reste à ce jour la base de données la plus utilisée dans le domaine d'identification des RCPGs. [25]

I.7 Objectif de l'étude de la fonction des protéines

Le rôle des protéines va au-delà régulation des processus cellulaires, les protéines étant directement impliqué dans le maintien de la santé d'un organisme. En fait de nombreuses maladies peuvent être retracées à protéines dysfonctionnelles, par exemple l'insuline produite dans le pancréas est une

hormone, cela aide à réguler le métabolisme du sucre dans le corps, une protéine d'insuline qui ne fonctionne pas cause un manquant qui cause le diabète.

L'étude de la fonction des protéines est donc importante car nous permet de mieux comprendre comment les protéines affectent la santé et la maladie.

En outre la compréhension de la fonction des protéines contribue également au processus de concevoir de nouveaux médicaments et options de traitement, par exemple la découverte de la protéine tumeur-nécrosis factor-alpha (TNF-alpha) joue un rôle fonctionnel dans la réponse inflammatoire a finalement conduit au développement de médicaments ciblant le TNF-alpha pour le traitement de l'arthrite. [26]

Les protéines jouent un rôle aussi important dans la santé humaine des efforts sont en cours pour découvrir et annoter la fonction des protéines séquencées, la fonction des protéines peut être déterminée en utilisant à la fois des méthodes expérimentales et des systèmes informatiques.

I.8 Conclusion

La bioinformatique est un domaine de recherche en plein essor beaucoup d'avancées sont réalisées du fait qu'elle combine diverses disciplines comme l'informatique, la biologie, les mathématiques, les statistiques, etc.…

Dans ce chapitre nous avons présenté la discipline protéomique qui est le principe de notre étude qui consiste à prédire la fonction de protéines à partir de leurs structures ainsi qu'à partir des interactions protéine-protéine, il est donc important de résoudre ces problèmes pour mieux pouvoir identifier les protéines.

A cet effet beaucoup de méthodes et techniques ont été développées afin d'y apporter des solutions efficaces tout en exploitant les tâches de la fouille de données telles que la classification que nous allons voir dans le chapitre suivant.

Chapitre II :
Classification

II.1 Introduction

Le champ le plus important dans le domaine de la bioinformatique est l'application de la fouille de données pour la résolution de problèmes biologiques. Il existe deux types d'approches la classification (apprentissage supervisé), qui comprend également la découverte des règles de classification et le clustering (apprentissage non-supervisé).

La classification fait partie de la technique de fouille des données qui est une étape importante dans l'analyse des données et consiste à regrouper les données en classes homogènes.

Nous avons présenté dans ce chapitre tout d'abord la signification de la Fouille des données, les techniques utilisées, définition de la classification, les méthodes de la classification à la fin les domaines d'applications de la classification.

II.2 Fouille de données

La fouille de données est un domaine qui est apparu avec l'explosion des quantités d'informations stockées, avec le progrès important des vitesses de traitement et des supports de stockage. La fouille de données vise à découvrir, dans les grandes quantités de données, les informations précieuses qui peuvent aider à comprendre les données ou à prédire le comportement des données futures.

Fouille de données est le processus d'analyse des modèles de données cachées selon différentes perspectives de catégorisation en informations utiles qui sont collectées et assemblées dans des zones communes tel que les entrepôts de données pour une analyse efficace des algorithmes d'exploration des données facilitant la prise de décision commerciale et autres informations, il est impératif de réduire les coûts et d'augmenter les revenus.

La fouille de données est souvent définie comme étant le processus de découverte des nouvelles connaissances en examinant de larges quantités de données (stockées dans des entrepôts) en utilisant les technologies de reconnaissance de formes de même que les techniques statistiques et mathématiques.

La fouille de données vise à exploiter ces données pour extraire des modèles en estimant les relations entre les variables (entrées et sorties) de ses systèmes. [27]

II.2.1 Les techniques de fouille des données

Les techniques principales d'exploration des données :

- **Classification:** La classification c'est la technique d'exploration des données la plus couramment utilisée, elle contient l'ensemble d'échantillons pré-classés afin de créer un modèle permettant de classer le grand ensemble de données.
- **Clustering:** Le clustering c'est l'une des techniques les plus anciennes utilisées dans l'exploration des données, et lui aussi, c'est d'identification de données similaires qui sont identiques les unes aux autres.
- **La regression:** c'est la méthode la plus aisée utilisée pour estimer les valeurs continues, son objectif c'est de trouver le meilleur modèle qui décrit la relation entre une variable de sortie continue et une ou plusieurs variables d'entrée.
- **L'apprentissage des règles d'association :** c'est une méthode d'apprentissage automatique basée sur des règles permettant de découvrir les relations intéressantes entre des variables parmi les grandes bases de données. [28]

II.3 Classification

Est une tâche très importante dans la fouille des données, et qui consomme beaucoup de recherches pour son optimisation. La classification est l'une des techniques les plus utilisées dans l'analyse des bases de données. Elle permet d'apprendre des modèles de décision qui permettent de prédire le comportement des exemples futurs. La classification consiste à inférer à partir d'un échantillon d'exemples classés une procédure de classification. Un système d'apprentissage effectue la recherche d'une telle procédure selon un modèle. [27]

II.3.1 Classification supervisée et non supervisée

Le processus de classification peut se découper en deux phases la première est une phase d'apprentissage pendant laquelle l'algorithme cherche des règles de classification (au sens large), la seconde phase consiste à appliquer les règles de la classification découvertes à un ensemble d'objets afin d'identifier la classe d'appartenance de chacun des objets.

- **La classification supervisée:** appartiennent à la modélisation prédictive dans l'apprentissage supervisé, les données d'apprentissage (observations, mesures) sont accompagnées d'étiquettes indiquant la classe de l'observation, les nouvelles données sont classées en fonction de l'ensemble de formation.

- **La classification non supervisée:** les techniques non supervisées appartiennent à la modélisation descriptive, dans l'apprentissage non supervisé (clustering) la classe de données d'apprentissage est inconnue, nous cherchons à trouver des classes ou des groupes d'observations ou de mesures. [29]

II.4 La classification non supervisée

II.4.1 Clustering

Le clustering regroupe un ensemble de techniques qui visent à regrouper les enregistrements

d'une base de données en des groupes selon leur rapprochement les uns des autres

en ne se basant sur aucune information antérieure.

Le Clustering est un regroupement en classes homogènes consistant à représenter un nuage des points d'un espace quelconque en un ensemble de groupes appelé Cluster.

Un système d'analyse en clusters prend en entrée un ensemble de données et une mesure de similarité entre ces données, et produit en sortie un ensemble de partitions décrivant la structure générale de l'ensemble de données.

Les types de Clustering : Il existe deux grands types de clustering :

- Le clustering hiérarchique.
- Le clustering non-hiérarchique (par partitionnement)

II.4.1.1 Le clustering hiérarchique

Dans ce type de clustering le nombre de clusters ne peut être connu à l'avance. Quant aux méthodes hiérarchiques elles sont ascendantes ou descendantes, la classification hiérarchique consiste à construire un arbre de classes appelé dendrogramme,

Il existe deux classes de ce type d'algorithmes :

Les algorithmes ascendants ou encore aggloméra tifs considèrent chaque objet de l'ensemble de données comme des classes initiales et à chaque étapes on fusionne deux classes qui optimisent un critère de similarité. Ainsi, en d'autres termes, chaque enregistrement comme étant un cluster indépendant puis rassemblent les plus proches en des clusters plus importants, et ainsi de suite jusqu'à atteindre un seul cluster contenant toutes les données.

A l'inverse La deuxième classe est celle des algorithmes descendantes ou divisibles qui commencent à partir d'un ensemble de données et le subdivisent en sous ensembles puis subdivisent chaque sous ensemble en d'autres plus petits, et ainsi de suite, pour générer en fin une séquence de clusters ordonnée du plus général au plus fin. [30]

II.4.1.2 Le clustering partitionnel

Les algorithmes de partitionnement de clustering construisent directement en sortie une partition de l'espace des objets en K classes. Le principe général par définition d'une partition cela signifie que chaque classe doit contenir au moins un objet et que chaque objet doit appartenir à une classe unique pour que cela se réalise le nombre de classes K est requise. Ces algorithmes génèrent une partition initiale puis cherchent à l'améliorer en réattribuant les objets d'une classe à une autre. [30]

II.4.2 Exemple des méthodes de classification non supervisée

II.4.2.1 K-means

L'algorithme des K-means est l'algorithme de clustering (ou partitionnement) le plus connu et le plus utilisé, tout en étant très efficace et simple l'objectif de cet algorithme est de subdiviser l'ensemble des éléments donnés en entrée sur un certain nombre de classes. Ces classes doivent être des intersections vides et chacune doit être représentée par un noyau.

L'algorithme des K-means regroupe les éléments en entrée par similitude afin d'obtenir les classes les plus homogènes possible ainsi nous pouvons dégager deux critères d'évaluation à savoir la séparation des classes et l'homogénéité de leurs éléments, cette dernière propriété constitue le critère de choix des classes que propose l'algorithme des K-means. [31]

II.4.2.2 K-médoides

Les méthodes des k-médoide se différencient de la méthode des k-moyennes par l'utilisation des médoides plutôt que des centroides pour représenter les classes. Dans l'algorithme de k-médoides une classe est représentée par un de ses objets prédominants, Le médoide d'un groupe est l'objet possédant la distance médiane la plus faible avec les autres objets du groupe c'est un algorithme itératif combinant la réaffectation des objets dans des classes avec une intervention des médoides et des autres objets. [32]

II.5 La classification supervisée

II.5.1 Objectif de classification supervisée

L'objectif de la classification est d'identifier les classes auxquelles appartiennent des objets à partir de leurs caractéristiques ou attributs des criptifs. Dans le cadre d'une classification supervisée c'est d'apprendre à l'aide d'un ensemble d'entraînements une procédure de classification qui permet de prédire l'appartenance d'un nouvel exemple à une classe. [33]

II.5.2 Définition de classification supervisée

Soit m un ensemble d'exemples de données étiquetées. S = {(x1, y1), ..., (xm, ym)}. Chaque donnée xi est caractérisée par m attributs et par sa classe yi. On cherche une hypothèse telle que:

- Hypothèse satisfait les échantillons $\forall i \in$ {1,…, m} hypothèse (xi) =yi
- Hypothèse contient de bonnes propriétés de généralisation.

Le problème de la classification consiste donc en s'appuyant sur l'ensemble d'exemples à prédire la classe de toute nouvelle donnée x$\in R$m. [33]

II.5.3 Méthodes de classification supervisée

De nombreux algorithmes d'apprentissage sont adaptés au problème de la classification supervisée on peut les citer :

Les Séparateurs à Vastes Marges (SVMs), les réseaux de neurones, les méthodes des k plus proches voisins, les arbres de décision…etc.

II.5.3.1 Arbre de décision

Les arbres de décision représentent une méthode très efficace d'apprentissage supervisé. Un arbre de décision est un modèle de classification présenté sous la forme graphique d'un arbre, Il s'agit de partitionner un ensemble de données en des groupes les plus homogènes possible du point de vue de la variable à prédire. On prend en entrée un ensemble de données classées, et on fournit en sortie un arbre qui ressemble beaucoup à un diagramme d'orientation où L'extrémité de chaque branche est une feuille qui présente le résultat obtenu en fonction des décisions prises à partir de la racine de l'arbre jusqu'à cette feuille.

