Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

وزارة التعليم العالي و البحث العلمي

BADJI MOKHTAR-ANNABA UNIVERSITY
UNIVERSITY BADJI MOKHTAR-ANNABA

جامعة باجي مختار- عنابة

Faculté des Sciences de l'Ingéniorat
Département d'Informatique

Année : 2019/2020

THÈSE

Présentée en vue de l'obtention du diplôme de

Doctorat LMD 3éme cycle

Intitulée :

Fouille de données bio-inspirée en bioinformatique

Filière : **Informatique**
Spécialité : **Système Informatique**

Par
Safia Bekhouche

Devant le Jury

Pr Halima BAHI	Professeur à l'Université d'Annaba	Président
Pr Yamina MOHAMED BEN ALI	Professeur à l'Université d'Annaba	Directeur de Thèse
Pr Nabiha AZIZI	Professeur à l'Université d'Annaba	Examinateur
Dr Zineddine KOUAHLA	MCA à l'Université de Guelma	Examinateur

A l'âme de ma sœur qui m'a souhaité toujours le meilleur sans récompense.

***A** tous ceux qui m'ont aidé de prés ou de loin à accomplir ce travail.*

***A** tous ceux qui sont chères.*

***A** tous ceux qui sont dans mon coeur.*

Remerciements

Mes premiers remerciements vont à Madame Mohamed Ben Ali Yamina, le directeur de ce travail. Je suis heureuse d'avoir eu la chance durant ces dernières années de partager ses réflexions, de disposer de son expertise scientifique et de son expérience pour mener à bien ce travail.

Je tiens aussi à remercier Mr Kouahla Zineddine, MCA à l'université de Guelma, Mme Azizi Nabiha, professeur à l'université Badji Mokhtar Annaba, de m'avoir fait l'honneur d'examiner cette thèse.

Je tiens aussi à remercier Mme Bahi Halima, professeur à l'université Badji Mokhtar Annaba, de m'avoir fait l'honneur de présider cette thèse.

Je souhaite associer à ces remerciements ma famille, mes amis, mes collègues et tous ceux que je connais de près ou de loin.

ملخص

تعتبر مستقبلات البروتين G (GPCRs) أكبر عائلة من مستقبلات سطح الخلية .لا يزال الكثير منهم أيتامًا. يعد التنبؤ بوظيفة GPCRs مهمة مهمة للغاية في المعلوماتية الحيوية .وهو يتألف من تعيين الفئة الوظيفية المقابلة للبروتين .تتطلب خطوة التصنيف هذه طريقة تمثيل بروتين جيدة وخوارزمية تصنيف قوية.

ومع ذلك ، يمكن زيادة تعقيد هذه المهمة بسبب العدد الكبير من ميزات تسلسل GPCR في معظم قواعد البيانات الحالية ، والتي تنتج انفجارًا اندماجيًا.

من أجل تقليل التعقيد وتحسين التصنيف .اقترحنا استخدام الاستدلال الفوقي المستوحى من الحيوية لاختيار الميزات واختيار أفضل زوج) استراتيجية استخراج الميزات (FES) وخوارزمية استخراج البيانات.((DMA)

اخترنا أيضًا استخدام خوارزمية BAT لاستخراج الخصائص ذات الصلة والخوارزمية الجينية لاختيار أفضل زوج

قارنا النتائج التي تم الحصول عليها بخوارزميتين موجودتين .النتائج التجريبية تشير إلى كفاءة النظام المقترح.

كلـمـات مفاتيـح : *GPCR، التنبؤ الوظيفي ، اختيار الميزات ، خوارزمية BAT ، الخوارزمية الجينية ، استراتيجيات استخراج الميزات ، التصنيف ، خوارزميات استخراج البيانات ، تمثيل البروتين.*

Résumé

Les Récepteurs Couplés aux Protéines G (RCPG) sont la plus grande famille de récepteurs de surface cellulaire; beaucoup d'entre eux sont encore orphelins. La prédiction de la fonction des RCPG représente une tâche très importante de la bioinformatique. Elle consiste à attribuer à une protéine, la classe fonctionnelle correspondante. Cette étape de classification nécessite une bonne méthode de représentation des protéines et un algorithme de classification robuste.

Cependant, la complexité de cette tâche pourrait être augmentée en raison du grand nombre de caractéristiques, produites de différentes MRP, qui produisent une explosion combinatoire.

Dans la littérature, il existe diverse stratégies d'extraction de caractéristiques pour représenter une chaine protéique, le choix de l'une ou de l'autre n'est pas évident, puisqu'il n'existe aucune stratégie adéquate pour toutes les données.

Au début, nous avons fait une étude empirique des MRP les plus utilisées dans les travaux de la classification des RCPG, cette étude est basée sur différents algorithmes de data mining, implémentés dans l'environnement Weka. Cette proposition est effectuée pour analyser la diversité des résultats de la précision et de taux d'erreur de la classification d'une méthode à l'autre, où nous avons conclu l'influence de le bon choix de la MRP et de l'algorithme de classification à la fois.

Afin de réduire la complexité et d'optimiser la classification en évitant le problème d'explosion combinatoire, nous avons proposé d'utiliser deux méta-heuristiques bio-inspirées à la fois pour la sélection des caractéristiques et pour le choix du meilleur couple (stratégie d'extraction de caractéristiques (MRP), algorithme d'exploration de données (AFD)).

Nous avons opté également d'utiliser l'algorithme de chauve-souris (Bat) pour extraire les caractéristiques pertinentes et l'algorithme génétique (AG) pour choisir le meilleur couple. Nous avons comparé les résultats obtenus avec deux algorithmes existants. Les résultats expérimentaux indiquent l'efficacité du système proposé.

Mots clés : RCPG, prédiction de la fonction, sélection de caractéristiques, algorithme BAT, algorithme génétique, stratégies d'extraction de caractéristiques, classification, algorithmes d'exploration de données, représentation de protéines.

Abstract

G Protein Coupled Receptors (GPCRs) are the largest family of cell membrane receptors; many of them are still orphans. Predicting the function of GPCRs is a very important task in bioinformatics. It consists of assigning the corresponding functional class to the protein. This classification step requires a good protein representation method and a robust classification algorithm.

However, the complexity of this task could be increased due to the large number of features, produced from different PRM, which leads to a combinatorial explosion.

In the literature, there are various feature extraction strategies to represent a protein chain, the choice of one or the other is not obvious, since there is no adequate strategy for all the data.

At the beginning, we made an empirical study of the most used MRPs in the work of the classification of RCPGs, this study is based on different data mining algorithms, implemented in the Weka environment. This proposal is carried out to analyze the diversity of the results of the precision and error rate of the classification from one method to another, where we concluded the influence of the correct choice of the MRP and the algorithm classification at a time.

Furthermore, in order to reduce the complexity and optimize the classification by avoiding the problem of combinatorial explosion, we proposed to use two bio-inspired meta-heuristics both for the selection of characteristics and for the choice of the best pair (strategy feature extraction (FES), data mining algorithm (DMA)).

We also opted to use the bat algorithm (Bat) to extract the relevant characteristics and the genetic algorithm (GA) to choose the best pair. We compared the results obtained with two existing algorithms. The experimental results indicate the efficiency of the proposed system.

Keywords: GPCR, Function Prediction, Feature Selection, BAT algorithm, Genetic Algorithm, Feature extraction Strategies, Classification, Data Mining Algorithms, protein representation.

Table des Matières

Table des Illustrations

Liste des Tableaux

Table des Abréviations

AA	Amino Acid (Acides Aminés)
AAC	Amino Acid Compisition (Composition en acides aminés)
ACO	Ant Colony Optimization (Optimisation par colonies de fourmis)
AFD	Algorithme de Fouille de données (Data Mining Algorithm DMA)
ADN	Acide Désoxyribonucléique
AG	Algorithme Génétique
Am-PseAAC	Amphiphilic Pseudo Amino Acid Composition
ARN	Acide Ribonucléique
ARNm	Acide Ribonucléique messager
BA	Bat Algorithm (Algorithme de chauves-souris)
BAG	Bagging
BN	Bayesien Network
BPSO	Binary Particule Swarm Optimization (optimisation de l'essaim de particules binaires)
DC	Dipeptide Composition
DPSO	Discrete Particule Swarm Optimization (optimisation des essaims de particules discrètes)
DT	Decision Table (Table de décision)
Fit	Fonction Fitness
FS	Feature Selection (Traduction de Sélection de caractéristiques)
GPCR	G-Protein Coupled Receptors (Traduction de RCPGs)
GPCRDB	G-Protein Coupled Receptor Data Base (Base de données de RCPGs)
KNN	k-Nearest Neighbors (k-plus proches voisins)
LB	Logit Boost
LD	Local Descriptors
MOA	Metaheuristic Optimization Algorithms
MRP	Méthode de représentation de protéines
NB	Naive Bayes
PseAAC	Pseudo Amino Acid Composition (Composition en pseudo acides aminés)
PSO	Particle Swarm Optimization (Optimisation par essaims particulaires)

RCPGs	Récepteurs Couplés aux Protéines G
RF	Random Forest
SVM	Support Vector Machine (Machine à vecteurs de support)
7TM	Seven Transmembranaire
Weka	Waikato Environment for Knowledge Analysis

Introduction générale

La bioinformatique est un domaine de recherche actif au cours des trois dernières décennies et attire continuellement l'attention des chercheurs en informatique et en biologie. Les objectifs de la bioinformatique étaient de stocker et de gérer les données biologiques et de développer des outils de calcul sophistiqués qui sont utiles dans l'analyse et la modélisation [LUS 01]. Le volume de données rassemblées dans le cadre du Projet de Génome Humain (Human Genome Project) [BEN 00] et de divers autres projets de séquençage réussis augmente de façon exponentielle, ce qui a soulevé de nombreux défis pour la communauté de recherche.

Les données se composent généralement d'acide désoxyribonucléique (ADN), d'acide ribonucléique (ARN) et de protéines. Les protéines constituent l'élément le plus fondamental de tout organisme vivant. Il comprend 20 acides aminés qui jouent un rôle important dans les fonctions cellulaires, notamment le transport des nutriments, la régulation du métabolisme et la construction musculaire. Une protéine peut adapter quatre types de conformations différents en raison de certains changements structurels afin d'exécuter des fonctions à l'intérieur de la cellule dans le corps humain [IQB 14]. Chaque protéine inconnue doit être annotée pour connaître sa structure et sa fonction, tandis que la vitesse des expériences in vitro est considérablement réduite à mesure que de plus en plus de nouvelles séquences sont ajoutées constamment dans les bases de données de protéines.

Cependant, les méthodes expérimentales rencontrent des difficultés pour annoter de nouvelles protéines car elles demandent beaucoup de travail et prennent beaucoup de temps.

À ce stade, trois approches sont couramment utilisées. La première approche fait la prédiction selon la similarité de séquence primaire (L'alignement de séquences). Il s'agit d'une approche largement utilisée en raison de la grande quantité de séquences découvertes. Cependant, il échoue parce que la structure primaire est la moins préservée dans l'aspect évolutif des structures. Cela signifie que les protéines peuvent avoir un degré élevé de similitude dans leurs chaînes, mais remplir des fonctions complètement différentes. La deuxième approche est liée à la structure tertiaire, qui est beaucoup plus préservée que la structure primaire. La fonction d'une protéine est directement liée à cette structure. Malgré cela, il a été observé que les similitudes structurelles ne correspondent pas toujours aux similitudes catalytiques. La troisième approche est basée sur l'utilisation de caractéristiques physico-chimiques pour représenter les acides aminés présents dans la structure primaire. Ils sont calculés sur la base de l'interaction de toutes les structures d'une protéine [PEA 04].

L'extraction de ces caractéristiques physico-chimiques nécessite une méthode de représentation de protéines. Cependant, il existe une multitude de stratégies différentes dans

leurs principes de fonctionnement et fournissent des vecteurs d'attributs numériques qui se diffèrent en terme taille et contenu. Plusieurs travaux ciblent la classification des protéines par cette approche [BHA 04c, GAO 06, SEC 07, NEM 09], et les résultats obtenus montrent son efficacité, mais il existe toujours quelques lacunes de ces stratégies qui limitent le fonctionnement des algorithmes de classification telle que: l'information des propriétés physico-chimiques contenue dans le vecteur de caractéristique, ainsi que la taille de ce dernier. Autres chercheurs visent à développer de nouvelle méthodes d'extraction de caractéristiques [REH 11] pour améliorer la classification des RCPG..

Quelques soit la MRP choisie, les vecteurs d'attributs produits peuvent comporter des informations inutiles, redondantes ou ambigües, tout simplement un bruit est toujours présenté dans les chaines numériques, ce qui peut diminuer la performance de l'algorithme de classification choisi et perturber l'apprentissage automatique. De plus, les caractéristiques de séquence non informatives ajoutent du bruit à la tâche de la classification et masquent les informations contenues dans les caractéristiques de discrimination.

Pour remédier ce problème, il faut essayer d'éliminer n'importe quelle information inutile présente dans le vecteur et préserver uniquement les caractéristiques pertinentes qui comportent des informations cruciales.

La sélection de caractéristiques (FS) est un processus qui sélectionne un sous-ensemble d'entités (c'est-à-dire des attributs ou des variables) à partir des entités d'origine. Cependant, il n'est pas possible de garantir que les caractéristiques sélectionnées par les techniques de FS sont le meilleur sous-ensemble de caractéristiques possible dans la base de données.

Il s'agit d'une méthode de prétraitement couramment utilisée pour préparer les données à une taille traitable avant de pouvoir être traitées par un algorithme de data mining. Les algorithmes sophistiqués d'exploration de données peuvent souvent échouer ou avoir des problèmes de calcul importants lorsqu'ils traitent directement avec un très grand nombre d'attributs [BER 08]. L'importance de la FS a été mentionnée dans plusieurs études [JEN 05, CHE 06, SAE 07, WAN 07]. Diverses méthodes de fouille de données ont été appliquées pour classer la fonction des protéines car elles ont l'avantage de découvrir des connaissances utiles à partir de séquences de protéines. Dans la bioinformatique, la connaissance de la fonction des protéines est un lien crucial dans le développement de nouveaux médicaments, et même le développement de produits biochimiques synthétiques tels que les biocarburants [PAN 06]. Récemment, il y a eu des progrès significatifs dans le développement de méthodes alternatives de prédiction fonctionnelle pour réduire la dépendance aux approches basées sur l'homologie.

Par conséquent, il est souhaitable d'explorer des méthodes supplémentaires qui prédisent la fonction des protéines indépendamment de la similitude des séquences. Parmi les approches qui méritent d'être exploitées, citons le data mining et l'apprentissage automatique en raison de leurs capacités à distinguer les protéines appartenant à différentes classes fonctionnelles. Cependant, les performances d'un algorithme de classification sont fortement influencées par les échantillons de représentation et nécessitent donc une attention particulière avant que l'extraction n'ait lieu.

Des études antérieures ont montré que des informations utiles sur la fonction sont contenues dans un spectre de caractéristiques de séquence protéique [JEN 02, KOT 07]. En biologie moléculaire, un flux massif de nouvelles données crée un problème sur la façon d'extraire des caractéristiques significatives. Pandey et coll. [PAN 06] identifié divers défis et tendances émergentes pour la recherche à venir en biologie computationnelle. Dans leur enquête, ils soulignent la nécessité d'explorer les domaines de l'exploration de données et de l'apprentissage automatique afin d'exploiter les données biologiques disponibles pour prédire la fonction des protéines de manière plus précise et plus efficace.

Etant donné la difficulté et l'importance des problèmes protéomiques posés, surtout en utilisant des données complexes, plusieurs approches ont été développées et exploitées pour y apporter des solutions efficaces et fiables en terme temps et mémoire, citons parmi eux : KNN, les réseaux de neurones et SVM. Malgré que ces méthodes ont montrées leurs preuves et ont données de bons résultats, les chercheurs se sont tournés, ces dernières années vers les méthodes bio-inspirées qui restent toujours le bon refuge de tous problèmes complexes peuvent engendrer une explosion combinatoire.

Ces méthodes s'inspirent de la nature, la physique et de la science de la vie, elles se basent essentiellement sur la biologie (corps humains, animaux, insectes) pour résoudre les problèmes computationnels. Elles connaissent un grand succès dans tous les domaines d'application.

Les Algorithmes Génétiques (AG) sont la famille la plus utilisée dans le domaine de la prédiction de la fonction des protéines pour résoudre diverses problèmes tel que: la sélection de caractéristiques [NAV 12, LEI 14, SAN 18]. De plus, les données de spectrométrie de masse ont été exploitées à l'aide d'un AG [JEF 04] pour produire des modèles discriminatoires qui distinguent les individus en bonne santé de ceux atteints de cancer.

Une autre méta-heuristiques bio-inspirée très efficace et prometteuse à cause de sa flexibilité, et sa capacité à contrôler itérativement les paramètres, est l'algorithme de chauves-souris (Bat Algorithm). Bien que cette approche donne de bons résultats dans divers

domaines d'application, elle n'est pas largement utilisée dans le domaine de la bioinformatique en général et spécialement dans la protéomique.

Problématique et Objectifs

L'identification de la fonction des RCPG est un domaine d'intérêt actuel dans la recherche pharmaceutique et biologique. Les médicaments actifs au niveau des RCPG ont des avantages thérapeutiques dans un large spectre de maladies humaines, sur les quelques 500 médicaments commercialisés cliniquement, plus de 30% sont des modulateurs de la fonction des RCPG, représentant environ 9% des ventes pharmaceutiques mondiales, ce qui fait les RCPG les plus efficaces de toutes les classes cibles en termes de découverte de médicaments [DRE 00, NGO 16]. Des efforts intenses ont été consacrés à l'identification de nouvelles fonctions RCPG pour les orphelins. Cependant, pour de nombreux RCPG, ces efforts n'ont pas abouti à des résultats fiables.

A ce stade, plusieurs questions ont été posées: quelles sont les étapes nécessaires pour une bonne identification de la fonction protéique? Quelle est la méthode de représentation de protéines (MRP) adéquate qui peut être utilisée pour extraire des caractéristiques pertinentes et construire des vecteurs d'attributs numériques? Quel algorithme d'exploration de données (AFD) doit être sélectionné pour effectuer une classification précise? Comment éviter l'explosion combinatoire des algorithmes de classification en raison de la nature complexe des données protéiques?

Bien que de nombreuses approches de prédiction de la fonction des RCPG aient été proposées, un grand nombre de RCPG sont encore orphelins. La méthode la plus ancienne est la recherche de la similarité des séquences dans les bases de données de protéines, cette technique est principalement basée sur l'alignement de séquences par paires comme l'outil BLAST [ZHA 12]. Mais il est difficile d'identifier les séquences de RCPG avec succès car il n'y a aucune similitude de séquence partagée significative. Cependant, deux protéines peuvent avoir des séquences très différentes et occupent une fonction similaire, ou avoir des séquences très similaires et remplir des fonctions différentes [NEM 09]. Pour résoudre ce problème, certaines approches statistiques et d'apprentissage automatique ont été développées pour la prédiction des RCPG [SEC 07].

Il existe trois problèmes majeurs dans la tâche de la prédiction computationnelle de la fonction protéique avec des algorithmes de classification, qui sont le choix de l'algorithme de classification et le choix de la méthode de représentation des protéines, ainsi que la sélection des attributs pertinents pour éviter le problème d'explosion combinatoire. Ce sont des

problèmes ouverts, même dans tout problème de classification car il existe des choix multiple et on ne sait pas lequel est le meilleur.

Généralement, il existe plusieurs stratégies pour extraire les attributs d'une séquence protéique, et le choix de la méthode de représentation protéique peut être aussi important que le choix de l'algorithme de classification, contrairement à peu de travaux [KIN 01] qui sont souvent négligés la stratégie d'extraction de caractéristiques utilisée et plus axée sur l'algorithme d'exploration de données à utiliser. D'autres chercheurs ont développé une stratégie d'extraction de caractéristiques hybrides [REH 11] qui peut combiner à la fois la stratégie de composition en pseudo-acides aminés (PseAAC) et la représentation énergétique multi-échelles en exploitant la capacité de discrimination dans les domaines spatial et de transformation pour la classification des RCPG, tandis que certains auteurs [SEC 10; NAV 12] ont comparé les précisions prédictives de quelques méthodes de représentation des protéines dans la classification des séquences.

La transformation de la chaîne protéique peut donner un énorme vecteur d'attributs numériques, la taille et les composantes de ce dernier, influencent fortement la précision prédictive et le taux d'erreur de la classification. Pour améliorer ces tarifs, il est strictement nécessaire d'éliminer les bruits «redondances ou informations inutiles» présents dans les exemples à classer. En outre, des ensembles de données avec des centaines et des milliers d'attributs peuvent causer la malédiction de la dimensionnalité (Curse of dimensionality) et des problèmes d'explosion combinatoire [CHE 14].

L'une des techniques les plus réalisables pour faire face à ce problème est la sélection de caractéristiques (FS) [SAE 07; BAG 16] pour optimiser le modèle de classification et améliorer les mesures de performance. Cette technique est largement utilisée dans différents domaines pour améliorer les résultats tels que: la catégorisation de texte [AGH 08], la reconnaissance faciale [KAN 08], la prédiction de la fonction des protéines [NEM 09] et elle est surtout utilisée dans le Big Data et l'exploration de données [Li 17; TUP 17].

Plusieurs chercheurs visent à optimiser la classification des RCPG, soit en sélectionnant des attributs et des classifieurs [SEC 10], soit en utilisant des méta-heuristiques bio-inspirées [NAV 12; HOL 08; HOL 09 ; NEM 09].

Dans ce travail, nous utiliserons deux approches bio-inspirées différentes qui montrent leur efficacité dans plusieurs domaines [NAY 18], la première est une méta-heuristique d'optimisation appelée l'algorithme de chauves-souris (Bat) introduite par Yang [YAN 10] et, est utilisée dans [YAN 12] pour l'optimisation de l'ingénierie globale, [GAN 13] pour les tâches d'optimisation contraintes, [CAI 19] pour l'optimisation à grande échelle et dans

[GUO 19] pour résoudre un problème d'optimisation de fonction globale. La seconde est une approche évolutive appelée l'algorithme génétique, qui a été utilisé pour un grand nombre d'applications de modélisation dans les domaines chimiques et biologiques [PED 96, JUD 97], de plus Judson a développé un AG pour la prédiction de la structure des protéines [JUD 08], et dans [SAN 19] les auteurs proposent une méthodologie utilisant une AG multi-objectifs pour la sélection de caractéristiques afin de prédire la fonction des protéines.

Dans notre travail, nous abordons trois problèmes: la sélection de la stratégie d'extraction de caractéristiques (FES), la sélection de l'algorithme d'exploration de données (DMA) et la sélection de caractéristiques pertinentes en utilisant deux algorithmes bio-inspirés différents qui sont: l'algorithme génétique (GA) pour le choix de meilleur couple (FES, DMA) et l'algorithme BAT pour FS pour améliorer la classification en terme précision de taux d'erreur afin d'optimiser la prédiction de la fonction des séquences de RCPG.

Contenu du manuscrit

Ce document est divisé en deux parties essentielles:

1. Etat de l'art.
2. Problèmes et propositions.

Il est organisé en six chapitres dont les trois premiers sont théoriques et comportent, respectivement, une introduction à la bioinformatique et des généralités et des concepts basiques des RCPG, une description des méthodes et concepts des approches bio-inspirées utilisées: les algorithmes génétiques et l'algorithme de chauve-souris (BAT) ainsi qu‘une étude des approches d'extraction et de sélection de caractéristiques. Les trois derniers chapitres concernent nos contributions ainsi les études expérimentales effectuées, qui résument une étude analytique des méthodes de représentation de protéines les plus utilisées dans la littérature, afin de constater leur influence aux résultats de la classification dans le troisième chapitre, et l'autre consiste en l'optimisation de la prédiction de la fonction de protéines en utilisant le choix de couple pertinent (MRP/AFD) par l'AG dans le chapitre quatre et une méthode de FS par l'algorithme de chauve-souris dans le cinquième et le dernier chapitre.

Nous avons fini par une conclusion, les futurs travaux et les perspectives.

Chapitre 1. ***Généralité sur la bioinformatique et les RCPG***

Dans ce chapitre, nous introduisons le domaine de la bioinformatique qui est multidisciplinaires, et utilise diverses disciplines pour résoudre les problèmes biologiques, son objectifs ainsi les bases et les sources d'informations biologiques. Par la suite nous

expliquons les techniques de data mining utilisées dans la protéomique qui découle de la génomique, ensuite nous présentons des exemples de classifications supervisée en bioinformatique, et nous parlons spécialement des travaux connexes à la prédiction de la fonction des protéines. En second lieu, nous discutons de la famille protéique des récepteurs couplés aux protéines G, qui sont la famille la plus importantes des protéines membranaires. Nous débutons par une définition des RCPG, leurs aspects biologiques, chimiques et pharmacologiques leur structure et leur mécanisme qui consiste en la transduction du signal et l'activation de la protéine G. Nous présentons en plus la classification la plus connue des RCPG.

Chapitre 2. ***Les Algorithmes Génétiques et l'algorithme de chauve-souris (BAT)***

Ce chapitre est partitionné en deux parties : la première est consacrée à la présentation des AG, son évolution biologique, le passage des facteurs aux code génétique et le cycle de vie naturel d'un individu, ensuite nous passons à l'évolution artificielle qui comporte cinq étapes nécessaires (Initialisation et terminaison, la sélection, la reproduction avec mutation, la survie et la convergence), nous les expliquons attentivement dans cette section. De plus, nous détaillons les différents domaines d'application des AG y inclus la bioinformatique, et nous finalisons cette partie par une étude des avantages et des inconvénients des AG.

Quant à la deuxième partie, qui représente la définition de l'algorithme de chauves-souris, ses concepts basiques, ses caractéristiques ainsi que les diverse domaines d'application.

Chapitre 3. ***Extraction et Sélection de caractéristiques.***

Nous abordons dans ce chapitre les deux paradigmes essentiels de notre étude, l'extraction de caractéristiques et la sélection de caractéristiques (FS). Nous présentons en premier lieu, les méthodes d'extraction de caractéristiques utilisées pour représenter une chaine alphabétique sous forme d'un vecteur d'attributs numériques, nous expliquons ces méthodes qui sont divisées en sept catégories, ainsi que les variantes de chacune. En second lieu, nous abordons la notion de la sélection de caractéristiques, sa définition, ses techniques d'évaluation, aussi nous présentons une revue de quelques méthodes de FS. Nous nous intéressons à identifier la pertinence d'une caractéristique et aussi à monter la grande utilisation de cette technique (FS) dans le domaine de la bioinformatique.

Chapitre 4. ***Etude analytique des stratégies d'extraction de caractéristiques.***

Ce chapitre est consacré à notre première contribution, qui consiste à faire une étude de différentes méthodes de représentation de protéines les plus utilisées dans le domaine

d'identification de la fonction des RCPG. Nous commençons ce chapitre par l'explication de notre contribution et notre proposition ainsi que toutes les étapes de sa réalisation. Après, plusieurs expérimentations ont été effectuées à l'aide de différents outils, ensuite une description détaillée de : la base de données utilisée dans ce travail, l'environnement Weka, l'outil Protr qui sert à nous fournir des vecteurs d'attributs des méthodes AAC, PseAAC, Am-PseAAC et LD. De plus, une illustration détaillée de la méthode LD que nous avons implémenté pour produire des vecteurs de 210 éléments numériques, nous terminerons cette section par la présentation des mesures de performance utilisées pour évaluer notre contribution. Les résultats expérimentaux de la classification des RCPG à trois niveaux, et leurs discussions sont présentés à la fin de ce chapitre. Ces résultats montrent l'atteinte de notre objectif, qui est de tester l'influence du choix d'une MRP ou d'une autre aux résultats de la classification.

Chapitre 5. ***Un algorithme évolutionnaire pour la sélection du meilleur couple (MRP/AFD)***

Ce chapitre présente notre deuxième contribution principale, portant sur la prédiction de la fonction des RCPG en proposant un algorithme génétique pour la sélection du meilleur couple (Méthode de Représentation de Protéines / Algorithme de Data Mining). En premier temps, nous débutons par l'exposition de l'architecture générale du système proposé en détaillant toutes les étapes de réalisation de notre proposition: le codage des individus, l'initialisation des paramètres, le calcul de la fonction fitness, la méthode utilisée pour la sélection, le croisement et la mutation. En seconde lieu, une étude expérimentale a été faite en présentant la base de données, l'étape de prétraitement qui est nécessaire pour améliorer la qualité de données utilisées, ensuite nous citons les outils intéressants à l'expérimentation, et le jeu de paramètres de l'AG, de plus une explication de la recherche de la meilleure solution, nous terminons par une discussion détaillée.

Chapitre 6. ***Sélection de caractéristiques par l'algorithme de chauves-souris pour la prédiction de la fonction des RCPG.***

Ce chapitre concerne notre troisième contribution principale, qui consiste en la proposition d'une approche bio-inspirée basée sur l'algorithme de chauves-souris pour la sélection de caractéristiques afin d'optimiser la classification des RCPG, nous entamons ce chapitre par la présentation d'un framework bio-inspiré pour la classification des séquences protéiques, ensuite, nous détaillons ces trois principaux modules et les étapes nécessaires à sa

réalisation, Différentes expérimentations ont été effectuées avec divers jeux de paramètres afin d'adapter, aux mieux, les algorithmes utilisés à la problématique abordée, une brève explication des outils et méthodes utilisés dans ces expérimentations, ainsi les résultats obtenus et une discussion détaillée. A la fin de ce chapitre, nous présentons une comparaison de notre proposition avec d'autres méthodes bio-inspirées utilisées pour la sélection de caractéristiques.

Première partie

État de l’art

Chapitre 1
Généralités sur La Bioinformatique et les RCPG

1.1. Introduction

La biologie moléculaire est entrée depuis 1995 dans l'ère de la génomique : on dispose maintenant de l'information génétique exhaustive sur un nombre croissant d'organismes vivants et il est aujourd'hui possible d'aborder de manière globale, un certain nombre de problèmes complexes dont on n'avait jusqu'à présent qu'une connaissance fragmentaire: voies métaboliques, interaction de la cellule avec l'extérieur, mécanismes globaux de régulation et de contrôle. Une nouvelle discipline est également née de la connaissance de ces séquences complètes de chromosomes: la génomique comparée. Il est maintenant possible de comparer deux organismes vivants à l'échelle de leur génome, de déterminer les gènes qu'ils ont en commun ou qui leur sont propres.

La bioinformatique est une discipline récente, traite différents aspects de ce nouveau champ de la connaissance et s'appuie bien sûr à la fois sur les concepts de la biologie et de l'informatique, et sur des outils issus de la chimie et de la physique. Elle se concentre surtout sur l'étude des séquences d'ADN et sur le repliement des protéines, donc elle travaille surtout au niveau moléculaire.

Chez l'homme, près de 865 gènes correspondent à des récepteurs couplés aux protéines G (GPCR). Il s'agit de la plus grande famille de récepteurs transmembranaires chez l'humain, représentant 3,4% du génome codant pour des protéines. Les GPCR sont impliqués dans la réception et l'analyse de stimuli extracellulaires. Ces récepteurs sont impliqués dans des processus aussi divers que la neurotransmission, le métabolisme cellulaire, la différenciation cellulaire, la sécrétion et la réponse inflammatoire. La diversité de ces stimuli activant les GPCR est étonnante. Ceux–ci peuvent être des hormones, des ions, des neurotransmetteurs et même des odeurs et de la lumière. L'activation d'un GPCR entraîne une réponse intracellulaire via une protéine G hétérotrimérique permettant l'adaptation de la cellule à son environnement. Des dérèglements dans la signalisation ou l'expression de ces récepteurs sont en causes dans de nombreuses maladies.

1.2.1 Les bases biologiques

Avant d'entamer notre problématique et les contributions associées, il est indispensable d'avoir des pré-requis biologiques nécessaires pour la compréhension et le développement des systèmes bioinformatiques, Dans notre étude, notre intérêt s'est porté sur la branche de la protéomique qui découle de la génomique.

Une cellule, généralement de 10 à 30 millionièmes de mètre (10 à 30 µm) de diamètre pour l'homme, contient de nombreuses structures spécialisées appelées organites (figure 1.1). La membrane cellulaire contrôle le passage des substances dans et hors de la cellule et enferme les organites cellulaires ainsi que les substances cellulaires.

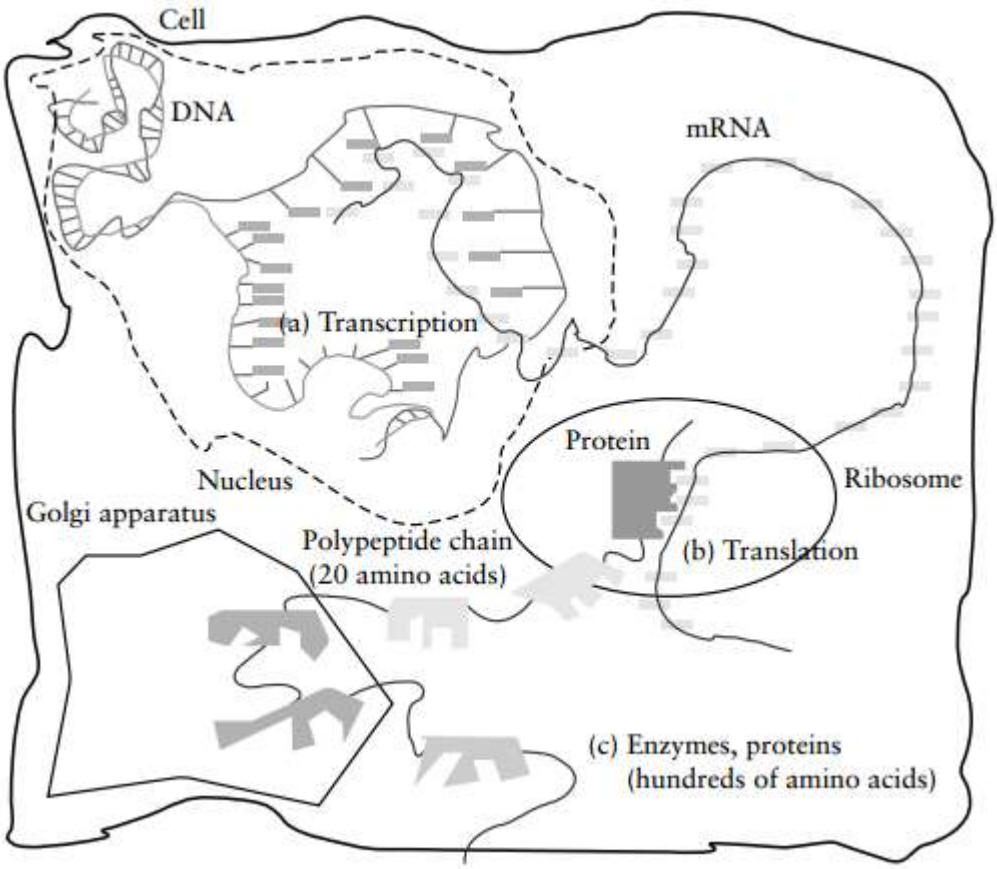


Figure 1.1: Un aperçu d'une cellule humaine typique [KEE 05].

Toutes les informations qui dirigent chaque fonction cellulaire sont stockées dans de grosses molécules d'**ADN** trouvées dans le noyau. Une cellule ne peut pas fonctionner sans ADN.

Les informations qu'il contient doivent en quelque sorte être mises à la disposition du reste de la cellule et être transmises à toutes les nouvelles cellules. Bien que chaque cellule contienne le complément complet d'ADN, grâce à un processus qui n'est pas encore clairement compris, certaines parties de l'ADN sont activées ou désactivées dans les cellules, ce qui entraîne différents types de cellules produisant différentes **protéines** pour une croissance et un fonctionnement normaux de l'organisme comme un ensemble.

Ce qui est montré sur la figure (1.1) est une cellule eucaryote, qui a un noyau lié à la membrane. Le corps humain compte environ 200 types différents de cellules eucaryotes [KEE 05]. **Le processus de transcription** (Figure 1.1 (a)) commence par l'ouverture de l'ADN double brin pour révéler les bases codant pour un gène. Une copie du gène est

fabriquée appelée ARN messager (ARNm) qui quitte le noyau. L'**ADN** double brin se ferme après la transcription. Au niveau des ribosomes, **le processus de traduction** commence (figure 1.1 (b)) par lequel trois bases copiées à la fois (codon) sont mappées sur un acide aminé. L'ARNm est brisé et peut rentrer dans le noyau pour une transcription supplémentaire de l'**ARNm**. La séquence croissante d'acides aminés (séquence polypeptidique) peut être modifiée par l'appareil de Golgi avant **la production finale d'enzymes, de protéines** et d'autres produits traduits (figure 1.1 (c)).

Dans ce qui suit, nous allons expliquer et détailler toutes les notions nécessaires à la synthèse des protéines qui conduit à la compréhension du fonctionnement des cellules vivantes.

1.2.1. ADN

L'ADN est un acronyme qui signifie Acide Désoxyribonucléique. Il s'agit d'une molécule présente chez tous les êtres vivants et qui porte l'information génétique nécessaire au développement et au fonctionnement de l'organisme. L'ADN du noyau prend la forme de grosses molécules appelées chromosomes, constituées de combinaisons de quatre types de nucléotides: adénine, guanine, thymine et cytosine (respectivement étiquetées «A», «G», «T» et «C»).

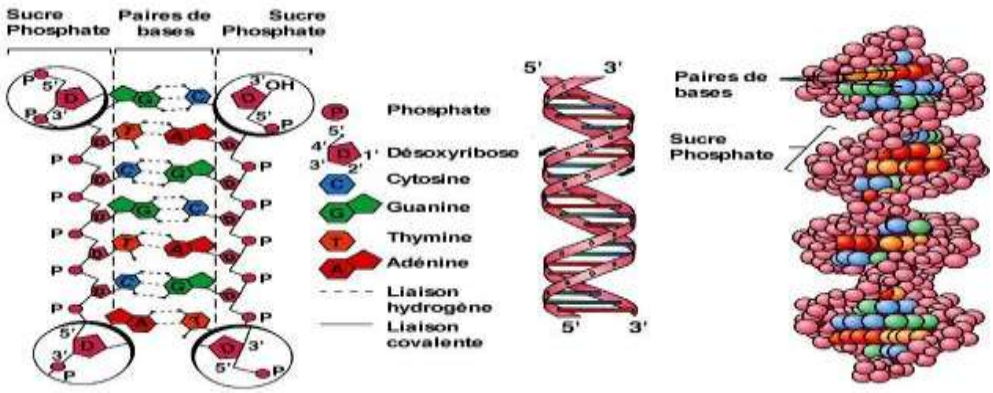


Figure 1.2: La structure en double hélice de l'ADN [1].

Un nucléotide d'ADN est formée par une présence d'un sucre pentose désoxyribose avec un acide phosphorique phosphate; plus une bases azotées.

Les 2 brins sont associes sur toute leur longueur de telle sorte que A est toujours en face de T et G en face de C .ces base apparient grâce a des liaisons H (hydrogène) trois liaisons H entre C et G deux liaisons H entre A et T : on dit que les 2 brins sont complémentaires. [ZEK 15].