Les feuilles intermédiaires sont appelées des noeuds, chaque nœud final (feuille) représente une décision (une classe) et chaque nœud non final (interne) représente un test, Chaque feuille représente la décision d'appartenance à une classe des données vérifiant tous les tests du chemin menant de la racine à cette feuille, Chaque noeud de l'arbre contient un test sur un attribut qui permet de distribuer les données dans les différents sous-arbres. Lors de la construction de l'arbre, un critère de pureté comme l'entropie (utilisé dans C4.5 ou Gini) est utilisé pour transformer une feuille en nœud.

L'objectif est de produire des groupes d'individus les plus homogènes possibles du point de vue de la variable à prédire, en prédiction un exemple à classer "parcourt" l'arbre depuis la racine jusqu'à une unique feuille, son trajet dans l'arbre est entièrement déterminé par les valeurs de ses attributs, il est alors affecté à la classe dominante de la feuille avec pour score la proportion d'individus dans la feuille qui appartiennent à cette classe. [34]

L'exemple suivant montre un ensemble de données avec quatre attributs : motivation, surprenant, durée, difficile et l'attribut à prédire Jouer.

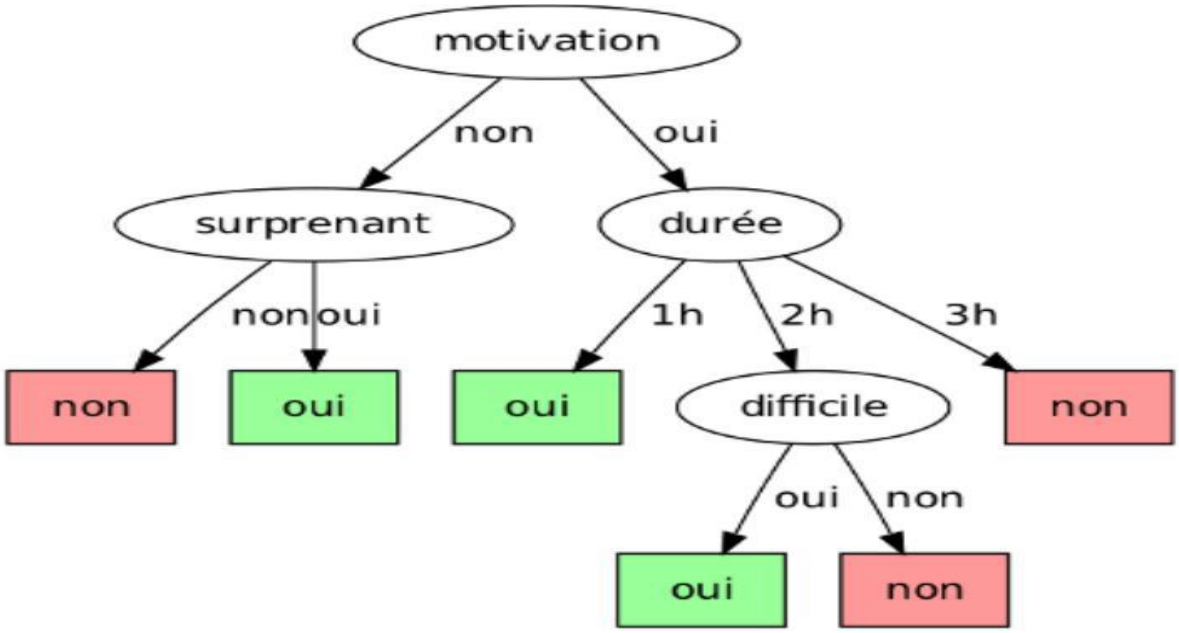


Figure II.1 : Exemple de classification avec l'arbre de décisions.

II.5.3.2 Séparateurs à Vaste Marge

Les Séparateurs à Vaste Marge ou Support Vector Machines (SVMs), les SVMs se situent sur l'axe de développement de la recherche humaine des techniques d'apprentissage. Les SVMs reposent sur une théorie mathématique solide à l'inverse des méthodes de réseaux de neurones. Elles ont été développées au sens inverse du développement des réseaux de neurones : ces derniers ont suivi un chemin heuristique de l'application et l'expérimentation vers la théorie ; alors que les SVMs sont venues de la théorie du son vers l'application.

Les SVMs sont des classifieurs qui reposent sur deux idées clés : La notion de marge maximale et la notion de fonction noyau, la première idée clé est la notion de marge maximale, on cherche l'hyperplan qui sépare les exemples positifs des exemples négatifs en garantissant que la distance entre la frontière de séparation et les échantillons les plus proches (marge) soit maximale, ces derniers sont appelés vecteurs supports. Afin de pouvoir traiter des cas où les données ne sont pas linéairement séparables la deuxième idée clé des SVM est de transformer l'espace de représentation des données d'entrées en un espace de plus grande dimension appelée espace de caractéristiques (possiblement de

dimension infinie) dans lequel il est probable qu'il existe une séparatrice linéaire ceci est réalisé grâce à une fonction noyau qui doit respecter certaines conditions et qui a l'avantage de ne pas nécessiter la connaissance explicite de la transformation à appliquer pour le changement d'espace. Les fonctions noyau permettent de transformer un produit scalaire dans un espace de grande dimension ce qui est coûteux sur une simple évaluation ponctuelle d'une fonction. [35]

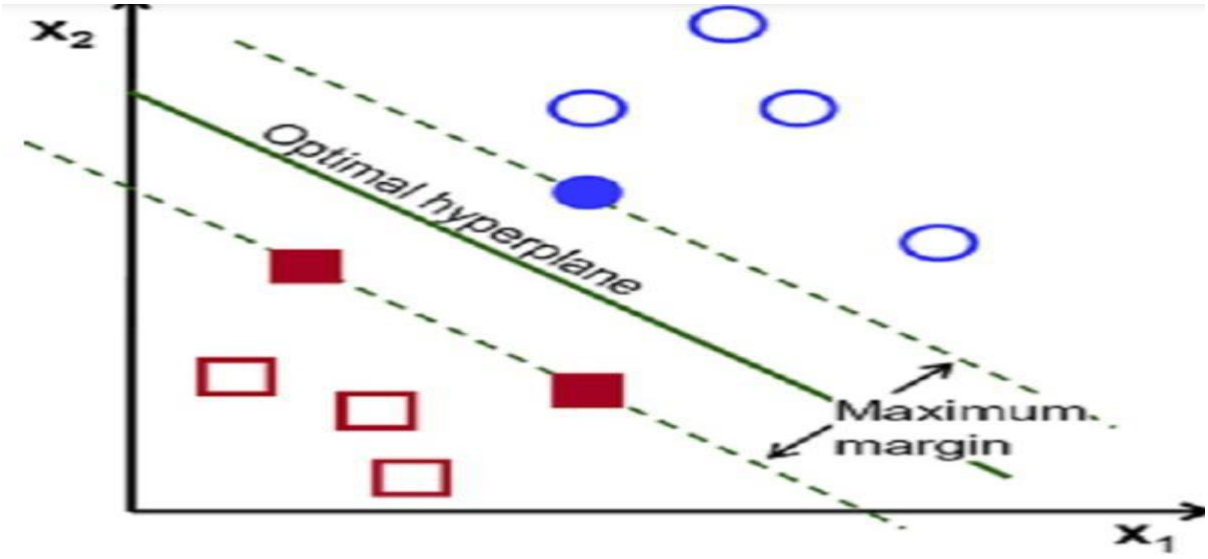


Figure II.2 : Le principe de SVM.

II.5.3.3 Les réseaux de neurones

Les réseaux de neurones sont inspirés de la méthode de travail du cerveau humain qui est totalement différente de celle d'un ordinateur. Le cerveau humain se base sur un système de traitement d'information parallèle et non linéaire, très compliqué, ce qui lui permet d'organiser ses composants pour traiter, d'une façon très performante et très rapide, des problèmes très compliqués tel que la reconnaissance des formes.

Un réseau de neurones est un modèle de calcul dont le fonctionnement vise à simuler le fonctionnement des neurones biologiques, Il est constitué d'un grand nombre d'unités (neurones) ayant chacune une petite mémoire locale et interconnectées par des canaux de communication qui transportent des données numériques, ces unités (neurones) peuvent uniquement agir sur leurs données locales et sur les entrées qu'elles reçoivent par leurs connections, les réseaux de neurones sont capables de prédire de nouvelles observations (sur des variables

spécifiques) à partir d'autres observations après avoir exécuté un processus d'apprentissage sur des données existantes.

Un neurone formel est l'unité élémentaire d'un système modélisé par un réseau de neurones artificiels, à la réception de signaux provenant d'autres neurones du réseau, le neurone formel réagit en produisant un signal de sortie qui sera transmis à d'autres neurones du réseau, ce signal est reçu comme une somme pondérée des signaux provenant des différentes neurones du réseau et le signal de sortie correspond à une fonction algébrique qui est une combinaison linéaire des entrées calculées comme suit :

$$v = \sum_{i=1}^{n} w_i x_i - w_0$$

Où Un ensemble de connexions avec les différentes entrées χ i, les Wi sont les poids synaptiques, W_0 le biais qui peut être considéré comme la pondération de l'entrée 0 fixée à 1. **υ** est appelé potentiel du neurone. La sortie du neurone sera :

$$y = f(v) = f(\sum_{i=1}^{n} w_i x_i - w_0)$$

f étant la fonction d'activation permettant de délimiter la sortie **y** du neurone. Elle opère comme une transformation qui est une combinaison affine des signaux d'entrée pour concevoir le signal de sortie. En classification les réseaux de neurones permettent d'introduire la notion de non-linéarité dans la séparation entre les classes grâce au choix de la fonction d'activation. [36]

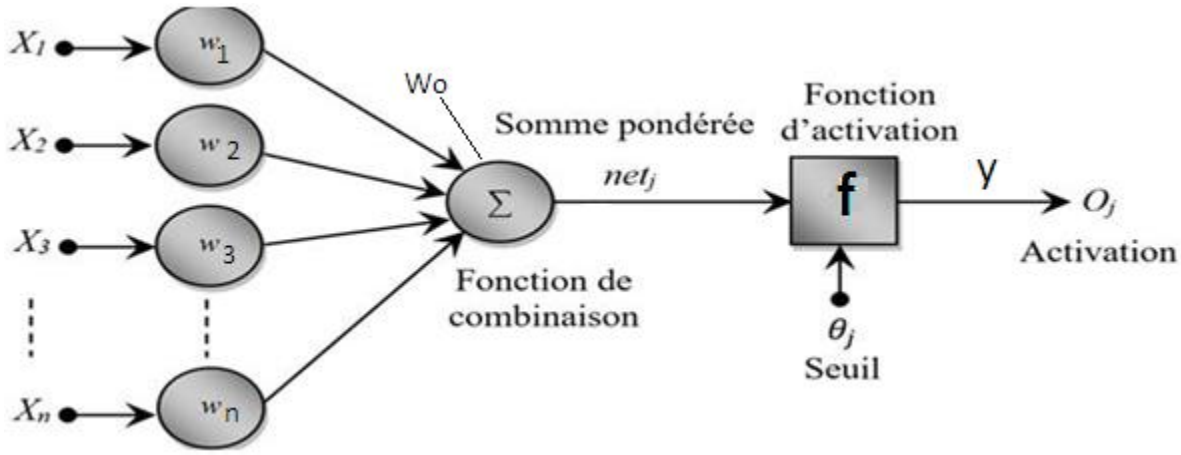


Figure II.3 : Représentation d'un neurone artificiel.