Les deux brins sont anti -parallèles, car leur polarité est inversé, dans une double hélice d'ADN un brin est dans le sens 5->3 alors que le brin complémentaire est en 3->5.

1.2.2 ARN

L'ARN ou l'acides ribonucléiques est un polymère constitué d'un ensemble des molécules monocaténaires (simple brin), il est différent de l'ADN par la présence d'un sucre pentose ribose et d'une base uracile **(U)** à la place de thymine **(T)**. Il existe trois types d'ARN (figure 1.3) :

- **ARN ribosomiques ou ARNr : (82%)** Ce sont des constituant principaux des ribosomes ; lieu de synthèse des protéines.
- **ARN de transfert ou ARNt : (16%)** Ils sont constitués d'un brin d'ARN replié sur lui-même ; joue le rôle d'un transporteur et d'enzyme de synthèse des polypeptides.
- **ARN messager ou ARNm : (2%)** c'est leur niveau que le message (l'information génétique) est véhiculé du noyau vers le cytoplasme est plus précisément au niveau des ribosomes.

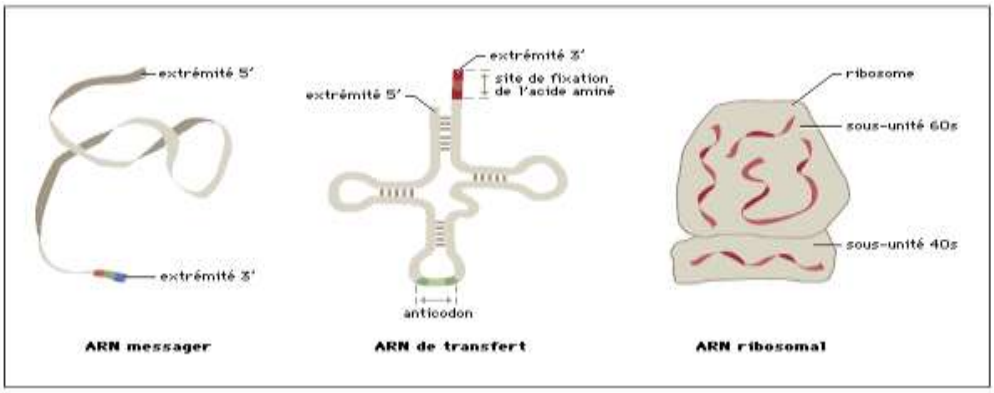


Figure 1.3: Différents types d'ARN [2].

1.2.3 Transcription

Le processus de transcription comprend trois étapes: l'initiation, l'allongement et la terminaison. La transcription commence par la double hélice en déroulant la figure 1.3 (a) et en exposant les bases qui représentent le début d'un gène. L'ARNm est ensuite formé, moyennant quoi une copie complémentaire du gène est réalisée. Puisque la transcription se déroule dans la direction 3 à 5, l'ARNm a une «polarité» opposée, c'est-à-dire que le début du gène se trouve maintenant à l'extrémité 5 de l'ARNm (figure 1.3 (b)), ou des parties du gène qui ne codent pas pour une protéine, sont éliminées, généralement par l'ARNm se repliant sur lui-même et formant des boucles qui sont coupées, laissant l'exonsin dans le transcrit. Ces transcrits contenant uniquement des exons peuvent être modifiés plus avant (figure 1.3 (c)) de sorte que des voies d'épissage alternatives pour le même gène soient

formées, c'est-à-dire qu'un gène peut donner lieu à de nombreux transcrits différents [KEE 05].

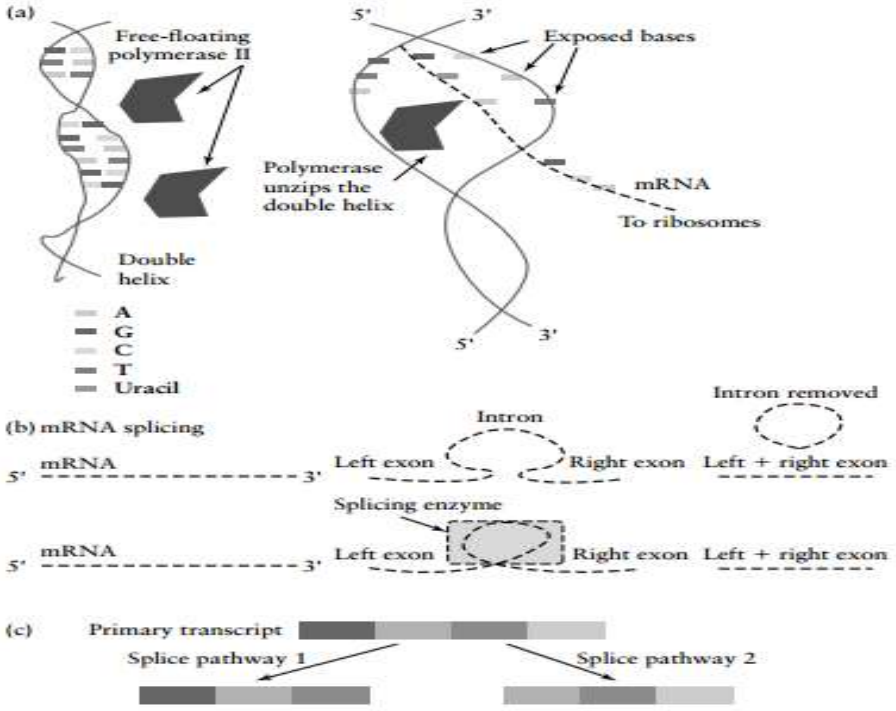


Figure 1. 4: Le processus de transcription. [KEE 05].

1.2.4. Les protéines

Les protéines ont été définies comme étant des macromolécules biologiques présentes dans toutes les cellules vivantes, elles sont composées par une ou plusieurs chaînes d'acides aminés, elles se répartissent en trois classes générales [DZI 10], sur la base de leur structure 3D globale et sur la base de leur rôle fonctionnel comme illustré dans le tableau (1.1). Alors, qu'arrive-t-il aux protéines / enzymes produites par le processus de transcription? Comme mentionné précédemment, les protéines remplissent de nombreuses fonctions vitales dans les organismes vivants. En tant que molécules structurelles, elles aident les cellules à former des tissus, les protéines transportent également des signaux d'une partie du corps à une autre ou d'une cellule à une autre, donc elles peuvent également agir comme un système de transport, transportant des molécules telles que l'oxygène dans le système circulatoire afin que toutes les cellules puissent avoir accès à cet élément important. Le système immunitaire humain dépend étroitement des protéines pour détecter l'arrivée d'agents pathogènes lorsqu'ils pénètrent dans le système humain et pour aider à monter une défense efficace du système immunitaire.

Après un certain temps, certaines réactions chimiques se produisent naturellement dans le tube à essai peuvent être notées. Il y aura un long délai car l'énergie d'activation nécessaire

pour démarrer une réaction chimique agit comme une barrière énergétique sur laquelle les molécules doivent être élevées pour qu'une réaction ait lieu. Une enzyme réduit efficacement l'énergie d'activation requise pour qu'une réaction se déroule. Une enzyme se verrouille sur une molécule, déclenche une réaction, puis est libérée inchangée.

Type	Description
Protéines Structurales	Protègent ou Soutiennent les structures cellulaires, les organes et les tissus
Enzymes	Facilitent les réactions chimiques
Protéines de transport	Transportent les molécules dans la cellule ou le corps
Hormones	Les messagers chimiques qui aident à réguler les organes

Tableau 1. 1: Types de protéines.

À partir de là, on peut voir que le processus de production d'enzymes / protéines, tel que déterminé par l'ADN, est absolument essentiel au bien-être continu d'un organisme, sinon les organismes en tant qu'êtres chimiques ne produiraient pas de réactions chimiques assez rapidement pour rester en vie (ex. respiration, digestion).

Selon la science biomoléculaire, ce que la vie signifie maintenant, c'est l'ensemble des gènes (ADN) qui codent pour la production de protéines appropriées qui augmentent le taux de réactions chimiques dans les cellules, où la nature et la vitesse des réactions sont déterminées par la nature des protéines. Les organismes d'une espèce particulière sont tous essentiellement les mêmes chimiquement; ce qui diffère, ce sont les enzymes produites par l'ADN hérité par leurs parents et d'autres facteurs (par exemple la mutation de bases et de gènes individuels par des moyens aléatoires). Ces enzymes contrôlent les processus cellulaires différemment pour différents membres de l'espèce, conduisant ainsi à des caractéristiques physiques différentes.

1.3. Objectifs de la bioinformatique

L'informatique peut contribuer de nombreuses manières à la recherche en biologie moléculaire et la pharmaceutique. En voici quelques exemples, pour donner une idée de la façon dont les systèmes informatiques peuvent être utiles en biologie et en conception de médicaments. [COH 04, RAO 08].

1. L'utilisation de la technologie informatique pour stocker des informations sur les séquences d'ADN et construire les séquences d'ADN correctes à partir de fragments identifiés par des enzymes de restriction (enzymes qui cassent l'ADN à certains points) a

été l'une des premières applications, découlant du projet du génome humain et d'autres projets avec séquençage de l'ADN de divers organismes. Alors que l'ADN d'un ensemble de 23 chromosomes pour un être humain est d'environ 3,5 gigaoctets, le H. le génome de la grippe n'est que de 1,9 Mbs, E. coli environ 4,6 Mbs et C.elegans environ 97 Mbs. Divers projets sont déjà en cours pour séquencer les génomes du poulet et du buffle, et ces projets, ainsi que plusieurs autres, entraîneront d'énormes besoins en matière de stockage et d'accès aux données.

2. Une fois que les séquences du génome sont stockées et consultées, il est nécessaire d'effectuer une analyse génomique comparative entre les bases de données afin d'étudier l'organisation et l'évolution des génomes. De telles analyses peuvent révéler des relations entre les organismes modèles, les cultures, les animaux domestiques et les humains. Des outils et des techniques de visualisation sont nécessaires pour réaliser ces analyses.
3. Les grandes bases de données doivent être structurées et organisées en utilisant une «ontologie» commune, ou un ensemble de termes qui sont structurellement liés les uns aux autres, afin que les chercheurs puissent accéder aux données de différentes bases de données en utilisant le même «langage de requête». The Gene Ontology Consortium [KEE 05] a produit des vocabulaires contrôlés pour décrire les gènes et les protéines qui, on l'espère, seront utilisés par tous les bio-informaticiens afin qu'une manière commune de se référer aux gènes et à leurs produits émerge.
4. De nombreux domaines de la biologie s'appuient sur des images pour communiquer leurs résultats. Des outils et des techniques sont nécessaires pour rechercher, décrire, manipuler et analyser les caractéristiques de ces images.
5. Une fois les bases de données des génomes créées, il est nécessaire de les maintenir et de vérifier que leur contenu est sans erreur et valide à mesure que les chercheurs ajoutent de nouvelles informations. Les anomalies doivent être identifiées et des mesures doivent être prises pour s'assurer que les bases de données sont aussi cohérentes que possible.
6. Des séquences protéiques sont ajoutées aux bases de données protéiques, et bien que celles-ci ne se développent pas aussi rapidement que les bases de données génomiques, il est nécessaire de stocker les séquences protéiques et leur structure ainsi que leur fonction. Même si un vocabulaire commun pour décrire les protéines est accepté, il existe un besoin majeur de lier les séquences protéiques avec leurs séquences sources d'ADN, étant donné les problèmes d'introns et d'ADN non codant. Il existe également un besoin d'outils permettant de prédire la structure d'une protéine à partir de sa séquence d'acides aminés

1.4. Techniques de data mining en protéomiques

Les protéines sont le résultat final de la traduction de l'ARNm par les ribosomes. Une fois que les séquences protéiques d'acides aminés quittent les ribosomes, elles se replient de manière complexe pour atteindre un état ou une conformation «natif» dans la cellule. L'état natif d'une protéine est une structure tridimensionnelle très stable qui aide à déterminer sa fonction biologique. En d'autres termes, une protéine ne peut fonctionner que si elle se replie de la bonne manière. Par exemple, les protéines catalytiques doivent se replier de manière à pouvoir se verrouiller sur une autre molécule. La détermination de la manière dont les protéines se replient dans des formes spécifiques s'appelle le «problème du repliement des protéines». Alors qu'une utilisation évidente des ordinateurs en bioinformatique est le stockage des informations de séquence d'ADN et la construction des séquences d'ADN correctes à partir de fragments identifiés par des enzymes de restriction (enzymes qui cassent l'ADN à certains points), des séquences protéiques et des séquences polypeptidiques qui composent cette protéine doivent également être stockés. De nouvelles séquences protéiques sont ajoutées aux bases de données protéiques à la suite de l'analyse des séquences d'ARNm, où l'ADN transcrit de manière redondante (introns) a été supprimé, et en traduisant les codons via le code génétique en lettres de l'alphabet des acides aminés. Cependant, ces séquences linéaires d'AA (séquence polypeptidique) ne nous disent rien sur la structure de la protéine ni sur son repliement. Le problème du repliement des protéines est important car il faut beaucoup d'efforts pour déterminer la structure d'une protéine réelle. Une vraie protéine doit être dénaturée (dépliée) pour que sa séquence d'acides aminés puisse être décrite.

1.4.1. Structure d'une protéine

La structure d'une protéine réelle est classiquement décrite de quatre manières (figure 1.5). La structure primaire d'une protéine (figure 1.5 (a)) est la séquence d'acides aminés produits au niveau des ribosomes. Puisqu'il y a 20 acides aminés, la structure primaire décrit l'ordre précis des acides aminés dans la protéine. La structure secondaire d'une protéine (figure 1.5 (b)) décrit les parties de la structure primaire (sous-séquences d'acides aminés) qui se replient en motifs réguliers et répétés, tels que des hélices α, des feuillets β ou des tours.
La structure tertiaire (figure 1.5 (c)) se compose des éléments de la structure secondaire qui construisent des unités plus complexes, telles qu'un motif $\alpha - \beta$, et fournissent une forme tridimensionnelle de la protéine. La structure tertiaire des enzymes est typiquement une forme compacte et globulaire, par exemple. Enfin, de nombreuses protéines sont constituées

de plus d'une chaîne polypeptidique. La structure quaternaire d'une protéine (figure 1.5 (d)) est une description de la manière dont plusieurs séquences polypeptidiques séparées se sont réunies pour former une protéine complexe.

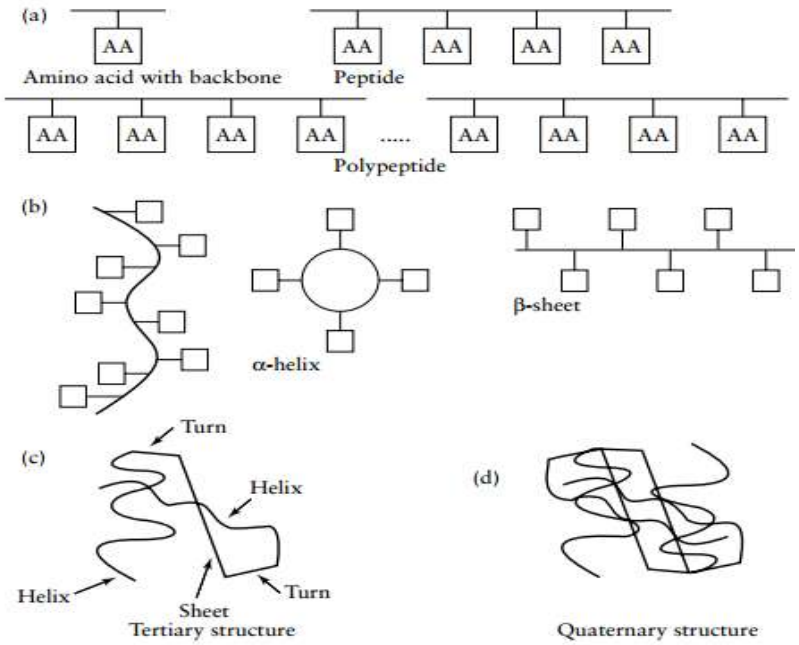


Figure 1.5: Structure des protéines [KEE 05].

1.4.2. Exemples de classification supervisée en bioinformatique

Dans les études bioinformatiques, la classification supervisée avec des variables d'entrée de haute dimension est fréquemment rencontrée [MA 08]. Des exemples se présentent couramment dans les études génomiques, épigénétiques et protéomiques.

- **Génomique: classification du cancer à l'aide de puces à ADN**

Le cancer est une maladie génétique qui peut être causée par des mutations ou des défauts de gènes. La technologie des puces à ADN permet d'étudier le génome à l'échelle mondiale. Des expériences d'expression génique sur des puces à ADN ont été menées pour identifier des bio-marqueurs dans les cancers, y compris le côlon, la prostate, le sein, la tête et le cou, la peau, le lymphome et bien d'autres. Il a été démontré que les régulations à la hausse ou à la baisse d'un sous-ensemble de gènes sont associés au développement du cancer. De nombreuses études sur les puces à ADN cancéreuses ont des phénotypes catégoriques d'intérêt - tels que la survenue du cancer, les stades ou les sous-types - ce qui conduit naturellement à une classification supervisée. Nous renvoyons à [GOL 99, WES 01] pour des exemples représentatifs.

- **Protéomique: classification du cancer par spectrométrie de masse**

En plus d'avoir des causes génétiques, le cancer peut également être considéré comme une maladie épigénétique. La régulation par la génétique implique une modification de la séquence d'ADN, tandis que la régulation épigénétique implique une modification de la structure de la chromatine et la méthylation de la région du promoteur. Des mesures épigénétiques, telles que les modèles de méthylation de l'ADN, ont été utilisées pour la classification du cancer. Des exemples incluent [PIY 02, ZUK 04] et leurs références

- **Protéomique: classification du cancer par spectrométrie de masse**

La spectrométrie de masse est une technique analytique qui mesure le rapport masse / charge des ions. Il est généralement utilisé pour trouver la composition d'un échantillon physique. Certains cancers affectent la concentration de certaines molécules dans le sang, ce qui permet un diagnostic précoce en analysant le spectre de masse sanguine.

Les caractéristiques mesurées avec des spectres de masse, souvent des statistiques sommaires des pics (par exemple, les contrastes de probabilité de pic dans [TIB 04]) - peuvent être utilisées pour discriminer les individus présentant différents phénotypes de cancer. Les chercheurs ont utilisé les spectres de masse pour la détection des cancers de la prostate, des ovaires, du sein, de la vessie, du pancréas, des reins, du foie et du côlon. Voir [YAN 04, DIA 05] pour des exemples.

- **Protéoméque: identification et classification des marqueurs protéiques**

La protéine est codée par des gènes et représentée par une séquence d'acides aminés. L'un des problèmes centraux de la bioinformatique est la classification des séquences protéiques en familles fonctionnelles et structurelles basées sur l'homologie des séquences. Il est généralement facile de séquencer les protéines, mais difficile d'obtenir une structure. Une solution analytique consiste à utiliser des techniques statistiques et à classer les données de séquence protéique en familles et superfamilles définies par des relations structure / fonction [LES 04, WES 04, WAN 06].