II.5.3.4 La classification bayésienne naïve

Ce sont des modèles graphiques dans lequel les connaissances sont représentées sous forme de variables, est représenté sous forme d'un graphe orienté acyclique, où les noeuds représentent les attributs et les arcs représentent les liaisons entre ces attributs (des probabilités conditionnelles). Deux attributs sont reliés par un arc si l'un cause ou influe sur l'autre : le prédécesseur est la cause et le successeur est l'effet.

Il existe plusieurs structures permettant d'employer les réseaux bayésiens comme classifieurs :

- Réseaux bayésiens naïfs.
- Réseaux bayésiens naïfs augmentés par un arbre.
- Réseaux bayésiens semi naïfs condensés etc…

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

Figure II.4 : Exemple sur classification bayésienne naïve.

II.5.3.5 k plus proches voisins (k-PPV)

L'algorithme des k-plus proches voisins (k-PPV) est un des algorithmes de classification les plus simples et utilisée en apprentissage supervisée, Les k-plus proches est diffère des méthodes d'apprentissages traditionnelles car aucun modèle n'est induit à partir d'exemples. Son principe est le suivant : une donnée

de classe inconnue est comparée à toutes les données stockées et ensuite une classe majoritaire parmi ses k-plus proches voisins est choisie pour la nouvelle donnée (elle peut donc être lourde pour les grandes bases de données) au sens d'une distance choisie.

K-plus Le seul outil dont on a besoin est une distance entre les éléments que l'on veut classifier. Si on représente ces éléments par des vecteurs de coordonnées, il y a en général pas mal de choix possibles pour ces distances, partant de la simple distance usuelle (euclidienne) en allant jusqu'à des mesures plus sophistiquées pour tenir compte si nécessaire de paramètres non numériques comme la couleur, la nationalité, etc...

L'échantillon d'apprentissage associé à une fonction de distance et à une fonction de choix de la classe suivant les classes des voisins les plus proches constituent le modèle élaboré, avant de classer un nouvel élément on le compare aux autres éléments en utilisant une mesure de similarité, les k les plus proches voisins sont alors considérés la classe qui revient le plus parmi les voisins est assignée à l'élément à classer.

Les voisins sont pondérés par la distance qui les sépare du nouvel élément à classer, le bon fonctionnement de la méthode dépend du choix d'un certain nombre de paramètre tel que le paramètre k qui représente le nombre des voisins choisis pour attribuer la classe au nouvel élément ainsi que la distance utilisée. [37]

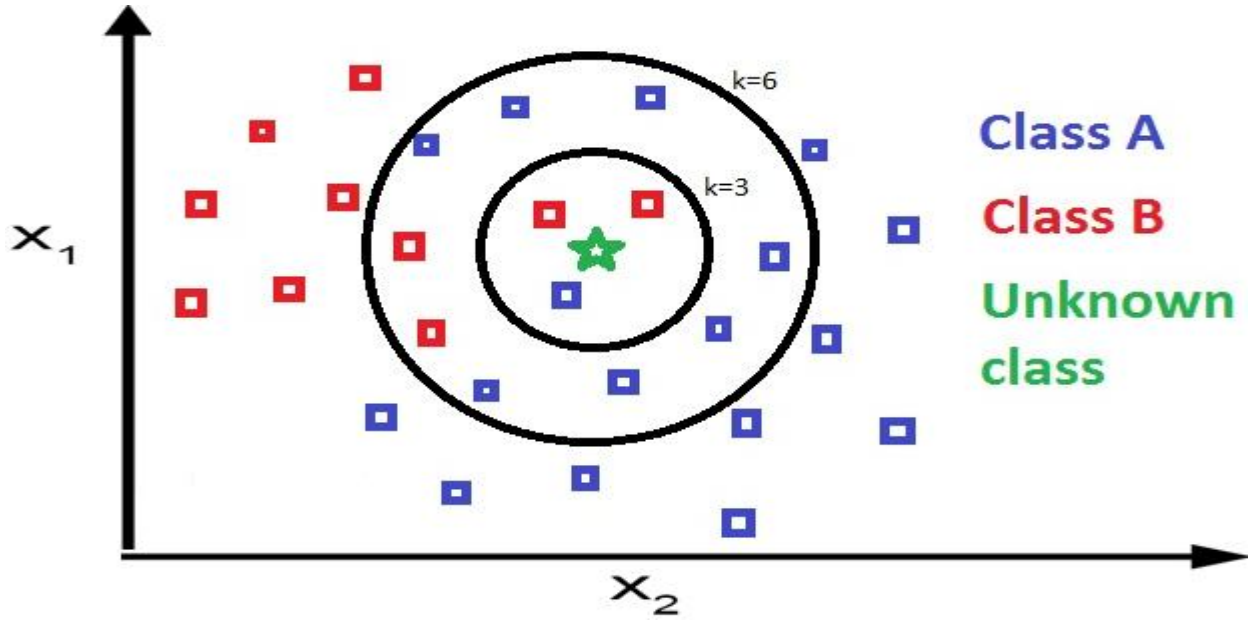


Figure II.5 : L'algorithme des k plus proches voisins.

Nous notons deux aspects importants de l'algorithme k-PPV :

D'une part à chaque nouvelle classification il est nécessaire de parcourir l'ensemble de la base d'apprentissage ce qui représente un algorithme qui n'est pas nécessairement très efficace (surtout qu'habituellement on cherche à avoir la base d'apprentissage la plus grande possible afin d'avoir un meilleur classifieur).

D'autre part un point crucial de cet algorithme est la fonction de distance utilisée pour mesurer la proximité des objets, il n'existe pas de distance/similarité universellement optimale et une bonne connaissance du problème traité guide généralement le choix de cette distance/similarité.

II.6 Domaines d'application

La classification comme dit préalablement joue un rôle dans presque toutes les sciences et techniques qui font appel à la statistique multidimensionnelle, à titre d'exemple les sciences biologiques, botanique, écologie, etc... Qui utilisent le terme "taxinomie" pour désigner l'art de la classification ainsi que les sciences de la terre et des eaux (géologie, géographie, étude des pollutions) font grand usage de classification, citons encore la médecine, l'économie, l'agronomie etc...

Dans toutes ces disciplines la classification peut être employée comme un domaine particulier mais elle l'est souvent vue comme une méthode complémentaire à d'autres méthodes statistiques. [38]

II.7 Conclusion

Dans ce chapitre nous avons vu la fouille de données est joue un rôle fondamental dans la compréhension des problèmes bioinformatiques.

Dans notre étude, nous nous sommes beaucoup de méthodes et techniques ont été développées afin d'y apporter des solutions efficaces, tout en exploitant les tâches de la fouille de données telles que deux phases de classification (supervisée et non supervisée clustering) et le qui demeurent indispensables.

Aussi nous avons, en détails les Méthodes classification supervisée.

Chapitre III :

Conception

III.1 Introduction

Dans ce chapitre, nous mettrons la conception d'un système de prédiction de fonction des RCPGs (récepteur couple protéine g) avec l'outil SVM multi classe, qui s'intéresse à comprendre et situer clairement les notions de base pour la conception, qui nous permettront la réalisation et la présentation de notre projet.

Ainsi, ce chapitre est consacré à notre Le processus de développement de notre système, nous permet de mettre en exergue une conception qui va décrire d'une manière non ambiguë son développement, la mise en place de l'architecture globale de notre système selon une vue interne, le détail des fonctionnalités de cette architecture avant de fournir sa réalisation.

III.2 Représentation du système

L'analyse de la base de données de la famille RCPGs par la méthode de classification supervisée SVM c'est le but de notre système. En ce qui concerne notre cas on possède une base de données, on a apriori la classe d'appartenance de chaque exemple de cette base de données.

Au début cette base de données est utilisée pour l'apprentissage du système servant ainsi à construire un classifieur ce qui permet son exploitation par la suit pour classer les exemples de la base de test à partir du nombre des exemples bien classés un taux de reconnaissance est calculé.

III.3 Conception globale

Globalement notre système est composé de quatre étapes qui se détaillent comme suit dans le schéma suivant :

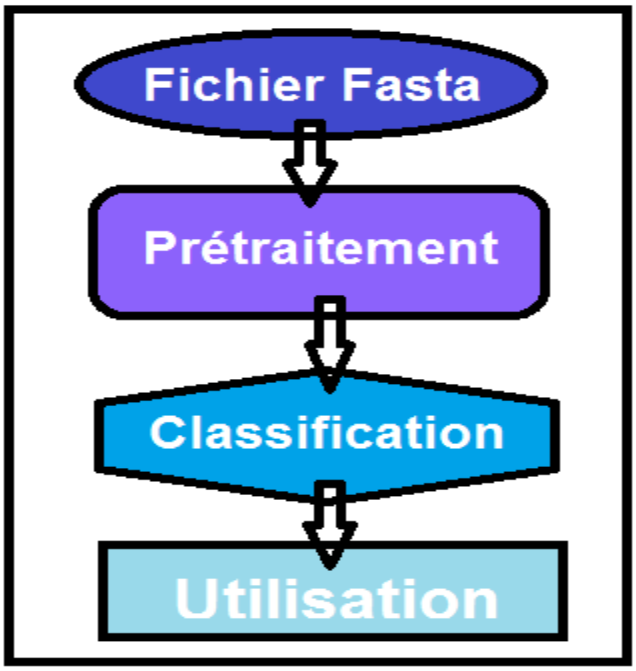


Figure III.1 : Conception globale du système.

III.3.1 Fichier d'accès

Le format FASTA (ou format Pearson) est un format de fichier texte utilisé pour stocker des séquences biologiques de nature nucléique ou protéique. Ces séquences biologiques sont représentées par une suite de lettres codées par des acides nucléiques.

Un fichier FASTA est composé au minimum de deux lignes. La ligne 1 décrit la séquence en commençant par le signe ">" suivi immédiatement de l'identifiant de la séquence et d'un commentaire séparé de l'identifiant par un espace.

La ligne 2 est constituée des lettres représentant les acides nucléiques ou les acides aminés de la séquence. Cette ligne possède cependant une longueur maximale de 120 résidus.

```
>Identifiant Commentaire
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
```

Figure III . 2 : Un fichier de format FASTA

III.3.2 Pré-traitement

La phase de prétraitement est nécessaire pour avoir un ensemble de données qui seront bien traités qui serviront à améliorer les performances. Dans cette phase de prétraitement il y a deux parties :

- la première partie consistant en une sélection des caractéristiques en utilisant PseAAC.
- la deuxième partie consiste en une normalisation.

III.3.3 Classification

Pour cette étape on a utilisé la méthode de classification supervisée SVM sert à séparer les données en données d'apprentissage et en données de test par la suite les données d'apprentissage permettent la construction d'un modèle de validation qui servira à tester les données du test, donc c'est la méthode utilisée pour la classification des données (prédiction de la fonction des protéines) dans notre travail.

III.3.4 Méthode de séparation

La méthode de la séparation holdout consiste à séparer les données en base d'apprentissage et la base du teste.

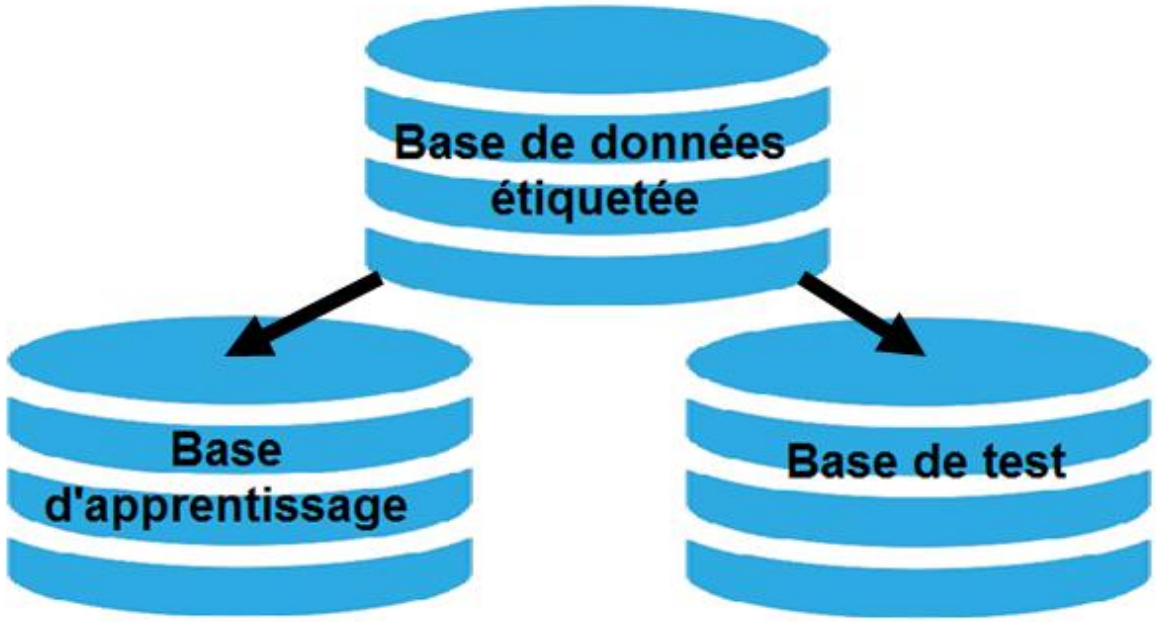


Figure III.3 : La séparation des données.

III.3.5 Utilisation

L'élaboration de cette étape est basée sur la mise en place d'un classifieur suivant la base d'apprentissage qui sera appliqué à la base du test, ce travail nous permettra de calculer un taux pour mesurer la performance de notre modèle de classification.

III.4 Conception détaillé

Pour obtenir une description des structures de données utilisées et leur communication entre elle il est nécessaire de mettre en place une conception détaillée pour parvenir en finale à une donnée classé.

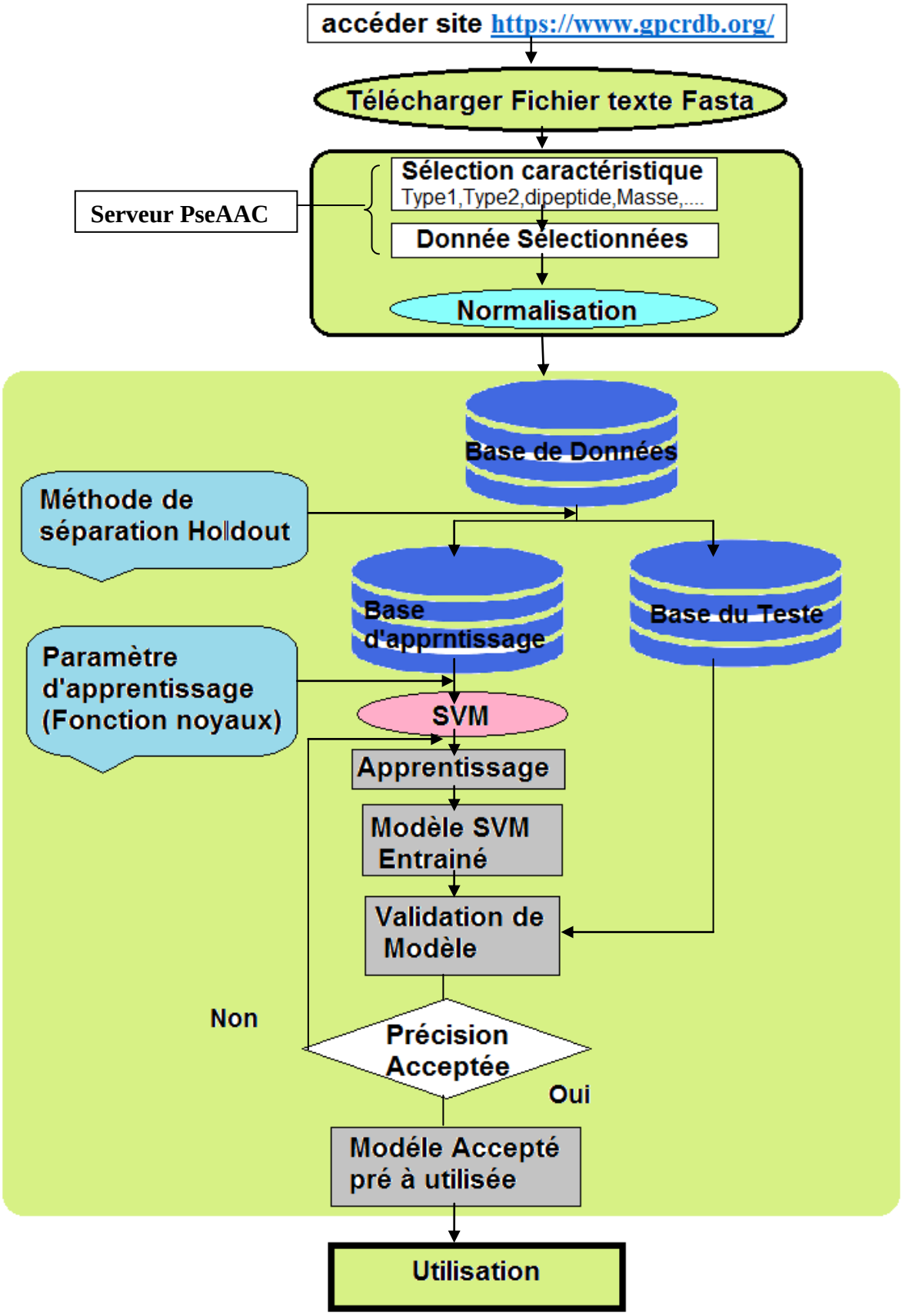


Figure III.4 : Schéma représentant la conception détaille du système.

III.4.1 Fichier d'accès

Un fichier FASTA multi-séquences ou multi-entrées est un fichier contenant plusieurs séquences d'une seule nature (nucléique ou protéique), chaque séquence étant précédée de sa ligne d'identification.

Les données brutes de récepteur couple protéine g sont stockées dans un fichier à l'extension fasta (figure IV.5).On obtient dans ce fichier une quantité de séquences biologique de la famille RCPGs, ces séquences biologique sont représentées par une suite de lettres codées par des acides nucléique. Ce fichier trouve son cheminement par un prétraitement selon notre base.

```
>>sp|Q5T6X5|GPC6A_HUMAN G-protein coupled receptor family A group 6 member A OS=Homo sapiens OX=9606 GN=GPRC6A
PE=1 SV=2MATTVPDGCRNGLKSKYYRLCDKAEAWGIVLETVATAGVVTSVAFMLTLPILVCKVQDSN
RRKMLPTQFLFLLGVLGIFGLTFAFIIGLDGSTGPTRFFLFGILFSICFSCLLAHAVSLT
KLVRGRKPLSLLVILGLAVGFSLVQDVIAIEYIVLTMNRTNVNVFSELSAPRRNEDFVLL
LTYVLFLMALTFLMSSFTFCGSFTGWKRHGAHIYLTMLLSIAIWVAWITLLMLPDFDRRW
DDTILSSALAANGWVFLLAYVSPEFWLLTKQRNPMDYPVEDAFCKPQLVKKSYGVENRAY
SQEEITQGFEETGDTLYAPYSTHFQLQNQPPQKEFSIPRAHAWPSPYKDYEVKKEGS

>sp|Q5T6X5|GPC6A_HUMAN G-protein coupled receptor family C group 6 member A OS=Homo sapiens OX=9606 GN=GPRC6A
PE=1 SV=1MAFLIILITCFVIILATSQPCQTPDDFVAATSPGHIIIGGLFAIHEKMLSSEDSPRRPQI
QECVGFEISVFLQTLAMIHSIEMINNSTLLPGVKLGYEIYDTCTEVTVAMAATLRFLSKF
NCSRETVEFKCDYSSYMPRVKAVIGSGYSEITMAVSRMLNLQLMPQVGYESTAEILSDKI
RFPSFLRTVPSDFHQIKAMAHLIQKSGWNWIGIITTDDDYGRLALNTFIIQAEANNVCIA
FKEVLPAFLSDNTIEVRINRTLKKIILEAQVNVIVVFLRQFHVFDLFNKAIEMNINKMWI
ASDNWSTATKITTIPNVKKIGKVVGFAFRRGNISSFHSFLQNLHLLPSDSHKLLHEYAMH
LSACAYVKDTDLSQCIFNHSQRTLAYKANKAIERNFVMRNDFLWDYAEPGLIHSIQLAVF
ALGYAIRDLCQARDCQNPNAFQPWELLGVLKNVTFTDGWNSFHFDAHGDLNTGYDVVLWK
EINGHMTVTKMAEYDLQNDVFIIPDQETKNEFRNLKQIQSKCSKECSPGQMKKTTRSQHI
CCYECQNCPENHYTNQTDMPHCLLCNNKTHWAPVRSTMCFEKEVEYLNWNDSLAILLLIL
SLLGIIFVLVVGIIFTRNLNTPVVKSSGGLRVCYVILLCHFLNFASTSFFIGEPQDFTCK
TRQTMFGVSFTLCISCILTKSLKILLAFSFDPKLQKFLKCLYRPILIIFTCTGIQVVICT
LWLIFAAPTVEVNVSLPRVIILECEEGSILAFGTMLGYIAILAFICFIFAFKGKYENYNE
AKFITFGMLIYFIAWITFIPIYATTFGKYVPAVEIIVILISNYGILYCTFIPKCYVIICK
QEINTKSAFLKMIYSYSSHSVSSIALSPASLDSMSGNVTMTNPSSSGKSATWQKSKDLQAQAFAHICRENATSVSKTLPRKRMSSI
```

Figure III.5 : Exemples de deux séquences de RCPGs (format fasta).