- **Protéomique: localisation des protéines**

La localisation subcellulaire d'une protéine est l'un des caractères fonctionnels clés car les protéines doivent être localisées correctement au niveau subcellulaire pour avoir des fonctions biologiques normales. L'imagerie automatisée des emplacements et des structures subcellulaires peut fournir la capacité de détecter des anomalies et de les associer à des protéines spécifiques. La prédiction de la localisation subcellulaire des protéines est une

composante importante de la prédiction basée sur la bioinformatique de la fonction protéique et de l'annotation du génome, et elle peut faciliter l'identification des cibles médicamenteuses [PAR 03, REY 05, YU 06]

1.4.3. Prédiction de la fonction des protéines

La découverte de la fonction protéique est un axe majeur de recherche en génomique, puisque les processus biologiques sont activés par ces molécules [ALB 02] Par exemple, l'hémoglobine est une protéine qui transporte l'oxygène dans le sang. La connaissance de la séquence protéique est également importante pour déterminer les anomalies pathogènes. La modification d'un acide aminé peut dans certains cas avoir des conséquences néfastes. L'anémie falciforme est une maladie caractérisée par une hémoglobine altérée. Cette mutation conduit à de nombreux symptômes: anémies, infections et risques d'accidents cérébro-vasculaires. Cette maladie, qui touche 50 millions de personnes dans le monde, est causée par la modification d'un seul acide aminé. Par conséquent, la connaissance des fonctions protéiques est très importante en sciences biomédicales, non seulement pour une meilleure compréhension de la biologie cellulaire en général, mais aussi parce que de nombreuses maladies sont causées par ou au moins associées à des défauts des fonctions protéiques [SIL 11]. Par conséquent, une méthode efficace de prédiction des fonctions protéiques peut potentiellement contribuer à générer de nouvelles connaissances biologiques qui peuvent conduire à un meilleur traitement et diagnostic des maladies, à la conception de médicaments plus efficaces, etc.

Bien que de nombreuses approches de prédiction GPCR aient été proposées au cours des deux dernières décennies, dans de nombreux cas, les techniques bioinformatiques conventionnelles, telles que l'alignement de séquences par paires ou en comparant des séquences à des motifs sont sans aucun doute valides, elles ne peuvent pas être optimales pour identifier le GPCR.

Premièrement, la séquence d'une superfamille de GPCR varie en longueur (entre 290 et 1 200 acides aminés), ce qui signifie que de nombreuses sous-familles ne peuvent pas être alignées efficacement. Le calcul sophistiqué est donc une approche plus efficace du problème de la classification GPCR, en utilisant des techniques basées sur l'exploration de données et l'apprentissage automatique.

Les méthodes de classification utilisées pour identifier et prédire la fonction GPCR sont nombreuses et variées. On peut distinguer deux approches essentielles: les méthodes basées sur la classification hiérarchique avec leurs deux types (arbre ou Graphe acyclique dirigé

(DAG)) [SEC 07, FRE 07, COS 08, SEC 10; SIL 11, NAK 17] et des méthodes basées sur la classification standard (classification plate) comme l'algorithme C4.5 utilisé dans [HUA 04], HMM [MUN 17], Classificateur SVM [KUM 10, SHR 10], les réseaux de neurones artificiels ont été utilisés dans [SEL 05] pour prédire les ligands GPCR.

Des outils en ligne ont également été développés. Par exemple, GPCRPred [BHA 04c] basée sur la méthode SVM pour la prédiction de deux niveaux pour les GPCR, familles et sous-familles, sur la base de la composition en dipeptide de la séquence primaire d'acides aminés, PRED-GPCR [PAP 04] fournit un complément à l'existant serveurs d'analyse de bases de données de modèles et potentiellement un outil de calcul pour la classification des familles GPCR [ZEK 11], GPCRTree est un serveur Web de classifications hiérarchiques en ligne [DAV 08]. C'est le premier serveur à implémenter une représentation indépendante de l'alignement des séquences de protéines et est également le premier à classer les séquences en utilisant une classification hiérarchique, PCA-GPCR [PEN 10], et le meilleur service Web est GPCR-MPredictor [NAV 12] qui repose sur une approche évolutive et prédisent le GPCR à cinq niveaux.

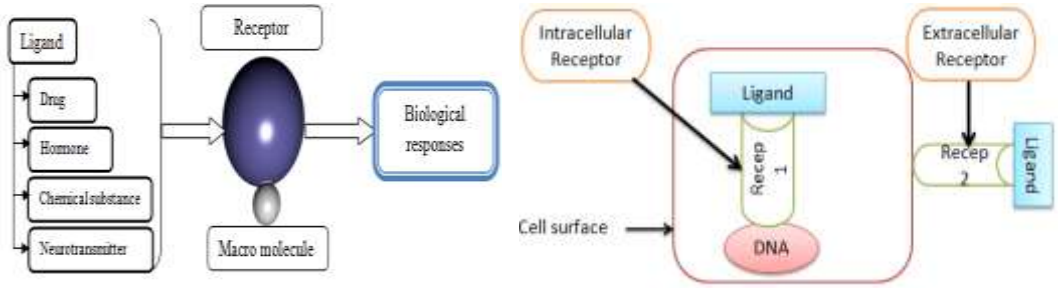
Une catégorie supplémentaire de méthodes consistant en les approches bio-inspirées est utilisée dans le domaine de la biologie moléculaire. Dans [SEC 09], les auteurs proposent un système immunitaire artificiel (SIA) qui résout le problème du clustering pour trouver le groupement optimal d'acides aminés pour le type de protéine. Cette méthode atteint une précision de 72,75% au troisième niveau, ce résultat marque une bonne amélioration par rapport à l'approche top-down [SEC 07], qui a fourni une précision prédictive de 70,46%. [GU 09] utilisent une approche Swarm Intelligence qu'ils exploitent, plus précisément, l'algorithme d'optimisation de l'essaim de particules binaires (BPSO) qui a une meilleure optimisation des performances sur les variables binaires discrètes (98,02%) que le PSO. En outre, il existe des travaux de Holden et Freitas [HOL 06] utilisant une approche bio-inspirée pour la classification hiérarchique, en particulier, l'algorithme hybride PSO/ACO qui a fourni une précision de 89,64%. Dans [COR 07], les auteurs utilisent également l'optimisation des essaims de particules discrètes (DPSO) pour la tâche de sélection des attributs et pour l'application à un ensemble de données complexes de classification fonctionnelle des protéines. Les résultats montrent que la sélection d'attributs offre de meilleures performances que si nous utilisions tout l'ensemble d'attributs.

1.5. Récepteurs Couplés au Protéine G "RCPG"

1.5.1. Définition des RCPG

En regardant le terme RCPG, nous pouvons extraire qu'il se compose de deux concepts essentiels qui sont: les récepteurs et la protéine G. Avant de définir le concept de RCPG, nous allons donner un aperçu rapide de ces deux notions.

- Récepteur: C'est une sorte de protéine cible [BUL 15] (enzyme, canal ionique; support, récepteur) utilisée pour s'accrocher au ligand qui peut être un médicament, une hormone, un neurotransmetteur ou une substance chimique pour donner les actions souhaitables en tant que site de reconnaissance. Les figures (1.7) (A) et (B) montrent le principe de fonctionnement des récepteurs ainsi leur emplacement dans la cellule.
- Protéines G, ce qu'on appelle parce qu'elles lient le GDP et le GTP. Ils sont hétérotrimériques [KIM 17], du fait de la constitution de trois sous-unités différentes qui sont: Gα, Gβ et Gγ. Il permet le transfert d'informations à l'intérieur de la cellule. Il participe ainsi à un mécanisme appelé transduction du signal. Pour effectuer leur fonctionnement, ils doivent être liés aux récepteurs comme illustré dans la figure (1.8) [DUC 17].



(A) (B)

Figure 1.6: (A): Principe de fonctionnement des récepteurs (B): Site des récepteurs

Les récepteurs représentent donc l'interface de la cellule avec son milieu extérieur et jouent un rôle crucial dans le maintien de l'homéostasie. Ils existent plusieurs types de récepteurs membranaires différenciés en fonction de leur structure et de leur mode d' action, notamment les récepteurs de type canaux ioniques (récepteurs ionotropiques) et les récepteurs associés à une activité enzymatique (récepteurs métabotropiques). Les récepteurs métabotropes sont eux-mêmes subdivisés en deux grandes classes de récepteurs selon le type d'enzymes associées, les récepteurs tyrosine kinases et les récepteurs couplés aux protéines G. Dans

notre travail, nous nous intéresserons spécifiquement aux récepteurs couplés aux protéines G (RCPG) aussi appelés récepteurs à sept domaines transmembranaires (7TM).

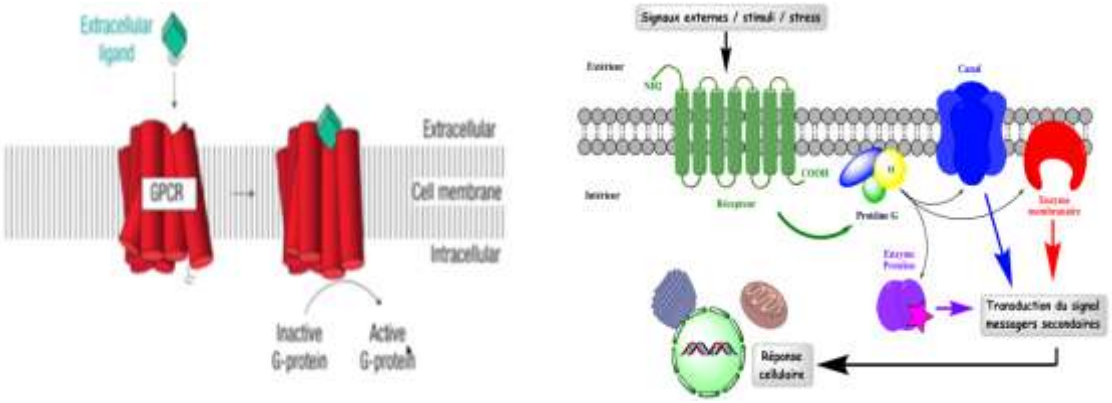


Figure 1.7: Mécanisme de signalisation de la protéine G hétérotrimérique [3].

1.5.2. Aspects biologiques

Les RCPG forment une superfamille de récepteurs présents sur les membranes cellulaires. Leur fonction générique est la transduction de signaux de l'extérieur vers l'intérieur de la cellule. Ils sont présents chez toutes les espèces animales ainsi que chez les bactéries et les champignons, mais dans ce mémoire nous ne nous intéressons qu'aux récepteurs humains. Les RCPG interviennent dans de nombreux processus physiologiques, comme la vision, l'olfaction, la régulation hormonale, la croissance cellulaire. L'analyse des résultats du séquençage du génome humain prédit l'existence de quelques 1000 RCPG, dont 400 sont des récepteurs non-olfactifs qui ne sont pas liés au mécanisme de l'olfaction. Sur 30 000 protéines prédites, les RCPG représentent 3%, pour cette seule superfamille.

La structure des RCPG est caractérisée par un domaine transmembranaire formé de 7 hélices α hydrophobes reliées entre elles par des boucles intra et extracellulaires. Les longueurs de ces boucles sont très variables en fonction des récepteurs, allant de quelques résidus à plusieurs centaines. On abrège souvent le nom de ce domaine par 7TM, et on utilise TM pour parler d'une des hélices α. Le domaine transmembranaire forme une cavité qui est le site actif d'une grande partie des RCPG, notamment ceux dont les boucles extracellulaires sont relativement courtes. Pour ceux chez qui ces boucles sont longues, ces dernières peuvent servir de site actif.

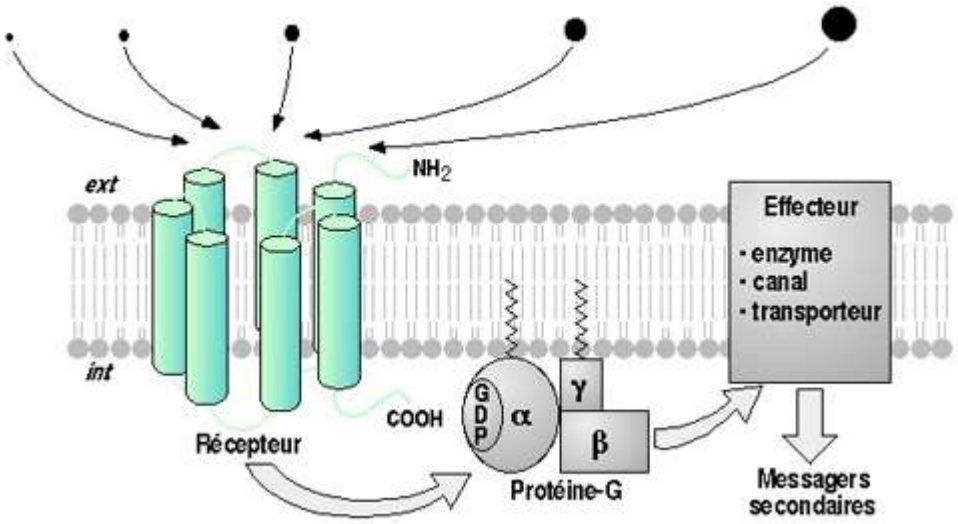


Figure 1. 8: Vue schématique d'un RCPG [4].

Seule la rhodopsine bovine a été résolue par cristallographie à l'heure actuelle, par l'´equipe de Palczewski en 2000, grâce à la diffraction par rayons X. Le récepteur est complexé avec son ligand par liaison covalente. Cette rare information structurale en est d'autant plus précieuse et sert de point de départ à plusieurs modélisations par homologie des RCPG.

Les RCPG sont aussi nombreux que diversifiés. Cette diversité apparaît dans les processus physiologiques qu'ils régulent, mais aussi dans les ligands qu'ils reconnaissent. Ainsi, leurs classifications reflètent cette diversité.

1.5.3. Aspects chimiques

A partir d'une même structure de base, les RCPG peuvent être actives par une grande ` diversité de ligands : ions de faible masse moléculaire (Ca2+), amines biogènes (dopamine, sérotonine), nucléoside et nucléotides (adénosine, ATP), peptides et hormones peptidiques (chimiokines, glucagon), lipides (sphingolipides, prostaglandines), mais aussi molécules olfactives et gustatives exogènes, et enfin le rétinal, cas particulier car li´e de façon covalente à son récepteur, et active par la lumière . Le schéma de la figure (1.10) présente quelques sites de fixation pour ces ligands, souvent dans la cavité transmembranaire, mais aussi au niveau des boucles extracellulaires.

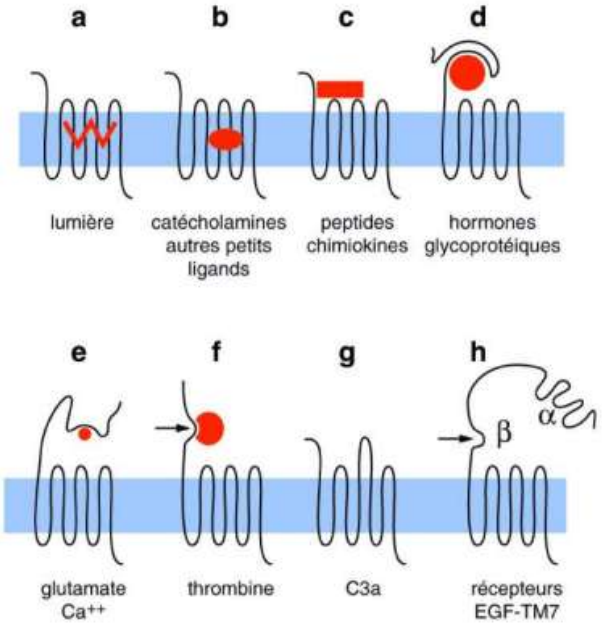


Figure 1.9: Schéma de quelques sites de fixation de ligands [5].

Lors d'une liaison entre un ligand agoniste et un récepteur, ce dernier est actif par la liaison, c'est-à-dire qu'il va changer de conformation par rapport à sa forme non liée au ligand. Ce changement conformationnel va impliquer une interaction avec une protéine G, liée au récepteur au niveau de son interface intracellulaire. Plus précisément, la protéine G, un hétérotrimère, va être scindée en deux parties et va en libérer une dans la cellule, débutant une cascade complexe d'´ev´enements produisant de nombreux messagers secondaires. La rhodopsine est un récepteur un peu particulier, car son ligand, le rétinal, lui est li´e de façon covalente. L'activation de la rhodopsine provient d'un changement conformationnel du rétinal qui, lorsqu'il est frappe par un photon, passe d'une forme cis à une forme trans. Le changement de conformation des RCPG de type rhodopsine fait intervenir un déplacement des hélices 3, 5, 6 et 7.

Les RCPG peuvent être présents sous forme non activée, sous forme activée avec un ligand, ou encore pour certains d'entre eux sous forme activée sans ligand (on parle alors d'activité intrinsèque ou constitutive). Toutes ces formes sont en ´équilibre et l'activité globale dépend de la proportion de ces formes en plus de l'activité des récepteurs, qui peut encore ˆêtre modifiée par des effecteurs allostériques ou par multimerisation des récepteurs.

1.5.4. Aspects pharmacologiques

Les RCPG sont distribues sur la surface de nombreuses cellules et régulent un grand nombre de processus physiologiques, comme la vision (sous-famille des opsines), l'olfaction (sous-famille des olfactifs), la croissance et la prolifération cellulaire (famille frizzled), l'adhésion

cellulaire (famille adhésion), la relaxation et la contraction des muscles lisses (sous-famille des prostanoıdes), la vasoconstriction (sous-famille des vasopeptides), et aussi la neurotransmission, les mécanismes hormonaux, des mécanismes inflammatoires, etc.

Chacun de ces processus physiologiques peut fonctionner de façon anormale. Les RCPG qui régulent ces processus interviennent donc dans la pathologie, soit parce qu'ils sont surexprimés ou sous-exprimés, soit qu'ils sont suractivés ou bloques. Dans le cas d'une surexpression ou d'une suractivité, il faut essayer de bloquer l'activité des récepteurs grâce à des ligands antagonistes ou agonistes inverses. Les ligands antagonistes se logent dans le site actif sans provoquer l'activation du récepteur, mais en prenant la place des ligands naturels ainsi empêchés ; l'activité globale est diminuée. Les ligands agonistes inverses sont utiles dans le cas ou le récepteur possède une activité constitutive sans être complexé avec un ligand. En se liant avec le récepteur (comme un ligand agoniste), ils vont bloquer son activité, d'ou le qualificatif d'inverse. L'activité globale est encore une fois diminuée. Dans le cas d'une sous-expression ou d'une sous-activité, il faut stimuler l'activité par des ligands agonistes, qui vont provoquer l'activation des récepteurs, ou supprimer une molécule inhibitrice. Les RCPG forment donc une classe de cibles très attractives pour les interventions thérapeutiques. Voici quelques chiffres pour donner un ordre d'idée de leur succès auprès des industries pharmaceutiques: 30% des médicaments les plus vendus modulent l'activité des RCPG; d'autres sources vont jusqu'à 60% des nouveaux médicaments mis sur le marché. Leur intérêt pharmacologique est triple : tout d'abord il y a les récepteurs déjà ciblés mais pour lesquels on peut améliorer la sélectivité et l'activité des médicaments ; ensuite il y a les récepteurs dont on connait des ligands endogènes, mais qui ne sont pas encore pris pour cible par des médicaments, et ils représentent la grande majorité(à l'heure actuelle, seuls 40 RCPG sont ciblés par des médicaments, sur les 1000 prévus par l'analyse du génome humain) ; enfin il existe environ 100 RCPG orphelins, dont on ne connait aucun ligand, et qui présentent un intéret potentiel à plus long terme

1.5.5. Structure des RCPG

Les RCPGs sont des protéines membranaires ayant une structure à sept domaines transmembranaires (TM). Les 7 TM sont des hélices α reliées par trois boucles intracellulaires (i1 à i3) et trois boucles extracellulaires (e1 à e3) (Figure 1.11).

Un domaine en α-hélice supplémentaire appelé hélice 8 (ou TM8) se situe après le TM7 et son orientation est parallèle à la surface interne de la membrane (Figure 1.10).

Deux cystéines sont souvent présentes après cette hélice. Elles sont généralement associées a un palmitate ce qui permet un ancrage supplémentaire à la membrane (Figure 1.10).