On peut télécharger ce fichier dans le site : https://www.gpcrdb.org/

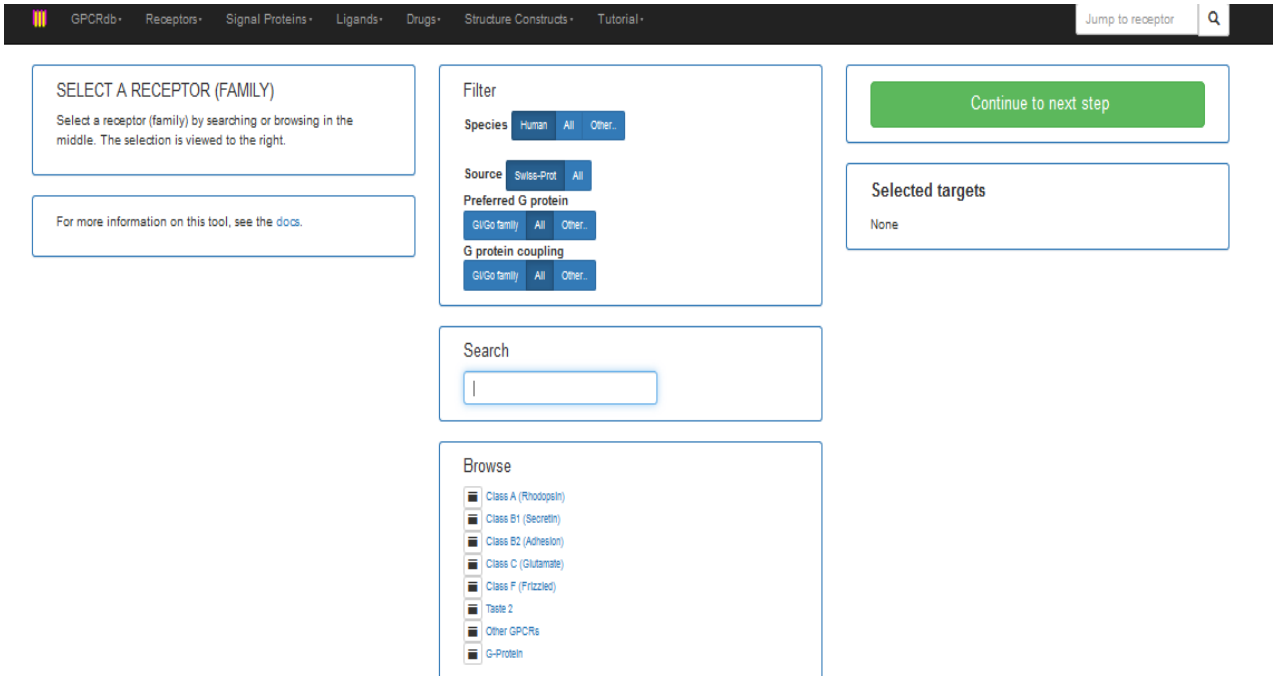


Figure III.6 : Le site de téléchargement de la famille RCPGs.

Pour télécharger le fichier de la famille Rhodopsine à partir du site sous-cité en haut on applique :

Pour la famille A (rhodopsine) : https://www.gpcrdb.org/family/001/.

Pour la famille B (sécrétine et adhérence) : https://www.gpcrdb.org/family/002/.

Pour la famille C (glutamate) : https://www.gpcrdb.org/family/004/.

Pour la famille F (crépu) : https://www.gpcrdb.org/family/005/

Pour la famille Taste 2 : https://www.gpcrdb.org/family/006/

Pour la famille Other GPCRs : https://www.gpcrdb.org/family/007/

Pour la famille G-Protein : https://www.gpcrdb.org/family/100/

III.4.2 Pré-traitement

III.4.2.1 Sélection des caractéristiques

Sélection des caractéristiques c'est le prétraitement des données biologique qui sont déterminées par la méthode la plus utilisée à savoir la composition en pseudo acides aminés.

Au lieu d'utiliser la composition d'acides aminés 20-D conventionnelle pour représenter l'échantillon d'une protéine, le professeur Kuo-Chen Chou a proposé la composition de pseudo acide aminé (PseAA) afin d'inclure les informations d'ordre de séquence. Basé sur le concept de la composition de pseudo-acides aminés de Chou, le serveur PseAA a été conçu de manière flexible, permettant aux utilisateurs de générer différents types de composition de pseudo-acides aminés pour une séquence de protéines donnée en sélectionnant différents paramètres et leurs combinaisons.

Les trois types de compositions PseAA ci-dessus sont actuellement pris en charge par PseAA, l'un est appelé composition PseAA Type1 , qui est également appelé type à corrélation parallèle et génère un vecteur (20+) -D pour chaque séquence protéique; le type2 est également appelé type à corrélation série et génère un vecteur (20+ *i* *) -D, où *i* est le nombre de attributs sélectionnés; la composition du dipeptide PseAA génère des nombres discrets 420-D pour représenter une séquence protéique.

Caractère d'acide aminé les valeurs d'hydrophobicité, d'hydrophilie, de Masse, de pK1 (alpha-COOH), pK2 (NH3) et pI (à 25 ^{o} C) utilisées par PseAA

Acide aminé	Hydrophobicité [a]	Hydrophilie [b]	Masse [c]	pK1 (a-CO2H) [d]	pK2 (NH3) [d]	pI (à 25 ^{o} C) [d]
UNE	0,62	-0,5	15,0	2,35	9,87	6.11
C	0,29	-1,0	47,0	1,71	10,78	5,02
ré	-0,90	3.0	59,0	1,88	9,60	2,98
E	-0,74	3.0	73,0	2.19	9,67	3.08
F	1.19	-2,5	91,0	2,58	9.24	5,91
g	0,48	0,0	1.0	2,34	9,60	6.06
H	-0,40	-0,5	82,0	1,78	8,97	7,64
je	1,38	-1,8	57,0	2,32	9,76	6.04
K	-1,50	3.0	73,0	2,20	8,90	9.47
L	1.06	-1,8	57,0	2,36	9,60	6.04
M	0,64	-1,3	75,0	2,28	9.21	5.74
N	-0,78	0,2	58,0	2.18	9.09	10,76
P	0,12	0,0	42,0	1,99	10,60	6.30
Q	-0,85	0,2	72,0	2.17	9.13	5.65
R	-2,53	3.0	101,0	2.18	9.09	10,76
S	-0,18	0,3	31,0	2.21	9.15	5.68
T	-0,05	-0,4	45,0	2.15	9.12	5,60
V	1,08	-1,5	43,0	2.29	9,74	6.02
W	0,81	-3,4	130,0	2,38	9.39	5.88
Oui	0,26	-2,3	107,0	2,20	9.11	5.63

Tableau III.1 : les valeurs d'acide aminé du serveur (PseAAC).

Le facteur de poids est conçu pour que les utilisateurs mettent du poids sur les composants PseAA supplémentaires par rapport aux composants AA conventionnels. Les utilisateurs sont autorisés à sélectionner le facteur de pondération de 0,05 à 0,70.

Le rang (ou niveau) compté de la corrélation le long d'une séquence protéique est généralement représenté par λ. Dans la composition PseAA de type 1, l'utilisateur obtiendra un λvecteur(20+) -D pour chaque séquence; dans la composition de PseAA de type 2, (20+ *i** λ) -Le vecteur D est généré, (où *i* est le nombre d'attributs d'acides aminés sélectionnés). Il est également important de noter que λla longueur de la séquence ne doit pas dépasser. Si l'utilisateur choisit λ= 0, la sortie sera la composition conventionnelle 20-D d'amino-aicd pour les deux cas. Pour des informations détaillées, veuillez cliquer sur lambda.

L'utilisateur doit saisir les séquences protéiques au format FASTA. Actuellement, PseAA accepte un maximum de 500 séquences de protéines pour chaque soumission.

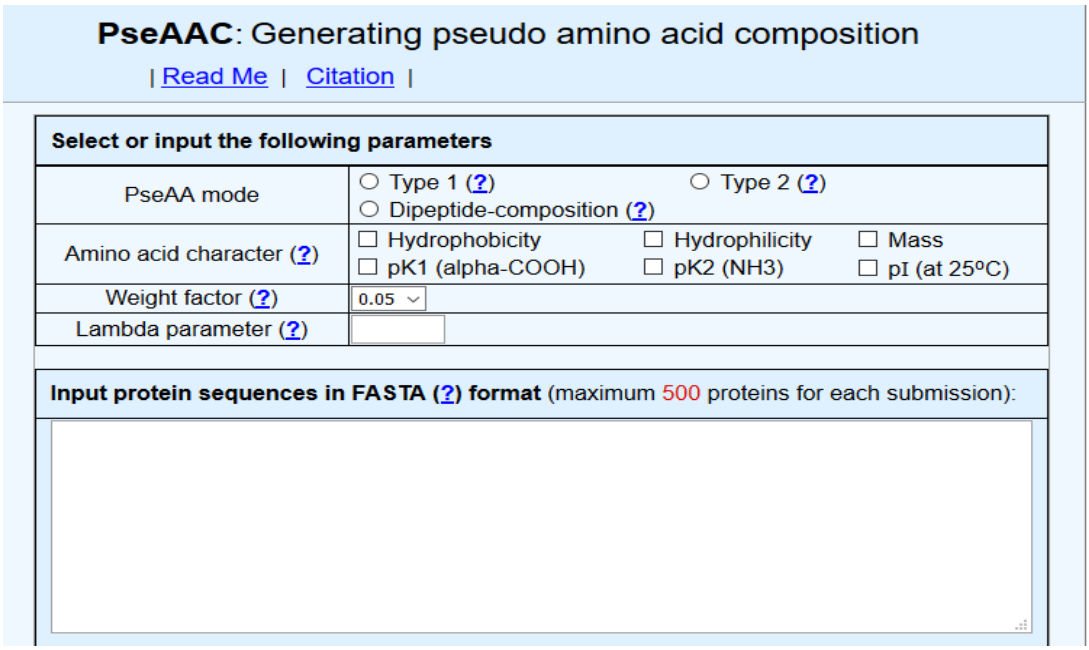


Figure III.7 : Serveur de composition en pseudo acides aminés (PseAAC).

III.4.2.2 La Normalisation

Les bonnes performances seront obtenues par une normalisation des données obtenues après application de la sélection des caractéristiques (en utilisant le PseAAC) soit effectuée.

Les approches les plus couramment utilisées pour la normalisation des caractéristiques est la normalisation (moyenne / variance) qui est la soustraction de la moyenne de la population et une mise à l'échelle. La formule de la normalisation est la suivante :

$$X_{ij} = \frac{\mathrm{X}_{ij} - \bar{X}_j}{\sigma_j}$$

Où Xijest la valeur de l'acide aminé i dans la j-nième séquence, Xj est la valeur de moyenne et σj est la déviation standard dans la j-nième séquence.

III.4.3 La méthode de classification SVM

La construction du classificateur SVM nécessite la spécification de quelques paramètres d'apprentissage tels que :

III.4.3.1 La méthode de la séparation des données

L'utilisateur doit prendre la méthode ci-après :

Holdout: partitionne les données exactement en deux sous-ensembles

- **Les données d'apprentissage « train » :** les données d'apprentissage permettent d'avoir un modèle entrainé.

- **Les données du test « test » :** l'ensemble de test est utilisé pour voir comment ce modèle fonctionne sur des données invisibles.

III.4.3.2 le choix de noyau (kernel)

L'utilisateur doit choisir le noyau à utiliser pour la construction du classificateur. Les noyaux les plus populaires sont : le noyau polynomial, le noyau gaussienne et le noyau mlp.

- Fonction polynomial: une fonction noyau pour le cas non linéairement séparable.

- Fonction base radiale gaussienne: une fonction noyau pour le cas non linéairement séparable.