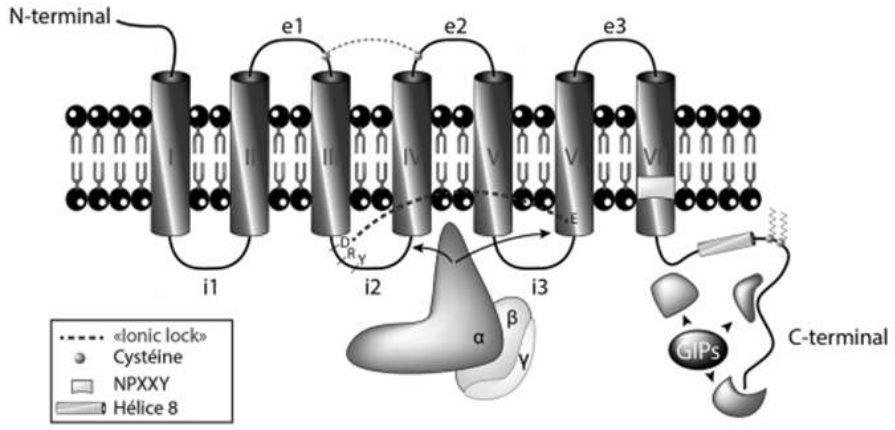


Figure 1.10: Structure générale des RCPGs [ZEK 15].

Les RCPGs ont une origine évolutive commune. Le succès évolutif de ces protéines a été considérable, le « bricolage évolutif » a généré des structures capables de reconnaître des messages très différents tels les photons, des petites molécules comme la sérotonine ou des grosses protéines comme les hormones glycoprotéine. Plus de mille des RCPGs ont été recensés dans les génomes d'êtres multicellulaires aussi évolutivement distants que Caenorhabitis élégants, les rongeurs ou l'homme. Chez la levure, deux types de RCPGs existent, assurant d'une part la reconnaissance du glucose (Gpr1) dont une fonction est de contrôler leur croissance et d'autre part celle des phéromones (Ste2-Ste3) dont la fonction est la reproduction sexuée : nourriture et sexe déjà. Chez l'homme, les récepteurs aux odeurs représentent cinq cents entités et les « endo-RCPGs » (récepteurs ayant un ligand endogène assurant la communication intercellulaire) environ trois cent soixante entités, soit au total 3 % du génome. Il existe encore quelques dizaines de récepteurs « orphelins » pour lesquels on ne connaît pas le ligand naturel. La déorphanisation de ces récepteurs est très importante car c'est souvent l'occasion de découvrir de nouvelles régulations physiologiques mais aussi de permettre la recherche de médicaments agissant sur ces récepteurs. Ces espoirs sont légitimes car les RCPGs sont la cible le 30-40 % des médicaments efficaces pour traiter les pathologies humaines, représentant 9 % des ventes. Quelques exemples de pathologies et de médicaments illustrent aisément ce point : douleurs (morphine), maladies mentales (antipsychotiques), hypertension (anti-angiotensine, β-bloquants), ulcères gastriques (antihistaminiques H2), migraines (inhibiteurs des récepteurs de la sérotonine 5-HT1D /1B, etc.

1.5.6. Transduction du signal par les protéines G

La protéine G est une protéine régulatrice qui active des enzymes qui elles-mêmes mobilisent des molécules « seconds messagers » pour activer d'autres enzymes ou canaux. Les protéines G interviennent dans les mécanismes de signalisation intracellulaire (ou de transduction du signal).

La protéine G est une GTPase ; elle est aussi appelée « protéine liant le GTP ». Lorsque le GDP est remplacé par un GTP, la protéine change de conformation. Le GTP est hydrolysé en GDP + phosphate inorganique (Pi). Le changement de conformation de la protéine lui permet d'interagir avec les protéines effectrices de la cascade de transduction du signal.

Les premières données conduisant à déterminer quelle protéine G est couplée à un RCPG donné, ont été obtenues au moyen d'expériences utilisant des toxines. Il a ainsi été rapporté que la protéine Gαs est sensible à la toxine cholérique (CTX) qui la bloque dans un état actif suite à l'ADP-ribosylation d'un résidu arginine du site liant le nucléotide guanlylique [GIL 78]. Parallèlement la protéine Gαi est inhibée par la toxine pertussique (PTX) qui, par ADP-ribosylation d'un résidu cystéine proche du site de liaison du récepteur, la bloque dans un état inactif [BOK 84].

Actuellement, nous disposons de diverses techniques pour déterminer la spécificité de couplage des protéines G aux récepteurs, telles que la co-immunoprécipitation, la mutagénèse dirigée, le BRET (Bioluminescence Resonance Energy Transfer) ou encore l'utilisation de modèles cellulaires ou animaux «knockout» pour certaines protéines ou sous-unités. Leurs utilisations font ressortir l'idée qu'un RCPG est capable d'interagir avec plusieurs types de protéine G [HER 03]. Un exemple de cette sélectivité multiple a été montré par le groupe de Hillhouse en 2001, avec le récepteur de la CRH (Corticotrophin Releasing Hormon) capable de se coupler à 5 protéines G différentes : G_s, $G_{q/11}$, G_o, et avec une efficacité réduite $G_{i1/2}$ et G_z (Grammatopoulos), [RAN 01].

Famille	Sous Type	Effecteur	Distribution	Toxine
$G\alpha_s$	$\alpha_{s\,(S),(L),(XL)}$ α_{olf}	↗ AC, ↗Src tyrosine Kinase ↗ AC	Ubiquitaire Neurones olfactifs, tractus digestif et urogénital	CTX CTX
$G\alpha_{i/o}$	$\alpha_{o1/2}$	↘ AC, ↘ canaux Ca^{2+}, ↗ canaux K^+	Neurones,astrocytes, cœur	PTX
	$\alpha_{i1/2/3}$	↘ AC, ↘ canaux Ca^{2+}, ↗ canaux K^+,↗ Src tyrosine kinase,↗ MAPK	Ubiquitaire	PTX
	α_z	↘ AC, ↘ canaux Ca^{2+}, ↗ canaux K^+,↗ Rap1GAP	Plaquettes, Neurones	?
	α_{t1-2}	↗cGMP-Phosphodiesterase	Cônes, batônnets, bourgeons du goût	PTX
	α_{gust}	?	bourgeons du goût, chemorécepteurs des voies aériennes	PTX
$G\alpha_{q/11}$	$\alpha_{q/11}$	↗PLC$\beta_{1/3}$	Ubiquitaire	YM-254890
	$\alpha_{14/15/16}$	↗PLC$\beta_{1/3}$	Cellules hématopoïétiques	?
$G\alpha_{12/13}$	$\alpha_{12/13}$		Ubiquitaire	?
	$\beta_{1\,à\,5}$ $\gamma_{1\,à\,12}$	↗PLCβ, ↗PLA_2,↗GIRK ↘ canaux Ca2+, ↘AC type I ↗AC type II,IV,VII ↗PI3Kinase,↗Src Kinase ↗GRK,↗JNK	$\beta_1\gamma_1$:cônes $\beta_3\gamma_8$:batonnets β_5: neurones $\beta_{5(L)}$: rétine autres $\beta\gamma$: ubiquitaires	

Figure 1.11: Diversité et fonctions des différentes sous-unités des protéines G [6].

Cependant, tous les RCPG ne sont apparemment pas capables d'activer plus d'une classe de protéines G. En effet, on estime que seulement 11% des RCPG peuvent activer plusieurs types de protéines G avec une efficacité différente, les autres RCPG seraient spécifiques d'une seule classe de protéine G avec 43% couplés à Gi/o, 33% couplés à $G_{q/11}$ et 25% préférentiellement couplés à Gs [WON 03].

1.5.7. Activation de Protéine G

Bien que leurs effecteurs différent et donc induisent des cascades de signalisation distinctes, les protéines G présentent un mécanisme commun d'activation et de désactivation : En absence de ligand et à l'état basal, le récepteur oscille entre une conformation inactive, majoritaire (notée R) et une conformation active mais non lié au ligand (notée R*). C'est sous la seconde conformation R* que le récepteur lie la protéine G hétérotrimérique, où la sous unité α est liée au GDP. Lors de la fixation du ligand (L) au récepteur, cet équilibre R-R* est déplacé en faveur de la conformation active du récepteur (L-R**). Dans cette

conformation active, le récepteur se lie à la protéine G et induit son changement conformationnel. Le récepteur activé agit comme catalyseur de l'ouverture du site GTPase de la sous unité α, ce qui aboutit à une réaction d'échange du GDP par du GTP. Ainsi selon la nomenclature établie pour les « petites G-protéines» de la superfamille Ras, le récepteur est une protéine de type GEF (Guanine Exchange Factor). Le processus moléculaire de l'échange se fait en deux étapes quasi simultanées, d'une part la libération du GDP lié à l'état basal à la sous-unité α, suivie de la liaison soit d'une nouvelle molécule de GDP ou de GTP. La liaison du GTP à la sous-unité α, entraîne un réarrangement structural du trio L-R-G et aboutit à la dissociation du complexe en trois : αGTP, βγ et le récepteur ligandé. La sous unité αGTP et le complexe βγ ainsi dissociés du récepteur interagissent alors avec leurs effecteurs intracellulaires. Comme mentionné précédemment, la sous-unité α possède une activité GTPasique intrinsèque qui restaure la forme αGDP et permet sa réassociation avec le complexe βγ et donc avec le récepteur.

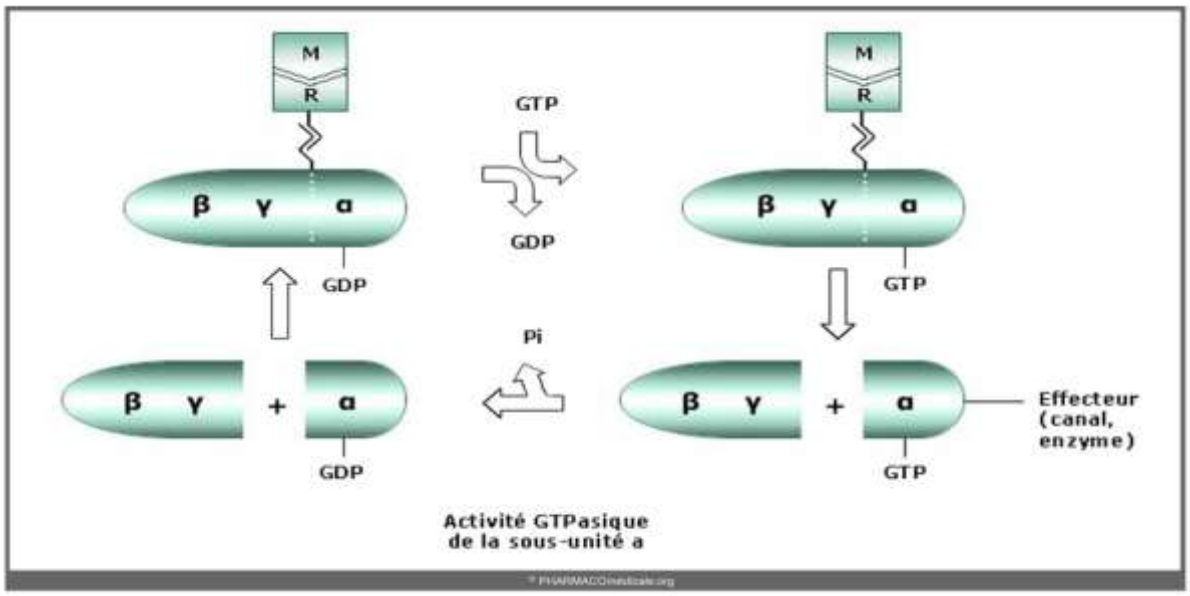


Figure 1.12: schéma du fonctionnement des protéines G [7].

Il faut souligner que les protéines G ne sont pas associées à un récepteur unique mais qu'elles peuvent diffuser au sein de la membrane et s'associer à différentes cibles. Enfin, il existe une relation entre les types de récepteurs et le type de protéine G avec laquelle ils interagissent. Les sous-unités a des différentes protéines G se distinguent les unes des autres par leur affinité pour différentes protéines effectrices et donc se différentient selon l'effet intracellulaire engendré.

- **ACTIVATION DES PROTEINES-G : LE CYCLE GTPASE :** Le signal apporté par le stimulus extracellulaire est transduit à l'intérieur de la cellule, par l'intermédiaire du

récepteur. Le récepteur transmet l'information aux protéines-G intracellulaires, qui à leur tour activent ou inhibent des effecteurs intracellulaires.

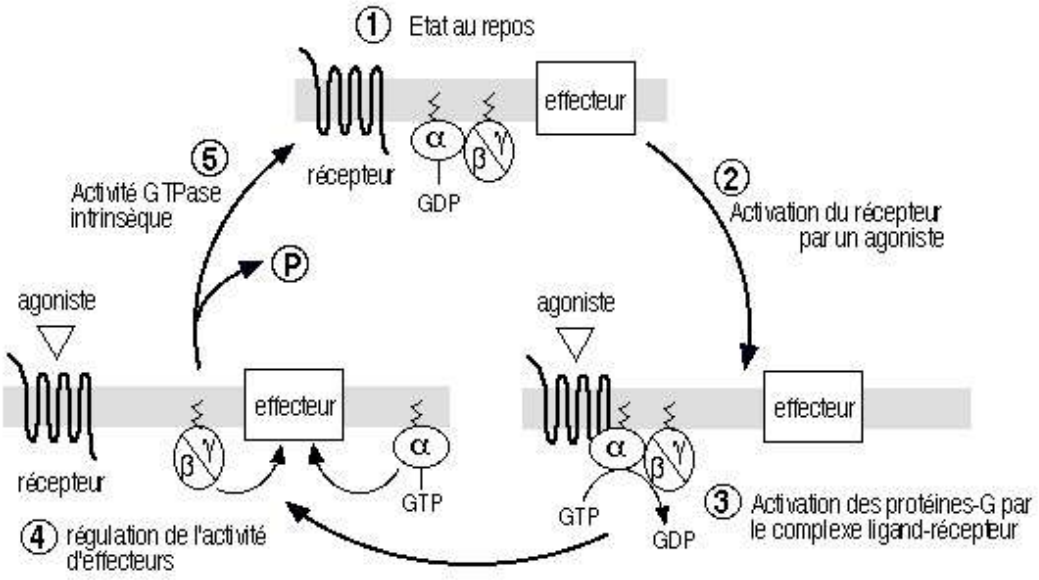


Figure 1.13: Cycle d'échange du GDP par du GTP [8].

Un RCPG au repos (1) est activé par la liaison d'un agoniste spécifique (2). Le changement de conformation du complexe agoniste-récepteur, induit par cette liaison, permet l'activation de l'échange du GDP par du GTP et donc l'activation de la protéine-G hétérotrimérique (sous-unités Ga et Gb/g) intracellulaires (3) qui vont aller réguler l'activité de divers effecteurs (4) membranaires ou cytosoliques. Le déclenchement de l'activité phosphatase, intrinsèque à la sous-unité Ga entraîne la réassociation des sous-untités Ga et Gb/g (5) et le retour à l'état initial (1).

Tant que le récepteur est activé par son ligand, et tant que le système ne subit pas une désensibilisation, le cycle d'échange du GDP par du GTP continue.

1.5.8. Classification de RCPG

Vu le grand nombre de GPCR, il a été nécessaire de les séparer en familles afin de mieux pouvoir les caractériser. Un système de classification basé sur l'homologie entre les différents GPCR fut proposé par Kolakowski [KOL 94]. Dans ce système, les récepteurs sont distribués dans 7 groupes soit les groupes : A, B, C, D, E, F et O

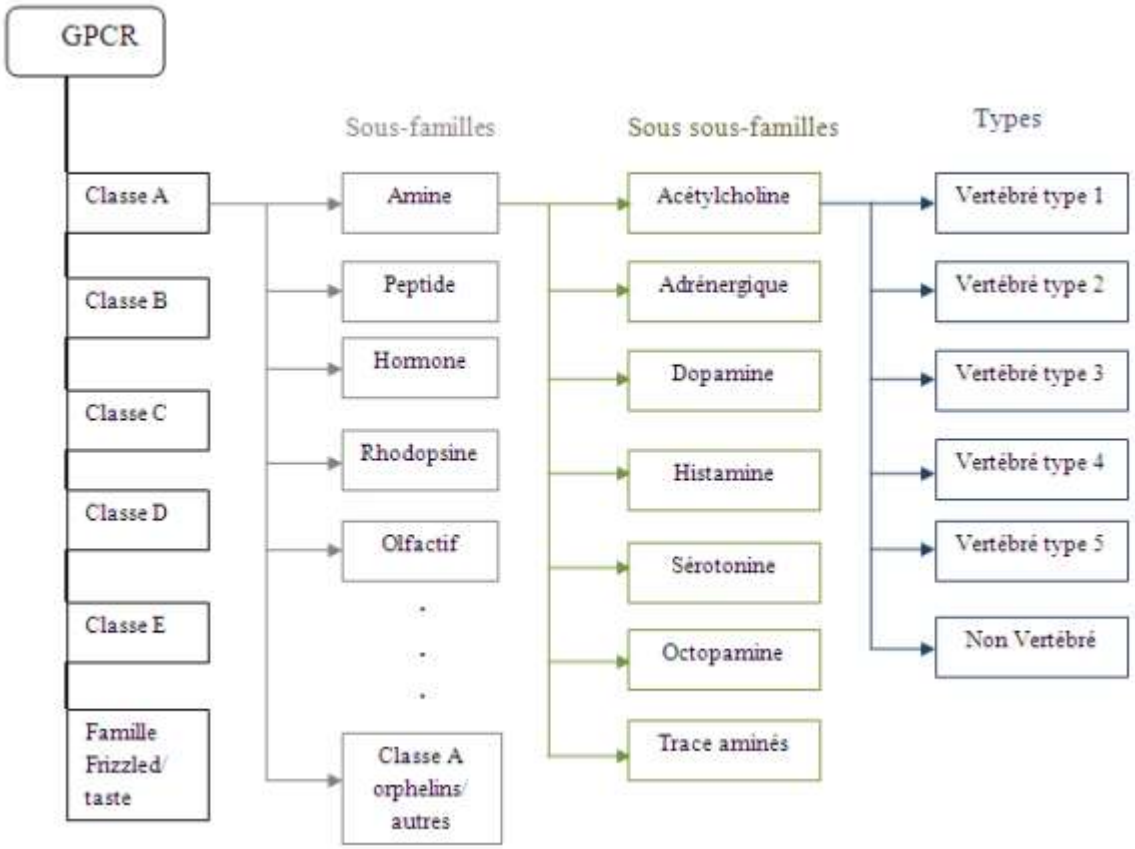


Figure 1. 14: Vue simplifiée de l'arbre des familles des récepteurs couplés aux protéines G (classification du système d'information GPCRDB) [HUA 04].

- ***GPCR de la classe A***

Ce sont les récepteurs de la famille de la rhodopsine et il s'agit de la plus grande famille de récepteurs comprenant environ 80% de l'ensemble des GPCR connus. Les récepteurs de classe A lient majoritairement des peptides, des amines biogènes ou des lipides. Même si les ligands des récepteurs de cette famille sont très divers, il y a une forte homologie de séquence entres les récepteurs membres de cette famille. Cette classe est divisée en trois sous-groupes suivant la taille et l'emplacement du site de liaison du ligand

- **La sous classe A1** comprend des récepteurs dont les ligands sont de petites molécules

comme le rétinal et pour lesquels le site de liaison est constitué par une cavité formée par les segments transmembranaires au sein du tonnelet central du RCPG

- **la sous-classe A2** comprend des récepteurs dont le site de liaison implique l'extrémité

N-terminale, les boucles extracellulaires et la partie supérieure des domaines extracellulaires. Les récepteurs de chimiokines font partie de cette sous-famille

- **la sous-classe A3** comprend des récepteurs d'hormones glycoprotéiques

(Thyréostimuline -TSH, Hormone folliculo-stimulante-FSH ou Hormone lutéinisant-LH). Elle est caractérisée par un grand domaine N-terminal qui sert de site de liaison du ligand.