- Fonction mlp: une fonction noyau pour le cas linéairement séparable.

III.4.4 Classification SVM multi classes (1vR)

Pour une classification multi-classes nous utilisons une stratégie Un Contre Tous (1vsR), l'idée de cette stratégie est de construire autant de classificateurs que de classes. Ainsi durant l'apprentissage tous les exemples appartenant à la classe considérée sont étiquetés positivement (+1) et tous les exemples n'appartenant pas à la classe sont étiquetés négativement (-1).

La sélection de la méthode de séparation avec la fonction noyau, cette opération mettra en place le modèle classificateur SVM qui est construit selon la base d'entrainement ensuite ce modèle est appliqué sur la base du test afin de classer nos données, les prédictions sur la base du test est donné en classifiant la base du test ce qui nous a permis d'avoir un taux de reconnaissance qui est calculé à partir du nombre des exemples bien classés.

III. 5 Conclusion

Dans ce chapitre la présentation de notre système de classification en axant essentiellement sur son rôle, également nous allons citer les différentes étapes en passant de la conception globale jusqu'à son la conception détail.

Le chapitre suivant mettra implémentations de notre système en apparence les étapes détaillées ayant trait à la conception. En suites la discussion et la comparaison sur les résultats obtenus.

Chapitre IV :
Implémentation

IV.1 Introduction

Dans le chapitre précédent, nous avons expliqué notre approche conceptuelle et discuté des différentes étapes de la sélection des données à la validation du modèle pour atteindre la phase de prise de décision.

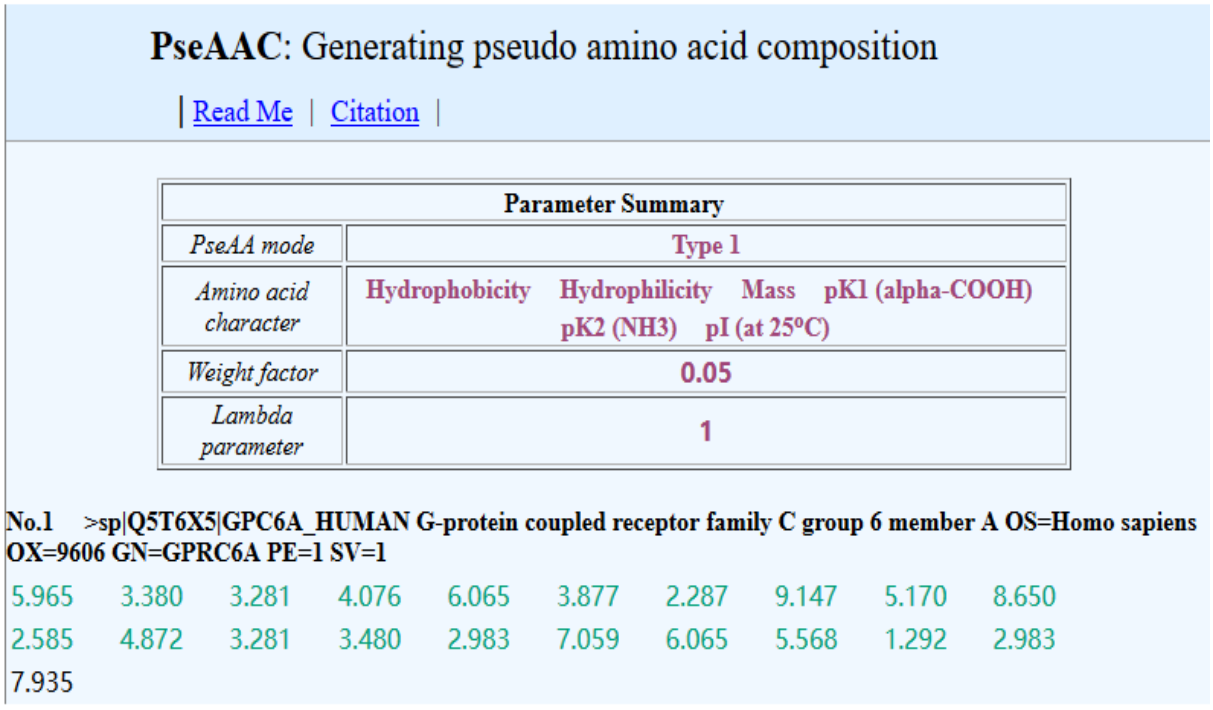
Ce chapitre a pour objectif de présenter l'implémentation du prétraitement, caractéristique de la Machine, l'environnement de la programmation et les fonctions principales utilisées pour développer et implémenter notre application ensuite la présentation de l'interface ainsi que les résultats de l'application, tout ce qui précède ce résume en deux étapes.

Premier étape ce constitue par :

IV.2 L'implémentation du prétraitement

IV.2.1 Sélection des caractéristiques

Dans cette phase on a téléchargé les familles (les récepteurs couples aux protéines G 'GPCRdb'), en suites de choisi serveur PseAAC mode de type1 auquel s'ajoute les caractéristiques d'acide Aminé qui sont : pK1 (alpha-COOH), pK2 (NH3) Hydrophobicité, Hydrophilie, Masse, pI (à 25°C) et Facteur de poids 0,05 qui font sortir 20 dimensions +λreprésentant une protéine.



PseAAC: Generating pseudo amino acid composition

| Read Me | Citation |

Parameter Summary	
PseAA mode	Type 1
Amino acid character	Hydrophobicity Hydrophilicity Mass pK1 (alpha-COOH) pK2 (NH3) pI (at 25°C)
Weight factor	0.05
Lambda parameter	1

No.1 >sp|Q5T6X5|GPC6A_HUMAN G-protein coupled receptor family C group 6 member A OS=Homo sapiens OX=9606 GN=GPRC6A PE=1 SV=1

5.965 3.380 3.281 4.076 6.065 3.877 2.287 9.147 5.170 8.650
2.585 4.872 3.281 3.480 2.983 7.059 6.065 5.568 1.292 2.983
7.935

Figure IV.1 La figure représente l'étape de sélection des caractéristiques.

IV.2.2 Normalisation

On a utilisé le language java pour la normalisation des données avec l'approche (moyenne / variance) la plus couramment utilisée pour la normalisation des caractéristiques.

```
for (int i = 0; i < dataFloatNormalis.length; i++) {
    for (int j = 0; j < dataFloatNormalis[0].length; j++) {
        dataFloatNormalis[i][j] = (dataFloatNormalis[i][j] - mean[j])/segma[j];
    }
}
```

Figure IV.2 La figure représente l'implémentation de l'étape de normalisation.

IV.3 Description sur le jeu de données

Dans notre application on a utilisé un ensemble de données pour mettre en place notre méthode de classification, l'ensemble des données qui sont appliquées sur la famille RCPG Se compose de sept (07) familles. Notre étude est axée sur sept familles (A, B, C,….) et chaque famille représente une classe avec 21 caractéristiques et la base comprend 9202 exemples.

Exemples de quelques séquences des récepteurs couples aux protéines G après les étapes de prétraitement des familles avec la PseAAC type 1 et la normalisation moyenne/ variance on obtient ce qui suit sur la figure IV.3.

```
0.50981593,0.13726988,0.1764571,0.09800626,0.1764571,0.2940952,0.09800626,0.19608891,
0.1764571,1.0,0.09800626,0.019631816,0.19608891,0.0,0.19608891,0.5490032,0.21572073,
0.43136507,0.03918723,0.09800626,0.4625315,ClassA
0.42184067,0.1561813,0.2968406,0.2968406,0.82815933,0.34368128,0.1561813,0.79684067,
0.5936813,1.0,0.18749999,0.46868134,0.35934067,0.18749999,0.17184067,0.78118134,0.42184067,
0.57815933,0.0,0.25,0.84491754,ClassC
0.6086924,0.15211765,0.043462187,0.06519329,0.30430925,0.28257817,0.13038658,0.36957645,0.0,1.0,
0.086924374,0.15211765,0.15211765,0.19565377,0.26084706,0.41303864,0.17392267,0.63042355,
0.02173109,0.10865548,0.4128169,ClassA
0.60002404,0.3142926,0.22860917,0.14292574,0.45709834,0.25717032,0.0,0.5714629,0.17148688,1.0,
0.20004801,0.22860917,0.3142926,0.20004801,0.37141493,0.45709834,0.5429018,0.79995203,0.028561149,
0.20004801,0.7159487,ClassA
```

Figure IV.3 Exemples des séquences après le prétraitement et la normalisation des familles.

La deuxième étape ce compose de :

IV.4 Présentation de la machine utilisée

Pour l'implémentation de notre système nous avons utilisé une machine présentant les caractéristiques suivantes :

- Processeur : Intel (R) Core (TM) i5-4310 U CPU @ 2.00 GHz 2.60 GHz.
- RAM : 4.00 GB.
- Système d'exploitation : Microsoft Windows 7 professionnel 32 bits.

IV.5 Outils et environnement de programmation

Dans ce cadre nous avons choisi comme environnement de programmation le langage MATLAB qui possède une richesse et offre une grande simplicité de manipulation dans le domaine bioinformatique, on a utilisé le weka avec la méthode supervisé réseau de neurone pour comparer les résultats.

IV.5.1 Matlab

Est une plate-forme de programmation spécialement conçue pour les ingénieurs et les scientifiques, celle-ci permet de faire le calcul matriciel, de développer et d'exécuter les algorithmes. Le MATLAB est devenu un logiciel de programmation largement utilisé il est présent dans plusieurs domaines tel que les systèmes d'apprentissage automatique, les appareils de surveillance médicale, dans le domaine de traitement du signal et de la vision par ordinateur. Avec ses nombreuses fonctions spécialisées pré codées et ses différentes «tool box » disponibles il permet une prise en main rapide et efficace. [39] Ce langage possède des avantages utiles tel que :

- Facilité de manipulation des matrices en ce qui concerne notre application.
- Langage interprété : pas de compilation donc pas d'attente pour compiler.
- Programmation infiniment plus rapide pour le calcul et pour l'affichage.
- Code facile à comprendre et très lisible.
- Le langage Matlab est un programme utile pour celui qui l'apprend en particulier dans le domaine bioinformatique.
- Il existe de nombreuses bibliothèques dans tous les domaines.

IV.5.1.1 Bibliothèque LIBSVM utilisée

LIBSVM est également une bibliothèque destinée aux machines à vecteurs de support (SVMs) développée dans le but de simplifier l'utilisation des SVM comme outil. Une utilisation typique de LIBSVM implique deux étapes:

- Premièrement : entrainent les données pour obtenir un modèle
- Deuxièmement : utilise le modèle pour prédire les informations d'un jeu de données de test. [40]

IV.5.1.1.1 Train SVM

Les paramètres de Train SVM sont :

- **Training set** : matrice de données d'apprentissage où chaque ligne correspond à une observation ou à une réplique et chaque colonne correspond à une caractéristique ou à une variable.
- **GIvAll** : qui peut être un vecteur de catégorie, numérique ou de logique, un tableau de cellules de vecteurs de caractères ou une matrice de caractères et chaque ligne représente une étiquette de classe.
- **Kernel function** : fonction du noyau que svm train utilise pour mapper les données d'apprentissage dans l'espace du noyau. [41]

```
u=unique(GroupTrain);
numClasses=length(u);
result = zeros(length(TestSet(:,1)),1);
%build models
for k=1:numClasses
    %Vectorized statement that binarizes Group
    %where 1 is the current class and 0 is all other classes
    G1vAll=(GroupTrain==u(k));
    models(k) = svmtrain(TrainingSet,G1vAll,'kernel_function',Noyeau);
end
```

Figure IV.4 Illustration de SVM train.

IV.5.1.1.2 SVM classifier

Les paramètres de SVM classifier sont :

- **Test Set** : matrice de données de test où chaque ligne correspond à une observation ou à une réplique et chaque colonne correspond à une caractéristique ou une à variable.
- **Models** : Structure de classificateur SVM crée à l'aide de la fonction svmtrain.[42]

```
% test cases
for j=1:size(TestSet,1)
    for k=1:numClasses
        if(svmclassify(models(k),TestSet(j,:)))
            break;
        end
    end
    result(j) = k;
    end
```

Figure IV.5 Illustration de SVM classifier.

IV.5.1.2 La boite à outil Guide

Les interfaces graphiques (ou les interfaces homme-machine) sont appelées GUI (Graphical User Interface) sous MATLAB, ils permettent à l'utilisateur d'interagir avec un programme informatique et de concevoir des interfaces utilisateurs pour des applications personnalisées avec différents objets graphiques (boutons, menus, cases à cocher ...), ces objets sont généralement utilisés à l'aide de la souris ou du clavier. Le GUIDE génère automatiquement le code MATLAB pour la construction de l'interface utilisateur qu'on peut modifier pour programmer notre application.

IV.5.2 L'outil weka

Weka est une collection d'algorithmes d'apprentissage automatique pour les tâches d'exploration de données. Les algorithmes peuvent être directement appliqués à un ensemble de données ou appelés à partir de votre propre code Java. Weka contient des outils pour le prétraitement, la classification, la régression, le regroupement, les règles d'association et la visualisation des données. [43]

IV.6 Description de l'application

IV.6.1 L'interface graphique

L'interface graphique facilite la consultation du système par les utilisateurs, l'exécution de notre système consiste à voir en premier une fenêtre d'accueil de notre système comme le montre la figure IV.6.

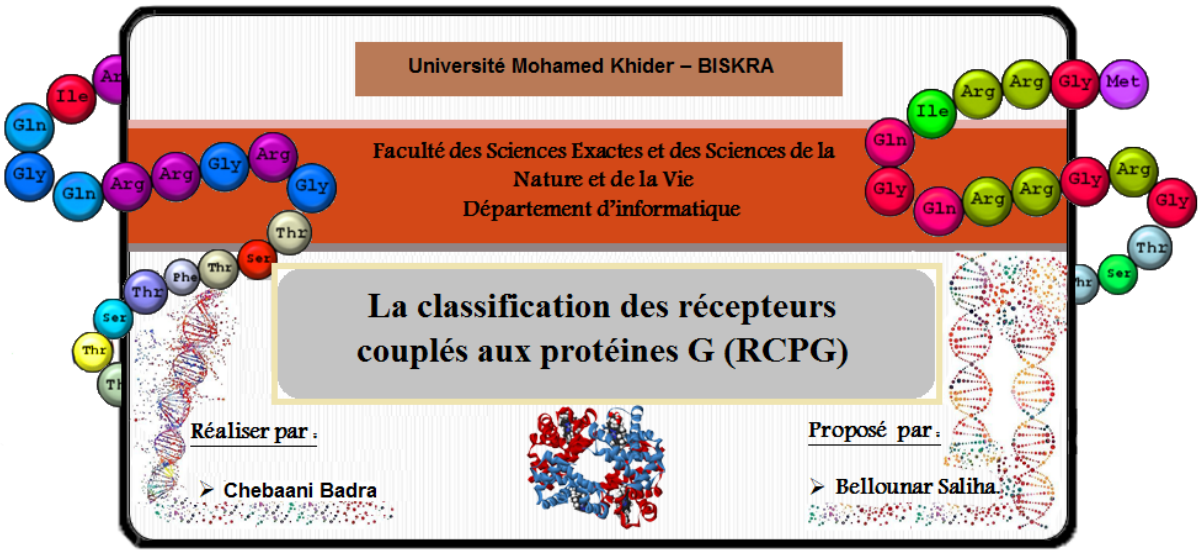


Figure IV.6 La fenêtre d'accueil de notre système

Explicatif du graphique :

1. Le titre du l'application.

2. entrée à la fenêtre principale. En appuyant sur le bouton Suivant apparait la fenêtre principale de notre système comme c'est précisé dans la figure IV.7.

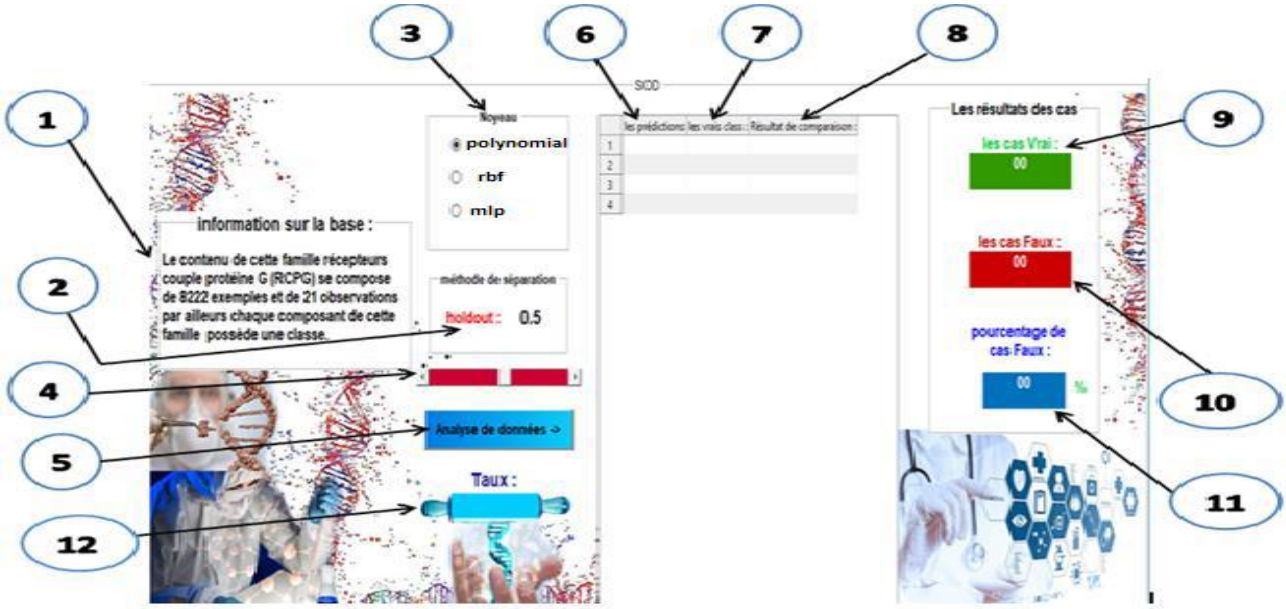


Figure IV.7 La fenêtre de classification de la famille RCPG.

Explicatif de la figure IV.7

1. Des informations sur le jeu de données.

2. Méthode de séparation HOLDOUT.

3. Une liste de la fonction noyau disponible pour la classification SVM.

4. Le nombre du ratio pour la séparation des données HOLDOUT.

5. bouton du démarrage de l'exécution.

6. Les prédictions du model SVM.

7. Les vrais classes de base du test.

8. Les résultats de comparaison (V ou F).

9. Les résultats des cas Vrai.

10. Les résultats des cas Faux.

11. pourcentage des cas Faux.

12. Le taux correct de la classification.

IV.6.2 Résultats Expérimentaux

IV.6.2.1 Résultats du language Matlab

Le tableau **IV.1** illustre les valeurs des différents taux de performances du système de prédiction des fonctions de la famille récepteurs couplés aux protéines G (RCPG)avec la méthode de classification supervisé SVM multi classe.

Noyaux	Méthode de séparation	nombre de ratio pour la séparation		Taux
		train	test	
Polynomial	HOLDOUT	0.9	0.1	98.301
Polynomial		0.7	0.3	97.812
Polynomial		0.5	0.5	97.2756
rbf		0.9	0.1	97.3301
rbf		0.7	0.3	97.2447
rbf		0.5	0.5	96.181
mlp		0.9	0.1	52.9126
mlp		0.7	0.3	53.0794
mlp		0.5	0.5	50.7419

Tableau IV.1 Illustre les valeurs des différents taux de performances.

La Figure **IV.8** représente le meilleur résultat de la classification SVM multi classe sur la base de la famille RCPG, en choisissons la méthode de la séparation HOLDOUT et le noyau polynomial.

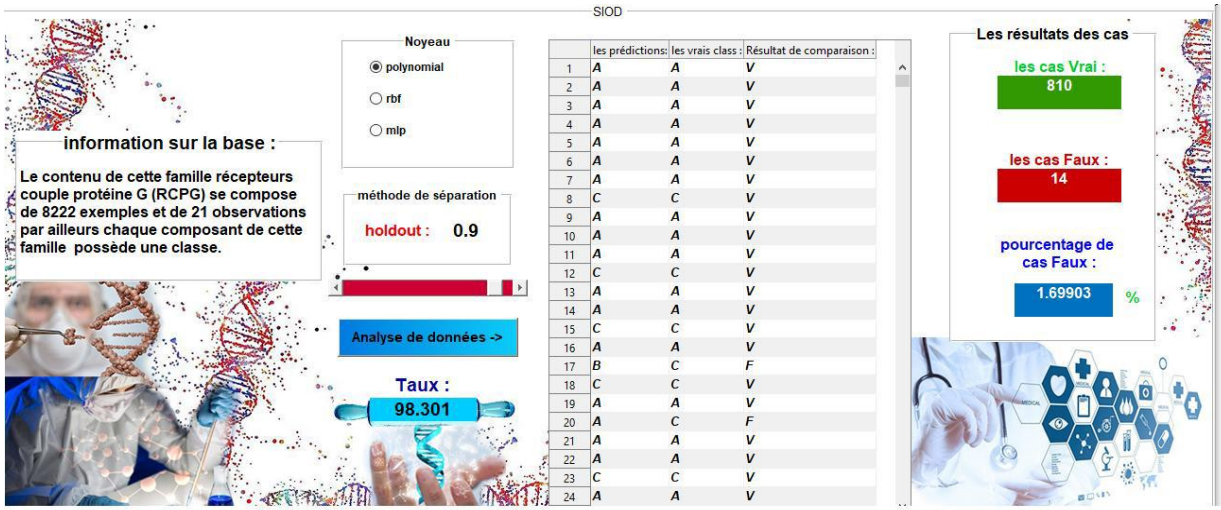


Figure IV.8 Le meilleur résultat de la classification SVM multi classe sur notre application.

IV.6.2.2 Résultats du weka

Le tableau V.2 illustre les valeurs des différents taux de performances du système de prédiction des fonctions de la famille récepteurs couplés aux protéines G (RCPG) avec la méthode de classification supervisé réseau de neurone .