- ***GPCR de la classe B***

Les récepteurs de classe B lient des peptides de grandes tailles comme la sécrétine ou la parathormone. Contrairement aux ligands des récepteurs de classe A, il y a beaucoup d'homologie entre les divers ligands de cette famille [CAR, 06]. Ces récepteurs sont organisés en deux domaines : un domaine extracellulaire impliqué dans l'affinité et la spécificité de la liaison du ligand et un domaine transmembranaire requis pour l'activation du récepteur. L'extrémité C-terminale du ligand va d'abord interagir avec le domaine extracellulaire du récepteur ce qui va ensuite permettre à l'extrémité N-terminale du ligand d'interagir avec le domaine transmembranaire pour activer le récepteur [PAL 11].

- ***GPCR de classe C, D, E, F et O***

Les récepteurs de classe C ou récepteurs de la famille des récepteurs métabotropiques du glutamate se distinguent par un massif domaine extracellulaire hydrophile contenant de nombreuses cystéines. Dans cette famille on retrouve principalement les récepteurs métabotropiques du glutamate mais aussi certains récepteurs du goût ou du neurotransmetteur GABA [DAS 06]. Les récepteurs de classe D sont des récepteurs répondant aux phéromones et sont utilisés par certains organismes pour communiquer chimiquement entres eux [NAK 05]. Les récepteurs de classe E, aussi appelés récepteurs de l'AMPC, sont des constituants de la signalisation chimiotactique des myxomycètes [PRA 06]. La classe F des GPCR correspond aux récepteurs *Frizzled* et *Smoothened*, deux récepteurs nécessaires à la liaison de *Wnt* en plus de jouer un rôle majeur dans la voie de signalisation *Hedgehog* [FOO 02]. Les récepteurs de classe O sont des récepteurs orphelins; c'est-à-dire que l'on ne connaît pas leur ligand endogène. Il est possible que plusieurs de ces récepteurs aient des propriétés ligand-indépendantes [GLO 05].

1.6 Conclusion

Les GPCR sont des protéines membranaires responsables de la transduction des signaux de l'extérieur vers la cellule. Répartis dans tout le corps, ils sont impliqués dans diverses fonctions physiologiques. Ils sont ciblés par une immense diversité de ligands possibles: photons, ions, amines biogènes, hormones, glycoprotéines, molécules gustatives, etc.

Un GPCR est composé de 7 hélices transmembranaires reliées entre elles par des boucles intra et extracellulaires. Jusqu'à présent, la rhodopsine bovine est la seule structure cristallographique connue, apportant une information structurelle précieuse.

Les GPCR présentent un grand intérêt pharmacologique. Leur diversité et les nombreuses fonctions qu'ils contrôlent les impliquent dans de nombreuses pathologies. Plus de 50% des nouveaux médicaments commercialisés ciblent les RCPG. De plus, de nombreux GPCR sont encore orphelins, sans ligand connu et constituent donc des cibles pharmacologiques potentielles.

Le fait d'identifier la fonction d‘un RCPG, permet de mettre en œuvre un médicament qui peut stopper son effet néfaste sur l‘organisme et ainsi contrôler la pathologie ou, mieux encore, l‘éradiquer. De ce fait, il est primordial d‘exploiter une ou plusieurs méthodes de classification afin de pouvoir identifier un RCPG.

Dans ce chapitre, nous avons synthétisé un ensemble de travaux sur les méthodes de classification utilisées pour l‘identification des RCPG. L‘étude bibliographique que nous avons effectuée, nous a permis de constater que les approches bio-inspirées récentes, telles que les AG et ce de l'algorithme BAT, ont été très peu utilisées dans le domaine de la prédiction de la fonction de RCPG, en comparaison avec d‘autres méthodes statistiques telles que SVM, KNN, les arbres de décision et les réseaux de neurones artificiels. Ceci nous a incités à améliorer l‘identification de RCPG avec les approches de l'AG et l'algorithme BAT qui, malgré leur succès dans beaucoup de domaines restent encore largement inexploités dans celui de l‘identification de fonction de RCPG.

Chapitre 2
Les Algorithmes Génétiques et L'algorithme Bat

2.1 Introduction

La nature a trouvé le moyen de résoudre certains problèmes en apparence insolubles. La vie est ainsi présente quasiment partout sur terre, des terres gelées aux fosses sous-marines (présentant des températures et pressions élevées) en passant par les airs. Cette réussite s'explique par la puissance de l'évolution biologique. Elle permet d'adapter en permanence les différentes espèces aux milieux à coloniser.

Les informaticiens ont imaginé comment cette évolution pourrait être utilisée pour résoudre des problèmes complexes. C'est ainsi que les algorithmes bio-inspirés sont apparus. Dans une première partie, les principes sous jacents à l'évolution biologique ou naturelle sont expliqués. Ils sont nécessaires pour comprendre le fonctionnement global et l'inspiration des algorithmes génétiques et des algorithmes de chauve-souris. Ensuite sera présenté comment ceux-ci fonctionnent, de manière globale. Ils peuvent être utilisés dans de nombreux domaines d'application présentés par la suite.

Ce chapitre se décompose en deux parties essentielles: la première décrit les algorithmes génétiques, leur principe de fonctionnement ainsi leur utilisation dans les différents domaines et surtout dans le domaine de la bioinformatique. Quant à la deuxième partie est dédiée à l'explication de l'algorithme de chauve-souris (Bat), ses caractéristiques, et ses domaines d'application. Une synthèse des avantages et des lacunes de ces algorithmes bio-inspirés est présentée à la fin de ce chapitre.

2.2 Les Algorithme Génétiques

Les algorithmes génétiques sont basés sur l'évolution biologique. S'il n'est pas nécessaire de comprendre tous les détails de celle-ci, il est cependant important de comprendre la source d'inspiration de ces algorithmes.

2.2.1 Evolution biologique

L'évolution biologique fut étudiée à partir de la fin du 18^{e} siècle. En effet, les preuves de cette évolution s'accumulaient, et les scientifiques voulaient comprendre les phénomènes sous-jacents. C'est au début du 19^{e} siècle qu'est apparue la paléontologie (le terme est employé à partir de 1822), science qui s'intéresse aux fossiles et aux formes de vie aujourd'hui disparues [MAT 14]. Les scientifiques trouvaient de nombreux squelettes et les classaient. Ceux-ci présentaient de fortes ressemblances entre eux, ou avec des formes de vie actuelles. Il semblait donc évident qu'il y avait eu une continuité, et que les espèces s'étaient

progressivement modifiées au cours des millénaire. Enfin, les grands découvreurs allaient d'îles en îles, et de nouvelles espèces étaient couramment découvertes. Il apparaissait que les individus situés sur des îles assez proches étaient eux-aussi proches physiquement. Au contraire, ils étaient beaucoup plus différents d'individus issus d'un autre continent [MAT 14]. Les espèces avaient donc évolué de manière différente mais graduelle. L'évolution biologique n'était donc plus un tabou au 19^{e} siècle mais une réalité scientifique. Cependant, il restait à savoir comment cette évolution pouvait avoir eu lieu.

2.2.2 Des facteurs au code génétique

Les travaux de Mendel ne furent malheureusement pas connus de suite de la communauté scientifique. D'autres scientifiques continuèrent leurs travaux sur le stockage de l'information génétique et les découvertes s'enchaînèrent à grande vitesse. C'est ainsi que l'ADN fut isolé en 1869, puis les chromosomes en 1879 par Flemming. En 1900, les lois de Mendel furent redécouvertes par plusieurs chercheurs de manière indépendante [MAT 14]. Il parut alors évident que c'était dans l'ADN des chromosomes que se situaient les facteurs de Mendel. On parle dès 1909 de gènes au lieu de facteurs, et une première carte des gènes de la drosophile (son chromosome X) fut d'ailleurs proposée dès 1913. La structure de l'ADN en double hélice fut découverte en 1952 par Watson, Crick et Wilkins. Le code génétique fut AFDis dans les années 1960 [MAT 14]. On comprend alors mieux le passage des gènes aux enzymes. Il se fait en deux temps: la transcription, qui transforme l'ADN en ARN (une sorte de négatif de l'ADN qui peut se déplacer jusqu'au lieu de production de l'enzyme), puis la traduction, qui permet de passer de l'ARN à la suite des acides aminés formant la protéine. Les principes basiques de la génétique sont alors tous présents: un individu possède des chromosomes, contenant des gènes. Chaque gène correspond à une enzyme, grâce à un code qui indique la composition de celle-ci. L'ensemble des gènes d'un être vivant est appelé son génotype. Les interactions entre toutes les enzymes créent l' individu, appelé phénotype. Voici un schéma récapitulatif montré dans la figure (2.1).

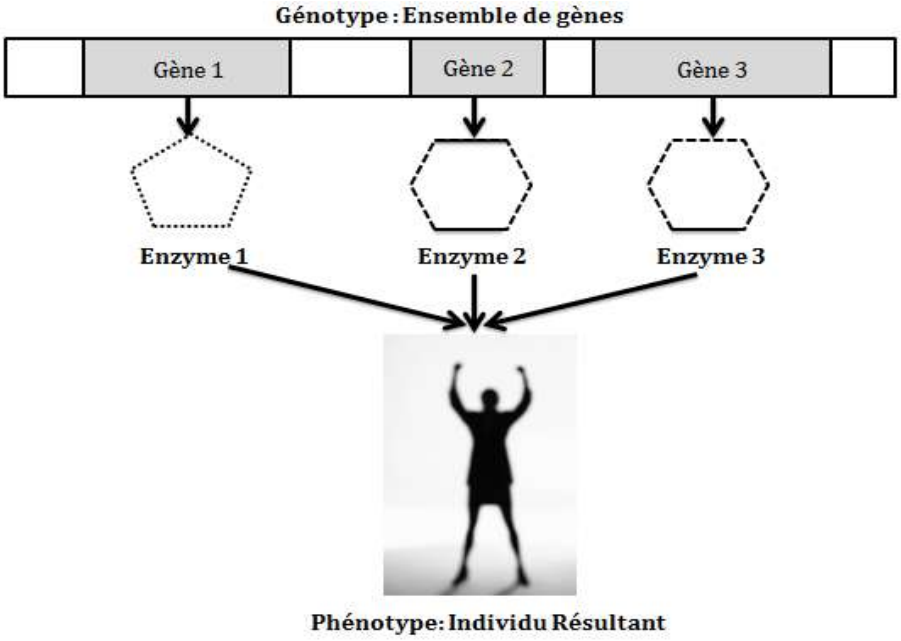


Figure 2. 1: Récapitulation du code génétique.

2.2.3 Cycle de vie

En résumé, un individu possède un ensemble de gènes (le génotype), présents à chaque fois en double exemplaire. La transcription puis la traduction permettent de transformer ces gènes en enzymes, qui vont réagir entre elles et créer l'être vivant (le phénotype). Lors de la reproduction, il va donner à son descendant la moitié de son capital génétique, qui sera mixé avec le capital génétique du deuxième parent. De plus, pendant ce processus, des mutations aléatoires peuvent se produire . L'individu ainsi créé va ressembler à ses parents, tout en leur étant légèrement différent. Selon les mutations qu 'il aura subies, il pourra être plus ou moins adapté que ses parents pour survivre dans son environnement. S'il est plus adapté, il aura plus de chances de survivre, sera plus résistant, ou plus attrayant, et va donc pouvoir ensuite se reproduire. Au contraire, si les mutations qu'il a subies le rendent moins adapté, il aura plus de difficultés à survivre. Les causes sont nombreuses : mort prématurée de l'individu, faiblesse, mauvaise résistance aux maladies, difficultés à se nourrir ou se déplacer... La sélection naturelle va donc avantager les mutations et les croisements d'individus intéressants pour la survie de l'espèce. Ceux-ci vont se disséminer dans la population et l'espèce va continuellement s'améliorer et s'adapter à son environnement. On peut résumer ce "cercle de la vie" par la figure suivante (2.2)

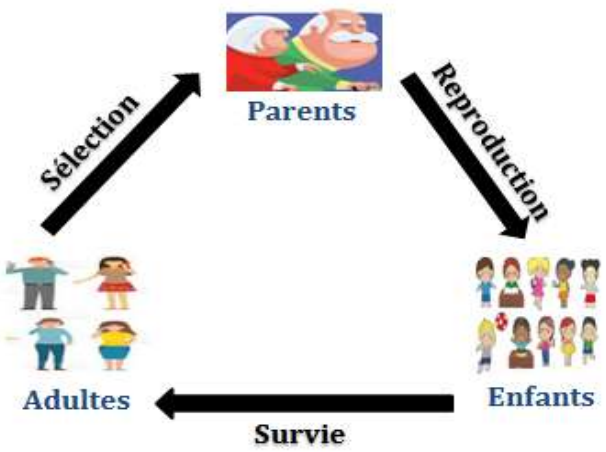


Figure 2. 2:Cycle de vie naturel d'un individu.

2.2.4 Evolution artificielle

L'évolution biologique vue précédemment est dite "désincarnée" : en effet, les principes de reproduction, de survie ou de sélection ne précisent pas comment les informations doivent être stockées ou transmises, ni même ce qui doit évoluer. Les chercheurs de domaines très divers s'y sont donc intéressés, que ce soit l'économie, la sociologie, la musique... etc. L'informatique n'est pas en reste, et cette évolution biologique peut être utilisée pour créer une évolution artificielle, permettant de résoudre des problèmes que des méthodes plus classiques ne permettent pas de résoudre. Les algorithmes évolutionnaires vont donc partir d'une population de solutions potentielles à un problème. Chacune est évaluée, pour lui attribuer une note, appelée fitness. Plus la fitness d'une solution est forte et plus celle-ci est prometteuse. Les meilleurs individus sont ensuite sélectionnés, et se reproduisent. Deux opérateurs artificiels sont alors utilisés: le croisement entre deux individus, appelé crossover, et des mutations aléatoires. Une étape de survie s'applique alors pour créer la nouvelle génération d'individus. Le processus est donc montré dans la figure (2.3).

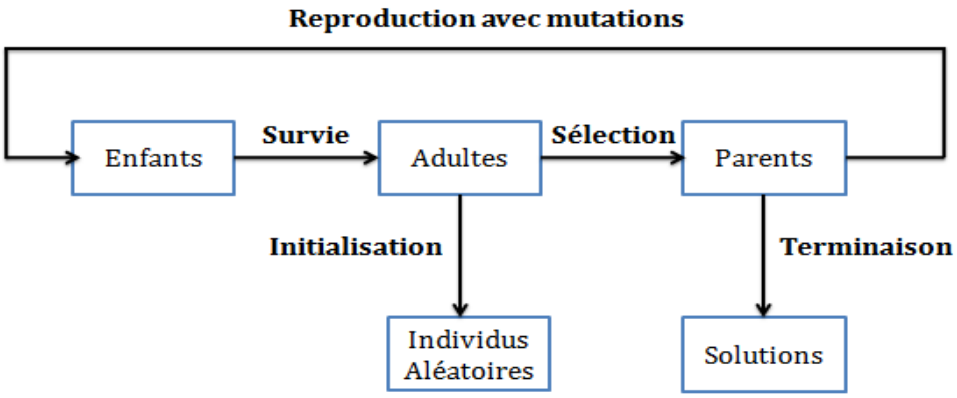


Figure 2. 3: Processus de l'AG.

Il est important de comprendre les enjeux et les principes de chaque phase utilisée dans un processus d'évolution artificielle. Chacune est ensuite étudiée de manière plus approfondie.

2.2.4.1 Initialisation et terminaison

Lors de la phase d'initialisation, une première population est créée. Pour la plupart des problèmes, on part de solutions aléatoires, qui sont donc en moyenne très peu adaptées au problème. Si on connait déjà des solutions acceptables au problème, il est possible de directement les injecter lors de l'initialisation. Le processus complet est alors plus rapide, nécessitant moins de générations. Il faut aussi définir un critère d'arrêt. Celui-ci permet de savoir à quel moment s'arrêter, pour donner à l'utilisateur les meilleures solutions trouvées. Ce critère peut porter sur un nombre de générations ou sur une fitness minimale à obtenir par les individus (un score à atteindre).

2.2.4.2 Sélection

La sélection consiste à déterminer quels sont les individus qui méritent d'être choisis comme parents pour la génération suivante. Il faut qu'en proportion, les meilleurs parents se reproduisent plus que les parents à fitness plus basse mais chacun doit quand même avoir une probabilité non nulle de se reproduire. En effet, c'est parfois en faisant muter ou en croisant des solutions en apparence "mauvaises" que l'on peut trouver une bonne solution à un problème.

2.2.4.3 Reproduction avec mutations

Lors de la reproduction, on choisit pour chaque enfant de 1 à N parents. Les informations génétiques des différents parents sont mixées avec l'opérateur de crossover. Une fois le mélange des parents effectué, on applique des mutations choisies aléatoirement au résultat, et dont le nombre dépend du taux de mutation de l'algorithme.

2.2.4.4 Survie

Les enfants étant créés, il faut maintenant obtenir une nouvelle génération d'adultes qui peuvent ou non se reproduire. Si la solution la plus simple consiste à remplacer toute la génération des parents par la génération des enfant s, il existe cependant plu sieurs autres stratégies de survie. On obtient donc à la fin de la survie une nouvelle population, et on peut reboucler tout le processus.

2.2.4.5 Convergence

La convergence vers la solution optimale est démontrée théoriquement. Cependant, rien ne précise le temps nécessaire pour converger vers cette solution, qui peut donc être supérieur à ce qui est acceptable. Il est donc important de bien choisir les différents opérateurs (sélection, mutation, crossover et survie) et les représentations, au nombre de trois : des gènes, des individus et de la population.

2.3. Applications des Algorithmes génétiques

Dans cette section, nous discutons de l'applicabilité d'un algorithme génétique au processus de recherche dans le data mining puisque Les AFD nécessitent une technique qui partitionne les valeurs de domaine d'un attribut dans un ensemble limité d'intervalles, simplement parce qu'il est impossible de prendre en compte toutes les plages possibles de valeurs de domaine. De plus, nous expliquons les applications de l'AG pour résoudre les problèmes de la bioinformatique et optimiser les performances des tâches de la fouille de données en bioinformatique.

2.3.1 Data mining et reconnaissance de formes

L'algorithme génétique est un outil efficace à utiliser dans l'exploration de données et la reconnaissance de formes. Il existe deux méthodes différentes pour appliquer l'AG à la reconnaissance de formes:

1. Utilisez AG comme classificateur directement dans le calcul.
2. Utilisez un AG pour optimiser les résultats, c'est-à-dire en tant qu'optimiseur pour organiser les paramètres dans d'autres classifieurs. La plupart des applications des AG dans la reconnaissance de formes optimisent certains paramètres du processus de classification [PAT 95].

Les AG ont été appliqués pour trouver un ensemble optimal de pondérations d'entités qui améliorent la précision de la classification. Tout d'abord, une méthode traditionnelle d'extraction de caractéristiques telle que l'analyse en composantes principales (PCA) est appliquée, puis un classificateur tel que KNN est utilisé pour calculer la fonction de fitness pour AG [PEI 98], [SIE 89].

2.3.2 Business

L'algorithme génétique a une large portée ans ce domaine. Une grande quantité de données doit être filtrée pour traiter les résultats pour optimiser le bénéfices des entreprises en utilisant diverses techniques d'exploration de données. Il existe de nombreux domaines dans lesquels elles peuvent s'appliquer:

- Optimisation: Donner un problème commercial avec certaines variables et une définition bien définie du profit, un algorithme génétique peut être utilisé pour déterminer automatiquement la valeur optimale des variables qui optimisent le profit [FOR 93].
- Prédiction : Les algorithmes génétiques ont été utilisés comme opérateurs de méta-niveau qui sont utilisés pour aider à optimiser d'autres AFD. Par exemple, ils ont été utilisés pour trouver les règles d'association optimales dans l'analyse de marché.
- Simulation : Parfois, un problème commercial spécifique n'est pas bien défini en termes de bénéfice ou de savoir si une solution est meilleure que l'autre. L'homme d'affaires a plutôt un grand nombre d'entités qu'il aimerait simuler via de simples règles d'interaction au fil du temps.

2.3.3. Sélection de caractéristiques

Les algorithmes génétiques se sont avérés être une méthode très efficace dans le problème de la sélection des caractéristiques. Plusieurs chercheurs ont visé les algorithmes génétiques pour sélectionner les meilleurs sous ensemble d'attributs pertinents à partir d'un ensemble plus grand de caractéristiques. Dans [SIK 07] les auteurs ont présenté un AG pour faire à la fois les tâches de data mining et de sélection de caractéristiques simultanément en faisant évoluer un code binaire parallèlement à la structure chromosomique utilisée pour faire évoluer les règles. De plus, [HUA 07] dans leur étude, un algorithme génétique hybride est adopté pour trouver un sous-ensemble de caractéristiques qui sont les plus pertinentes pour la tâche de classification.

2.3.4. Bioinformatique

L'alignement de séquences multiples (MSA) est utilisé dans l'analyse génomique, comme l'identification des motifs de séquence conservés, l'estimation de la divergence évolutive entre les séquences et l'inférence des relations historiques des gènes. Plusieurs recherches ont été menées pour déterminer le niveau de similitude d'un ensemble de séquences. En raison

du problème de la nature du NP-complet, un certain nombre de recherches utilisent des algorithmes génétiques (GA) pour trouver une solution à l'alignement de séquences multiples [BEN 16]. Aussi les auteurs ont présenté dans [PRA 14] une approche basée sur un algorithme génétique pour résoudre le problème de MSA efficacement en utilisant une nouvelle présentation des chromosomes basée sur la matrice de notations (Scoring Matrix). La recherche montrée dans [NIZ 11] se concentre sur l'application de l'algorithme génétique avec le concept d'auto-organisation pour MSA, L'auto-organisation fait référence au système qui fonctionne sans aucune intervention extérieure. L'algorithme proposé [NIZ 11], est un algorithme génétique cyclique (CGA) peut être développé avec une connaissance complète du problème et de ses paramètres. Une approche d' AG a été créée par Chowdhury et al. [CHO 16] pour trouver des gènes et annoter des génomes, aussi selon [MAT 02] il est possible d'améliorer encore la prédiction génétique et la découverte de nouveaux gènes grâce à l'utilisation des AG. Cependant, Une approche d'algorithme génétique proposée et développée par Hwang et al. [HWA 13] peut être utilisée pour prédire quels gènes sont essentiels. Par définition, des gènes essentiels sont nécessaires pour qu'un organisme vive, ce qui signifie que leur identification par des expériences en laboratoire est difficile. Bien qu'importante, la découverte de gènes et d'éléments régulateurs à elle seule présente encore des informations génétiques incomplètes. Afin d'obtenir des informations plus pratiques sous forme de prédictions phénotypiques, les interactions entre les caractéristiques génomiques doivent également être prises en compte. En raison du nombre exponentiellement élevé d'interactions possibles, ainsi que de la nature multidimensionnelle des données impliquées, les modèles statistiques ont très peu de pouvoir pour prédire ces associations, les AG peuvent être utilisés pour contourner ce problème car leurs populations explorent de nombreuses solutions possibles au problème, mais ne constituent pas une recherche exhaustive. Pour cette raison, Moore et al. [MOO 04] ont utilisé les AG pour développer une méthode de prédiction de modèles de risque de maladie. Une autre approche de Yang et al. [YAN 16] ont combiné les AG avec un algorithme de recherche local pour étudier l'influence des polymorphismes nucléotidiques (SNP) sur la maladie et prédire les gènes associés à la maladie.

Une autre méthode basée sur l'AG pour déterminer les gènes associés à la maladie a été développée par Tahmasebipour et al. [TAH 15] Ces chercheurs ont noté que les études communément utilisées par la Genome Wide Association (GWA) et les approches d'analyse comparative de la prédiction des gènes de maladies sont limitées par leur focalisation sur la prédiction de gènes d'une seule maladie [TAH 15]. La maladie pouvant être le résultat de

nombreux allèles en interaction, les SNP qui ont une faible contribution individuelle au risque de maladie peuvent avoir un effet important en présence d'autres allèles spécifiques. Afin de tenir compte de cette possibilité, les auteurs ont recadré le problème, représentant l'association de la maladie comme un réseau de nœuds représentant des gènes, de l'ARNm, des protéines et d'autres agents [TAH 15]. Azad et coll. [AZA 11] ont développé une méthode combinant des algorithmes génétiques avec une machine à vecteurs de support (SVM) pour classer les séquences en catégories de promoteurs et de non-promoteurs. Dans ce contexte, L'outil PROMOBOT qui est développé par [UMA 17] reste un outil de pointe pour l'identification des promoteurs végétaux avec une sensibilité moyenne de 85%.

Une approche basée algorithme génétique [OOI 03] a été appliquée aux données de puces à ADN de plusieurs lignées de cellules cancéreuses, suggérant que les AG peuvent fournir des applications diagnostiques pratiques dans le domaine de la médecine [GHA 15].

Pour la prédiction structurelle des protéines, un problème supplémentaire est que la similitude de séquence n'indique pas toujours une similitude structurelle [BYW 16], afin de maintenir une précision élevée et de minimiser les hypothèses, Pedersen et Moult [PED 97] ont utilisé des algorithmes génétiques pour réduire l'espace de recherche du problème. En faisant, moins d'évaluations structurelles des protéines seraient nécessaires, ce qui atténuerait coûts de calcul tout en maintenant la précision.

Un nouvel algorithme hybride ACO / GA a été proposé par [NEM 09] pour la sélection de caractéristiques qui combine des algorithmes génétiques (GA) et l'optimisation des colonies de fourmis (ACO) dans la prédiction de la fonction des protéines pour une recherche plus rapide et meilleure. L'algorithme hybride utilise les avantages des méthodes ACO et GA.

2.4. Avantages et limites des Algorithmes Evolutionnistes

2.4.1 Les avantages des AG

Les algorithmes génétiques sont un outil puissant et flexible pour relever de nombreux défis en bioinformatique. Tout problème où les solutions peuvent être classées de manière significative pour mesurer l'aptitude peut être résolu grâce à l'aspect évolutif des AG. Cela est particulièrement utile car de nombreuses autres méthodes d'apprentissage automatique reposent sur l'utilisation de données d'apprentissage. Ces méthodes sont souvent victimes de biais dans les données, faisant de la sélection des ensembles d'apprentissage un problème en soi [KUB 97]. Les AG ne nécessitent qu'une fonction de fitness et évitent donc ces problèmes; ces fonctions peuvent souvent être conçues avec une connaissance préalable

limitée d'un problème de bioinformatique donné. Cela rend les AG idéaux pour les problèmes où l'espace de recherche est mal compris, car ils peuvent tolérer un certain bruit dans la fonction de fitness [MAN 13]. En raison de ces caractéristiques, les AG peuvent être rapidement appliquées à de nouveaux domaines de recherche en bioinformatique où les connaissances sont imparfaites. Cette flexibilité conduit également les GA à être facilement hybrides avec d'autres approches. Dans de nombreux cas, la combinaison de méthodes préexistantes avec des AG peut conduire à des algorithmes encore meilleurs, comme le démontrent les travaux de [NOT 97, SÏR 10, HWA 13].

En plus des améliorations globales grâce à des méthodes combinées, des approches hybrides peuvent également être utilisées pour se concentrer sur des aspects particuliers d'une solution qui peuvent être souhaitables. En ce sens, les méthodes spécialisées qui peuvent être faibles en elles-mêmes peuvent être améliorées avec la force des AG tout en maintenant leur maîtrise dans des niches particulières.

Un autre aspect particulièrement utile des GA est leur capacité à réduire la complexité de calcul d'un problème donné. Du fait qu'ils n'explorent que certaines solutions, mais pas toutes, pour un problème donné, le nombre de calculs, et donc le temps d'exécution, requis est considérablement diminué. La complexité peut être réduite encore davantage grâce à l'utilisation de stratégies de réduction variables; ceux-ci intègrent certaines connaissances du domaine pour améliorer considérablement les performances d'une AG [WU 13]. Ceci est extrêmement utile en bioinformatique, car la plupart des problèmes sur le terrain impliquent des statistiques complexes pour arriver à une solution. Cet avantage peut être encore amélioré grâce à la parallélisation; le système basé sur la population des AG peut facilement être distribué sur plusieurs processeurs, ce qui réduit considérablement les exigences d'exécution [MAN 13].

L'avantage le plus grand et le plus simple des algorithmes génétiques est peut-être leur capacité à apprendre des solutions avec peu de connaissances préalables sur le problème. Même des définitions simples de la fonction fitness peuvent aboutir à des solutions extrêmement complexes et efficaces à un problème [ANG 94]. En conséquence, le processus sélectif dans les AG aboutit souvent à des solutions robustes qui n'auraient pas été prévues par la recherche dirigée par l'homme. Ces approches limitent également les hypothèses des chercheurs, car le processus d'apprentissage peut avoir lieu sans fournir à l'algorithme des connaissances explicites [ANG 94].

2.4.2 Inconvénients des AG

Malgré que les algorithmes génétiques sont un outil très efficace en bioinformatique, aucune méthode n'est parfaite, et les AG présentent quelques inconvénients. Bien que les AG utilisent souvent moins de connaissances a priori que les autres approches, ils exigent les chercheurs à concevoir une fonction de fitness [SPE 90]. Certains problèmes peuvent nécessiter des fonctions objectives complexes, ce qui signifie qu'une connaissance approfondie du problème peut encore être requise. La fonction fitness joue également un rôle déterminant dans l'orientation de l'algorithme vers une solution; des biais peuvent facilement être introduits dans la fonction qui modifie la sortie du programme. Si l'adéquation est mal définie ou si des approximations sont introduites, le processus de sélection suivra toujours ces règles, ce qui entraînera une erreur [SAS 02].

Cela se traduira par une solution médiocre qui ne fonctionnera que pour un cas de test et échouera dans les applications «réelles». Les décisions sur la taille de la population, le nombre de générations et les taux de croisement et de mutation peuvent également influencer l'apprentissage et nécessiter une mise au point pour obtenir les meilleurs résultats [MAN 13]. Contrairement à la fonction de fitness, ces décisions doivent toutes être prises de manière arbitraire, ce qui entraîne un processus minutieux d'essais et d'erreurs pour le réglage des paramètres. L'optimisation des GA peut être un problème en soi, et elle peut nécessiter un algorithme à exécuter plusieurs fois pour trouver la meilleure solution [MAN 13].

Une autre préoccupation est que si les GA peuvent être utilisés pour trouver des solutions solides, ce ne sont généralement que des «réponses». En d'autres termes, leur production n'enseigne pas directement aux scientifiques les règles du problème. La découverte des méthodes génétiques sous-jacentes et la découverte de l'organisation des génomes étant l'un des principaux sujets de recherche, c'est l'une des plus grandes lacunes des AG et de l'apprentissage automatique dans le domaine. Les GA peuvent être utilisés pour guider cette type de recherche en fournissant des exemples, mais ne fournissant que des réponses et jamais une explication sur les raisons pour lesquelles l'algorithme y est arrivé. En ce sens, les AG sont une «boîte noire» avec une entrée et une sortie [MAN 13].

Dans le cadre de la méthodologie bioinformatique, les AG sont relativement rapides pour faire des prédictions précises. Cependant, le fait qu'une population de grande taille et de nombreuses générations soient requises par les AG signifie qu'elles prennent encore beaucoup de temps à fonctionner; elles sont généralement plus lentes que les méthodes de recherche dirigée [MAN 13]. Ces contraintes d'exécution signifient que les GA ne sont pas la

meilleure option pour les problèmes simples qui ne nécessitent pas un apprentissage automatique substantiel à résoudre. Dans les cas où les règles du problème sont bien établies, un algorithme de recherche intégrant des connaissances préexistantes peut être plus efficace (Manning et al. 2013). Une solution de «supposition éclairée» peut également suffire dans certains cas, ce qui signifie que l'amélioration de la précision d'un GA peut ne pas valoir le coût d'exécution. Dans ces situations, les AG seraient extrêmement long à exécuter par rapport aux autres algorithmes, ce qui rend les méthodes traditionnelles plus adaptées. Les coûts d'exécution des GA peuvent être atténués par la parallélisation, mais ils peuvent encore être considérablement lents [MAN 13].

2.5. L'algorithme Bat

Un algorithme de chauve-souris (Bat) est un algorithme heuristique qui opère en imitant le comportement d'écholocation des chauves-souris pour effectuer une optimisation globale. L'algorithme Bat est largement utilisé dans divers problèmes d'optimisation en raison de ses excellentes performances. Dans l'algorithme chauve-souris, la capacité de recherche globale est déterminée par le paramètre volume et fréquence. Cependant, les chercheurs montrent que chaque opérateur dans l'algorithme ne peut améliorer les performances de l'algorithme qu'à un certain moment. Cet algorithme effectue le processus de recherche en utilisant des chauves-souris artificielles comme agents de recherche imitant le volume sonore naturel des impulsions et le taux d'émission de vraies chauves-souris. Pour améliorer les performances de l'algorithme Bat, différentes stratégies ont été proposées. Nous élaborerons dans les sections suivantes.

2.5.1. Concepts basiques

L'algorithme Bat, introduit par Yang [YAN 10, YAN 12a], est basé sur le comportement d'écholocation des chauves-souris. Les bats sont les seuls mammifères capables de voler. Il existe de nombreuses espèces de chauves-souris et elles sont de tailles différentes. Parmi eux, les micro-bats utilisent largement l'écholocation. Ils émettent des impulsions sonores fortes et écoutent l'écho qui rebondit sur les objets environnants. Les chauves-souris émettent des impulsions sonores de fréquence constante dans la gamme de 25 KHz à 150 KHz. Chaque sursaut sonore ultrasonique dure très peu de temps, généralement de 5 ms à 20 ms. Les micro bats émettent normalement environ 10 à 20 salves sonores par seconde. Le taux d'émission de ces impulsions sonores est augmenté (jusqu'à environ 200 impulsions par seconde) lorsqu'ils volent à proximité de leur proie. En cherchant une proie, les chauves-

souris produisent des impulsions sonores fortes (de l'ordre de 110 dB), mais lorsque les chauves-souris se rapprochent de leur proie, elles deviennent plus silencieuses. De telles capacités d'écholocalisation des micro-bats peuvent être associées à la fonction objective à optimiser et des algorithmes d'optimisation peuvent être formulés qui imitent ce comportement de chauve-souris pour trouver la solution optimale.

Le pseudo-code [YAN 12b] de BA est simple et peut être implémenté dans n'importe quel langage de programmation comme suit, et l'organigramme général de cet algorithme est présenté dans la figure 2.4:

Pseudo code de l'Algorithme bat

Objective function f(x),x=(x1....xd)^T

Initialize the bat population xi (i=1, 2,...., n) and vi

Define pulse frequency fi at xi

Initialize pulse rates ri and the loudness Ai

while(t<Max number of iterations)

Generate new solutions by adjusting frequency,

and updating velocities and locations/solutions

if(rand>ri)

Select a solution among the best solutions

Generate a local solution around the selected best solution

end if

Generate a new solution by flying randomly

if(rand<Ai &f(xi)<f(x∗))

Accept the new solutions

Increase ri and reduce Ai

end if

Rank the bats and find the current best x∗

end while

Postprocess results and visualization

Tableau 2. 1: Pseudo code de BA.

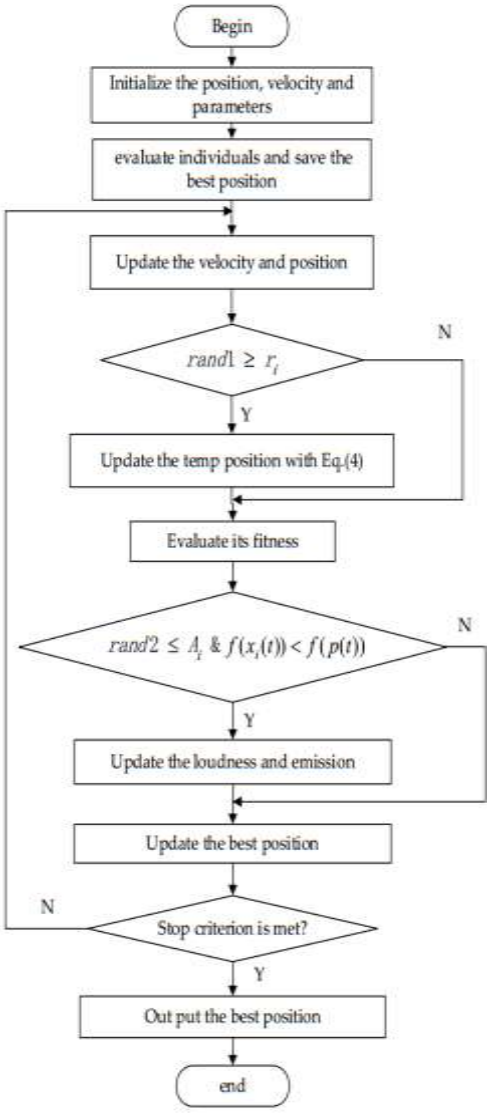


Figure 2. 4: Organigramme du BA

2.5.2. Caractéristique du BA

BA est simple, facile à implémenter et flexible. Il a été appliqué avec succès à un large éventail d'applications. Le rendement élevé de BA peut être attribué aux trois caractéristiques suivantes comprenant le réglage de fréquence, le zoom automatique et le contrôle des paramètres [YAN 13].

- **Réglage de fréquence:** BA utilise l'écholocalisation et le réglage de la fréquence / longueur d'onde pour résoudre les problèmes d'optimisation [CHA 15]. Bien que l'écholocation ne soit pas directement utilisée pour imiter la vraie fonction en réalité, les variations de fréquence sont utilisées. Cette capacité fournit certaines fonctionnalités qui pourraient être similaires à la fonction clé utilisée dans PSO et la recherche d'harmonie (HS). Il utilise une bonne combinaison des avantages majeurs de ces algorithmes et, par conséquent, est potentiellement plus puissant qu'ils ne le sont.
- **Zoom automatique:** Comparé à d'autres MOA (Metaheuristic Optimization Algorithms), BA présente des avantages distincts. Il a la capacité intégrée de transiter automatiquement dans les zones où il est possible d'obtenir des solutions prometteuses. Cette transition ou en plein essor s'accompagne du passage automatique d'un mouvement d'exploration diversifié à un mouvement d'exploitation local intensifié. BA montre un taux de convergence élevé, aux premiers stades des itérations, par rapport à d'autres algorithmes [ALS 14, CHA 15].
- **Contrôle des paramètres:** Le contrôle des paramètres est de la plus haute importance pour que toute les MOA fonctionnent efficacement. Les performances d'un algorithme dépendent largement des paramètres de l'algorithme. Dans BA, le contrôle des paramètres peut être effectué de telle manière que les valeurs des paramètres qui comprennent la sonie et le taux d'émission d'impulsions peuvent être modifiées au fur et à mesure des itérations. De cette façon, le BA fournit un mécanisme intégré pour passer automatiquement de l'étape d'exploration à l'étape d'exploitation lorsque la solution optimale approche.