Nombre couches cachées	Méthode de séparation	nombre de ratio pour la séparation		Taux
		train	test	
4, 6,4	HOLDOUT	0.9	0.1	96.2287
4, 6,4		0.7	0.3	94.6088
4, 6,4		0.5	0.5	94.7458
2, 4,2		0.9	0.1	94.6472
2, 4,2		0.7	0.3	94.3656
2, 4,2		0.5	0.5	94.3809
5, 9,5		0.9	0.1	96.8370
5, 9,5		0.7	0.3	96.1897
5, 9,5		0.5	0.5	95.7675

Tableau IV.2 Illustre les valeurs des différents taux de performances.

La Figure **IV.9** représente le meilleur résultat de la classification réseau de neurone (Multilayer Perceptron) sur la base de la famille RCPG à l'extension csv, en choisissons la méthode de la séparation HOLDOUT (percentage splite) avec le nombre de couches cachées 4, 6, 4.

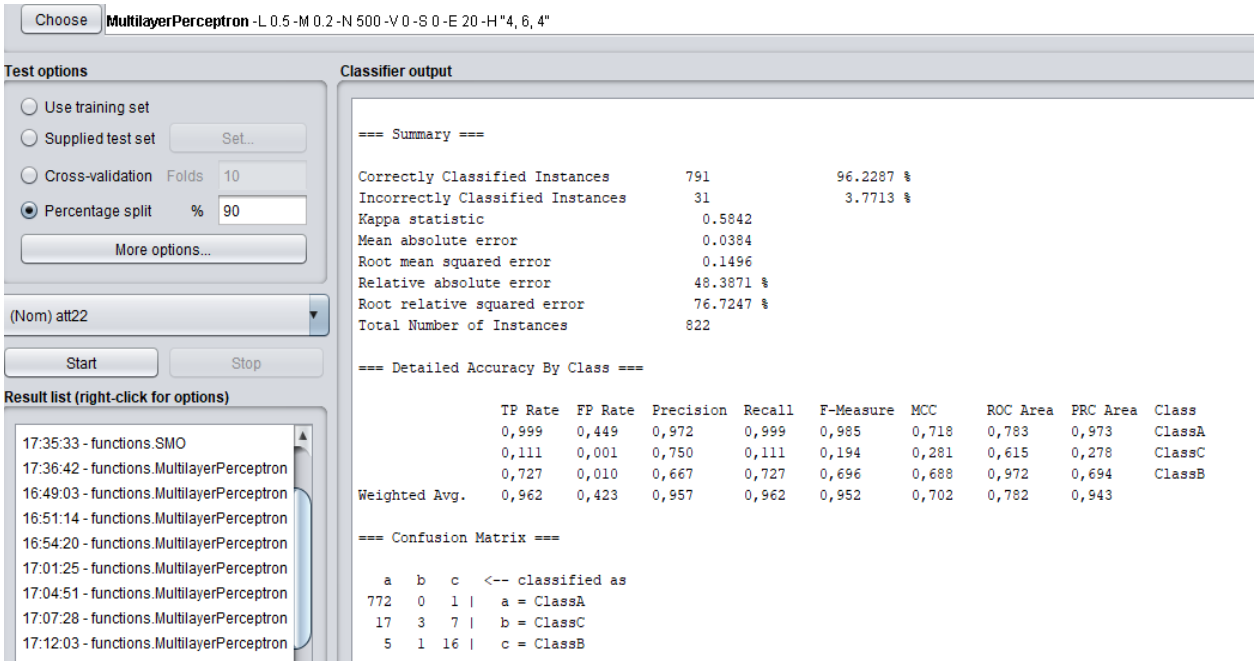


Figure IV.9 Le meilleur résultat de la classification réseau de neurone par l'outil weka.

IV.7 Discussion

Nous avons utilisé deux méthodes SVM multi classe et réseau de neurone pour la classification des données de la famille des récepteurs couplées aux protéines G. Nous remarquons que le meilleur taux obtenu dans SVM multi classe est de 99.1505% par noyau =polynomial et nombre de ratio= 0.9. Par réseau de neurone meilleur taux 96.2287% par le nombre de ratio =90 % et avec le nombre de couche cachées 4, 6, 4. Nous pouvons dire que la méthode SVM reste une méthode efficace et performante dans le domaine de l'apprentissage automatique.

IV.8 Conclusion

Dans ce chapitre, nous avons présenté tous les outils et toutes les étapes nécessaires à la mise en œuvre de notre projet que nous avons discuté en donnant des résultats sur notre méthode de classification qui était tout à fait acceptable et satisfaisante.

Conclusion

Conclusion générale

L'évolution des données biologiques durant ces dernières années a conduit à l'émergence d'un nouvel axe de recherche regroupant diverses disciplines comme l'informatique, les mathématiques, les statistiques et la biologie, sous le nom de Bioinformatique.

L'étude menée sur la classification des données ayant trait à la prédiction de fonction de la famille récepteur couplée aux protéines G (RCPG) par l'utilisation de la méthode de classification supervisée machine à vecteurs de support (SVM) est une méthode très performante.

Cette méthode mise au point a été testé sur un jeu de données, le constat de nos différents tests a fait ressortir des résultats très satisfaisants avec la méthode de classification SVM.

Par l'implémentation de notre application qui permet la classification des données de la famille récepteur couplée aux protéines G, a offert une interface interactive qui facilite l'utilisation de ses différents composants.

Par l'utilisation de l'outil weka on a pris la méthode réseau de neurone pour la détermination des taux de classification, la comparaison des deux résultats des deux méthodes (SVM et réseau de neurone) démontre que les deux méthodes proposées sont compétitifs.

Auparavant nous avons procédé à la classification des données de la famille récepteur couplée aux protéines G qui a donné des résultats très compétitifs, notre application pourrait dorénavant être encore améliorée pour cela il sera proposé d'analyser d'autres types de bases données et d'ajouter des algorithmes de classification supervisée.

La Bibliographie

[1] : https://www.scq.ubc.ca/what-is-bioinformatics/.

[2] : Rao, V.S., Das, S.K., Rao, V.J., Srinubabu, G.: Recent Developments in Life Sciences Research: Role of Bioinformatics, African Journal of Biotechnology, vol. 7, issue 5, 2008.

[3] : https://www.futura-sciences.com/sante/definitions/adn-mitochondrial genomique-156/.

[4] : J.D.Watson and F.H.C.Crick. Molecular structure of nucleic acids: structure for deoxyribose nucleic acid.Nature, (4356),1953.

[5] : https://ghr.nlm.nih.gov/primer/basics/gene.

[6] : https://ghr.nlm.nih.gov/primer/basics/genome.

[7] : A.LAYEB. Approche quantique évolutionnaire pour l‟‟aligne- ment multiple de séquences en bioinformatique, thèse, université des frères mentouri constantine,2005.

[8] : W. Dubitzky, M. Granzow, D. Berrar (Eds.), Fundamentals of Data Mining in Genomics and Proteomics, Springer Science+Business Media, 2007.

[9] : https://fr.wikipedia.org/wiki/Structure_des_prot%C3%A9ines.

[10] : Kerbellec, G.: Apprentissage d'Automates Modélisant des Familles de Séquences Protéiques.

[11] : Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., Walter, P: Molecular Biology of the Cell, Taylor and Francis Group, 2008.

[12] : Dzuida, D.M.: Data Mining for Genomics and Proteomics, Analysis of Gene and Protein Expression Data, John Wiley & Sons, 2010.

[13] : Kihara, D., Hawkins, T., Luban, S , B., Ramani, K., Agrawal, M.: Protein Function Prediction in Proteomics Era, Frontiers of Computational Science,2007.

[14] : https://www.ncbi.nlm.nih.gov/books/NBK26911/.

[15] : https://en.wikipedia.org/wiki/Protein/.

[16] : Cohen, J.: Bioinformatics-An Introduction for Computer Scientists, ACM Computing Surveys, vol. 36, issue 2, 2004.

[17] : Selzer, P.M., Marhöfer, R.J., Rohwer, A.: Applied Bioinformatics: An Introduction, Springer-Verlag Berlin Heidelberg, 2008.

[18] : http://ressources.unisciel.fr/biocell/chap11/co/module_Chap11_1.html.

[19] : https://www.britannica.com/science/G-protein-coupled-receptor/.

[20] : Rovati, G. E., Capra, V. & Neubig, R. R. (2007) The highly conserved DRY motif of class A G protein-coupled receptors: beyond the ground state, MolPharmacol. 71, 959-64.

[21] : Miller, L. J., Dong, M., Harikumar, K. G. &Gao, F. (2007) Structural basis of natural ligand binding and activation of the Class II G-protein-coupled secretin receptor, Biochem SocTrans. 35, 709-12.

[22] : Pantages, E. &Dulac, C. (2000) A novel family of candidate pheromone receptors in mammals, Neuron. 28, 835-45.

[23] : http://rcpg.chez.com/partie4.html/.

[24] : https://www.nature.com/scitable/topicpage/gpcr-14047471: Aude Sauliere. Etapes membranaires de la transduction du signal par les récepteurs couplés aux protéines G: organisation dynamique du récepteur mu aux opioïdes humain à la surface de neuroblastomes.

[25] : http://docs.gpcrdb.org/.

[26] : http://www.didier-pol.net/2INT-PRO.html.

[27] : Abdelhamid Djeffal. "Fouille de données Avancé".

[28] : https://www.digitalvidya.com/blog/what-is-data-mining/.

[29] : https://cloudtweaks.com/2014/09/use-supervised-unsupervised-data-mining/.

[30] : Guillaume Bouchard. Les modèles génératifs en classification supervisée et applications à la catégorisation d'images et à la fiabilité industrielle. Thèse de doctorat. University Joseph Fourier â Grenoble 1, 2005.

[31] : E.levine et E.Domany. Resampling method for unsupervised estimation of cluster validity.Neural Computation, 13(11):2573-2593; 2001

[32] : Nadjia Khatir. Clustering dans les bases de données. Mémoire de magistère. Informatique. Université d'oran Es Senia.

[33] : Arnaud Buhot.Etude de propriétés d‟‟apprentissage superviséet non supervisé par des méthodes de Physique Statistique.ÉCOLE DOCTORALE STIM.

[34] : H. Chouaib, Sélection de caractéristiques : méthodes et applications, Université Paris Descartes, Paris, France: Thèse de Doctorat, 2011.

[35] : B. Schölkopf and A. Smola. Learning with kernels. Support vector machines, regularization, optimization, and behond. MIT.Press, 2002.

[36] : W. E. Dietz, E. L. Kiech and M. Ali, "Classification of data patterns using and autoassociative neural network topology," In Proceedings of the 2nd international conference on Industrial and engineering applications of artifical intelligence and expert systems, vol. 2, no1, 1989.

[37] : P. Indyk and R. Motwani, "Approximate nearest neighbors: towards removing the curse of dimensionality," STOC '98 Proceedings of the thirtieth annual ACM symposium on Theory of computing, 1998.

[38] : Marc jamoulle, Michel Roland, Jacques Humbert, Jean-François Brûlet. Traitement de l‟‟information médicale par la Classification internationale des soins primaires, deuxième version : CISP-2. Care Edition, Bruxelles, 2000.

[39] : https://www.mathworks.com/discovery/what-is-matlab.html/.

[40] : Chih-Chung Chang and Chih-Jen Lin A Library for Support Vector Machines.

[41] : http://es.mathworks.com/help/stats/svmtrain.html.

[42] : https://www.analyticsvidhya.com/learning-paths-data-science-business-analytics-business-intelligence-big-data/weka-gui-learn-machine-learning/.