2.5.3. Domaines d'application

Même si BA a été introduit très récemment en 2010, la gamme d'études publiées disponibles en ligne utilisant ce nouvel algorithme a été augmenté de jour en jour. Dans la plupart des cas, sinon tous, les résultats et la qualité des solutions, obtenus à l'aide de cet algorithme, ont

surpassé ou correspondaient à ceux obtenus à l'aide d'autres MOA bien connus. Diverses applications en sciences de l'ingénieur employant un BA sont présentées dans cette section.

2.5.3.1. Ingénierie biomédicale

Le but de l'estimation des paramètres est de déterminer les paramètres du modèle qui donnent le meilleur ajustement à un ensemble de données expérimentales. La tâche d'estimation des paramètres est difficile et coûteuse en calcul en raison de la dynamique hautement non linéaire et de la mesurabilité limitée des systèmes biologiques. Dans [LIN 12], les auteurs ont discuté d'une méthodologie générale pour sélectionner de manière adaptative les valeurs des paramètres du modèle pour la reconstruction de la dynamique du système biologique et ont développé des algorithmes d'optimisation améliorés basés sur BA. Il intègre à la fois les effets de la dynamique chaotique et des vols de Levy. On a supposé que les séries chronologiques observées provenaient d'un système primaire avec des paramètres inconnus et qu'elle était utilisée pour piloter un système secondaire, de sorte que les deux systèmes soient couplés. Les paramètres du système secondaire ont été optimisés de manière adaptative par le vol chaotique Levy BA proposé pour lui faire suivre la dynamique du système primaire.

L'estimation de la pose humaine du corps entier à partir de séquences vidéo trouve des applications dans de nombreux domaines tels que l'animation de personnages de dessins animés, l'analyse de la démarche humaine, biomécanique sportive, robotique, interface homme-machine et reconnaissance gestuelle. Ce problème d'estimation de la pose humaine du corps entier dans les séquences vidéo a été abordé par Akhtar et al. [AKH 12] à l'aide de BA. Il a été formulé comme un problème d'optimisation non linéaire en 31 dimensions. La performance de BA a été comparée avec le filtre à particules (PF), le filtre à particules recuites (APF) et le PSO en utilisant un ensemble de données standard. Les résultats d'erreur de pose la plus probable / appropriée (MAP) suggèrent que le BA a mieux performé que les trois autres algorithmes employés. Expérimentalement, on constate qu'avec l'augmentation de la population de chauves-souris, la précision du suivi est augmentée aux dépens de la charge de calcul.

2.5.3.2. Ingénierie Informatique

Bat Intelligent Hunting (BIH) développé par Kim [KIM 10] fournit un cadre pour résoudre divers problèmes d'optimisation. Il modélise les comportements de chasse aux proies des chauves-souris. Les chauves-souris localisent et capturent leurs proies sans utiliser leur vue

en utilisant des techniques d'écholocation et d'approche de cible absolue constante (CATD). BIH implémente ces concepts pour converger vers la solution optimale et a été utilisé pour résoudre le problème de planification multiprocesseur. Ce problème concerne l'attribution d'un ensemble donné de tâches à un ensemble de processeurs afin d'optimiser des objectifs spécifiés, l'un avec mise à l'échelle de la tension prenant en compte l'énergie, et l'autre sans mise à l'échelle de la tension avec l'énergie pas considéré comme l'objectif. Dans les deux problèmes, plusieurs optimisations objectives ont été effectuées par les auteurs à l'aide de la fonction d'utilité additive pondérée normalisée, dans laquelle un ensemble d'importance de valeurs objectives (poids) ont été utilisés pour identifier un ensemble de solutions efficaces.

Komarasamy et Wahi [KOM 12] ont proposé un nouvel algorithme d'optimisation basé sur l'algorithme K-means et l'algorithme chauve-souris (KMBA). Cet algorithme n'exige pas que l'utilisateur donne à l'avance le nombre de cluster et de centre de cluster, il résout le problème de cluster K-means (KM). Cette méthode recherche le centre du cluster, qui est généré à l'aide du BA, puis forme le cluster à l'aide du KM. Il utilise les meilleurs résultats de KM et BA pour former un cluster et, ainsi, évite les lacunes de KM et BA. Les expériences ont été réalisées sur trois ensembles de données différents et ont montré la robustesse et l'efficacité du nouvel algorithme. Le nouvel algorithme améliore la vitesse de convergence de BA et aide KM à rester indépendant des centres initiaux.

Aussi un nouvel algorithme de chauve-souris avec plusieurs stratégies de couplage pour l'optimisation numérique est proposé par Wang et al. [WAN 19] pour résoudre ce problème. De plus, l'algorithme bat a été proposé pour optimiser les paramètres de la machine à vecteurs de support (SVM) qui réduisent l'erreur de classification [CUI 19]. Notamment, augmenter la précision du classifieur SVM et éviter le piège des optima locaux en utilisant l'algorithme Bat a été très utile dans la recherche biomédicale [THA 16, BAS 18].

2.5.3.3. Génie industriel et de production

Kaveh et Zakian [KAV 13] ont présenté une amélioration de BA pour effectuer l'optimisation de la taille des structures squelettiques constituées de fermes et de cadres. Divers problèmes d'optimisation comprenant la taille, la forme et la topologie ont été mis en œuvre pour démontrer la capacité du présent BA amélioré. Ces exemples de conception étaient associés à différentes contraintes qui comprenaient le déplacement, la fréquence et la contrainte. Les chargements dynamiques statiques et chronologiques ont également été pris en compte. Des problèmes d'optimisation discrets et continus ont été étudiés. Les résultats ont indiqué l'efficacité du BA pour l'optimisation de la conception des structures

squelettiques. L'outil d'ordonnancement basé sur l'algorithme de chauve-souris (BAST) a été développé par Musikapun et Pongcharoen [MUS 12] pour résoudre le problème d'ordonnancement multi-machines, multi-produits. L'objectif était de minimiser à la fois le coût des pénalités de précocité et de retard, de séquencer correctement les opérations requises pour fabriquer des composants et pour satisfaire la relation de priorité d'assemblage. Expérimentalement, il a été constaté que la qualité des solutions obtenues à partir de BAST peut être améliorée de manière significative après application des paramètres nécessaires identifiés par les outils statistiques. Les résultats obtenus avec un réglage optimisé BA ont surpassé ceux utilisant un réglage non optimisé. Taherian et al. [TAH 13] ont proposé un modèle basé sur un réseau amélioré de SVM. Il a été utilisé pour la prévision des prix sur le marché iranien de l'électricité en 2013. Dans ce modèle, les données d'entrée ont d'abord été regroupées par la technique des moyennes c-floues. Cela a permis de séparer les données en fonction du type de charge ou du jour de l'année. Ensuite, les données de formation appropriées ont été utilisées par le réseau SVM amélioré à court terme prévision des prix. BA a été utilisé pour optimiser les paramètres du réseau SVM. Les résultats ont été comparés à ceux des méthodes conventionnelles de réseaux de neurones, qui étaient basées sur un gradient et une SVM commune. Il a été constaté que les résultats obtenus en appliquant le modèle montraient une grande précision du modèle proposé.

2.5.3.4. La bioinformatique

On s'attend à ce que les données sur l'expression de gènes apportent une grande contribution à la production d'un diagnostic et d'un pronostic efficaces du cancer. Les données d'expression de gènes sont codées par de grands gènes mesurés, et seuls quelques-uns d'entre eux contiennent des informations précieuses pour différentes classes d'échantillons.

Récemment, plusieurs chercheurs ont proposé des méthodes de sélection génétique basées sur des algorithmes méta-heuristiques pour analyser et interpréter les données d'expression de gènes. Cependant, en raison du grand nombre de gènes sélectionnés avec un nombre limité d'échantillons de patients et une interaction complexe entre les gènes, de nombreuses méthodes de sélection de gènes ont connu des défis afin d'approcher les gènes les plus pertinents et les plus fiables.

Par conséquent, dans [AL-B 20], les auteurs ont été proposé une méthode Filtrer / Wrapper hybride, appelé rMRMR-MBA est proposé pour le problème de sélection de gène.

Dans cette méthode, la pertinence maximale de redondance minimale robuste (rMRMR) comme filtre pour sélectionner les gènes les plus prometteurs et un algorithme de chauve-

souris modifié (MBA) comme moteur de recherche dans une approche wrapper est proposé pour identifier un petit ensemble de gènes informatifs [AL-B 20].

2.5.3.5 Applications additionnelles

- Un système de dépistage des lieux de travail des entreprises à haut risque ergonomique a été créé par Khan, Nikov et Sahai [KHA 11]. Dans ce travail, une modification floue de BA a été obtenue pour le regroupement des lieux de travail des entreprises. Des grappes de lieux de travail à risque ergonomique faible, modéré et élevé ont été obtenues.
- Les lieux de travail présentant des niveaux de risque ergonomique modérés et élevés ont été supprimés et des solutions pertinentes ont été proposées. Cette méthode permet de réduire les efforts de calcul et de filtrer rapidement les lieux de travail présentant des problèmes ergonomiques majeurs au sein d'une entreprise.
- En outre, Yang et Gandomi [YAN 12b] ont utilisé BA pour résoudre divers problèmes d'ingénierie tels que la conception de fermes à trois barres, la conception de réducteurs de vitesse, l'identification des paramètres des structures, la poutre en porte-à-faux, la conception de l'échangeur de chauffage et le problème côté voiture. Pour la plupart des problèmes, les solutions optimales obtenues par BA étaient bien meilleurs que les meilleures solutions rapportées dans la littérature.

2.6. Conclusion

- Les algorithmes génétiques fournissent un ensemble d'outils incroyablement diversifiés et efficaces pour l'analyse bioinformatique. Leur capacité à résoudre des problèmes majeurs en génomique, transcriptomique et protéomique peut grandement accélérer la recherche en bioinformatique. Ceci est démontré par les performances toujours plus élevées des méthodes GA par rapport aux méthodes existantes dans ces catégories. Les méthodes GA sont également hautement adaptables et peuvent souvent surmonter les faiblesses grâce à l'utilisation d'approches hybrides. Enfin, des données de séquence supplémentaires continueront d'améliorer les GA à mesure que d'autres exemples d'apprentissage deviendront disponibles. La précision, l'efficacité et le potentiel de croissance de la méthodologie basée sur l'AG fournissent une solution robuste à l'analyse de données en bioinformatique.
- Les chauves-souris sont des créatures mystérieuses qui utilisent l'écholocalisation pour localiser et capturer des proies sans observer visuellement leur environnement.

Dans ce chapitre, une tentative a été faite pour fournir un état de l'art sur le BA. Cet algorithme modélise le comportement de chasse aux proies des chauves-souris pour trouver des solutions aux problèmes d'optimisation. Il utilise le réglage de fréquence et a un contrôle dynamique de l'exploration et de l'exploitation en faisant varier les taux d'émission d'impulsions et le volume sonore pendant les itérations. L'ajustement fin de ces paramètres affecte son taux de convergence. Il a attiré de nombreux chercheurs de divers domaines des sciences de l'ingénieur qui travaillent à la résolution de problèmes d'optimisation. Depuis sa création en 2010, il a été appliqué avec succès pour résoudre de nombreuses applications industrielles et d’ingénieries distinctes et semble être un algorithme très prometteur. Cependant, l'algorithme en est encore à ses débuts et des développements et des recherches supplémentaires sont nécessaires pour améliorer ses caractéristiques de performance globales.

Chapitre 3
Extraction et Sélection de caractéristiques

3.1. Introduction

L'extraction de caractéristiques est une tâche inévitable, en particulier dans l'étape critique du prétraitement des séquences biologiques. Cette étape consiste par exemple à transformer les séquences biologiques en vecteurs de motifs où chaque motif est une sous-séquence pouvant être vue comme une propriété (ou attribut) caractérisant la séquence. Cette sortie peut être utilisée pour appliquer des outils d'apprentissage automatique standard pour effectuer des tâches d'exploration de données telles que la classification. Plusieurs travaux antérieurs ont décrit des méthodes d'extraction de caractéristiques pour la classification des séquences de protéines, mais aucun d'entre eux n'a discuté de la robustesse de ces méthodes et leurs influences aux résultats de la classification [SAI 12].

L'extraction de caractéristiques protéiques joue un rôle important dans les domaines de la prédiction des fonctions, structures, des interactions des protéines et de l'analyse de similarité des séquences protéiques. Cependant, chacune des méthodes existantes a ses avantages, ses point forts et souffre des limites et des lacunes, donc le choix de la méthode adéquate pour représenter un tel ensemble de données reste toujours une question ouverte et un problème non encore résolu.

Tandis que La sélection des caractéristiques est une étape importante pour l'amélioration du système de la classification, car les attributs qui serviront d'entrées à l'algorithme d'apprentissage en sont dérivés. L'étape a pour objectif de choisir un sous-ensemble d'attributs parmi tous les attributs disponibles. Un attribut est jugé pertinent si la machine peut l'utiliser pour créer une capacité de séparation entre les différentes classes. Parmi les algorithmes utilisés pour la sélection des fonctionnalités, on peut citer les algorithmes exponentiels, qui effectuent une recherche exhaustive dans l'ensemble de solutions pour déterminer la meilleure solution possible. Ce n'est pas une méthode viable, car le temps de calcul augmente de façon exponentielle. Une autre technique est la sélection séquentielle, telle que la sélection avant (forward selection) [DEA 06] et l'élimination arrière (backward elimination) [BEN 07]). Son inconvénient est de ne pas tenir compte des interactions entre les caractéristiques. Deux algorithmes évolutifs font partie des méthodes de recherche stochastiques: les algorithmes génétiques [HUA 06] et l'optimisation des essaims de particules (PSO) [HUA 08]. L'avantage des algorithmes bio-inspirés par rapport aux méthodes séquentielles est que les premiers prennent en compte les interactions caractéristiques, et les seconds non.

C'est pour ça, nous avons opté une méta-heuristiques très robuste, et a des résultats prometteurs dans différents domaines qui est l'algorithme des chauves souris (Bat), cette méthode est implémentée pour résoudre le problème de la sélection de meilleurs attributs.
En effet, grâce aux méthodes stochastiques telles que l'AG et l'algorithme bat, il est plus facile d'apporter des solutions optimales aux problèmes bioinformatiques complexes en association avec des méthodes computationnelles.

3.2. Extraction de caractéristique

Les séquences protéiques sont des résidus d'acides aminés consécutifs, et nous les considérons comme des chaînes de texte avec un alphabet A de taille | A | = 20. De nombreuses méthodes d'extraction de caractéristiques ont été développées au cours des dernières années. Généralement, les types de caractéristiques peuvent être divisés en sept groupes: composition en acides aminés, auto-corrélation, composition, transition et distribution, ordre des quasi-séquences, composition en pseudo-acides aminés [LI 06, CAO 13], descripteurs d'entropie de Shannon, et d'autres tels que la triade conjointe [WEI 16] et les valeurs-z. La complexité de calcul de l'extraction de caractéristiques varie en fonction de la méthode d'extraction de caractéristiques, de la longueur moyenne des séquences et du nombre de séquences. Selon [ISM 16], l'auto-corrélation et la composition en pseudo-acides aminés peuvent prendre plus de temps de calcul que les autres méthodes.
Le nombre d'éléments *N* dans le vecteur d'attributs varie selon les types de stratégie choisie. Généralement, le vecteur de caractéristiques *V* de la séquence protéique de l'indice d'ordre *i* peut être représenté par:

$$V_i = \{f_1, f_2, f_3, \dots, f_N\}$$

où f_j est la valeur de nombre de caractéristique de l'élément j. L'élément caractéristique peut également être appelé attribut, variable ou colonne et ces termes sont utilisés de manière interchangeable.
En règle générale, les caractéristiques sont extraites de plusieurs séquences. Si le nombre de séquences est désigné par *M* et *i = 1 à M* est l'indice d'ordre d'une séquence, les caractéristiques extraites d'un ensemble de séquences protéiques peuvent alors être représentées par une matrice comme suit:

$$\begin{pmatrix} f_{1,1} & f_{1,2} & f_{1,3} & \cdots & f_{1,N} \\ f_{2,1} & f_{2,2} & f_{2,3} & \cdots & f_{2,N} \\ f_{3,1} & f_{3,2} & f_{3,3} & \cdots & f_{3,N} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ f_{M,1} & f_{M,2} & f_{M,3} & \cdots & f_{M,N} \end{pmatrix}$$

En utilisant une notation différente, cette matrice peut également être appelée une table avec *M* lignes et *N* colonnes. Alors que *M* est déterminé par le nombre de séquences, *N* dépend des méthodes d'extraction de caractéristiques.

Dans les paragraphes suivantes, nous passerons en revue certaines des méthodes d'extraction de caractéristiques qui peuvent être calculées directement à partir de la séquence alphabétique.

3.2.1 La Composition en Acides Aminés

L'AAC repose sur la fréquence des acides aminés ou d'un ensemble d'acides aminés dans la séquence qui peuvent capturer certaines informations qui aident à prédire le sujet d'intérêt. L'AAC est une famille de caractéristiques discrètes qui ne dépendent pas de l'ordre des résidus et elle comprend des mono acides aminés, des dipeptides, des tripeptides ou tout réglage d'acides aminés dans n'importe quel ordre. Il a été prouvé que les AAC réussissaient à prédire les groupes homologues de séquences [HAR 73, COR 79]. Par conséquent, il peut être utile là où l'homologue l'implique, comme dans la classification des séquences protéiques, la fonction protéique et la prédiction de la structure. La composition en acides aminés du dipeptide et du tripeptide peut capturer plus d'informations locales que composition monoacide aminé.

3.2.1.1. Uni-amino Acid Composition (UAAC)

La composition en monoacides aminés est définie comme la fréquence de chaque 20 acide aminé dans la séquence protéique. Il est fréquemment utilisé et constitue la forme la plus simple de composition en acides aminés. Le nombre d'éléments dans le vecteur de caractéristiques est de 20. Chaque élément représente la fréquence relative d'un acide aminé. La formule générale de l'UAAC de chaque séquence protéique est la suivante (Equation 4.1) [BHA 04a, CAO 13]:

$$f(Ai) = \frac{N(Ai)}{N} \quad (3.1)$$

Où : *N* représente la longueur de la séquence.

N (Ai) est le nombre total d'acide aminé *Ai* présent dans la séquence.

Cette technique est très simple et facile à mettre en œuvre mais toutes les informations d'ordre de séquence sont perdues.

3.2.1.2. Composition de Dipeptide (DC)

Avec DC, une valeur de caractéristique représente la fréquence d'un dipeptide, qui est le nombre d'occurrences d'un acide aminé dans deux positions adjacentes dans une séquence protéique.

Par exemple dans la séquence: MALMACC les fréquences du dipeptide: MA, AL, LM, AC et CC sont respectivement de 2, 1, 1, 1 et 1. Le nombre de dipeptides possibles est de 400, soit le nombre d'éléments caractéristiques. Les caractéristiques DC sont normalisées en divisant les fréquences par (N-1), où N est la longueur de la séquence, et en la multipliant par 100 [CAO 13]. Les caractéristiques DC sont calculées en utilisant la formule (3.2).

$$f(a_i a_j) = \frac{N(a_i a_j)}{\sum_{i=1}^{20} \sum_{j=1}^{20} N(a_i a_j)} * 100 \quad (3.2)$$

Où:

$N(a_i a_j)$: représente le nombre total du dipeptide $a_i a_j$ présent dans la séquence.

$\sum_{i=1}^{20} \sum_{j=1}^{20} N(a_i a_j)$: représente le nombre total des dipeptides existent dans la séquence = *(N-1)*

Le dipeptide ajoute une nouvelle signification à la composition en acides aminés car la fréquence de deux acides aminés contigus peut capturer certaines informations d'ordre local [BHA 04b].

Par conséquent, la composition dipeptidique convient aux cas où une information localisée est requise, comme une information d'homologie [PET 93].

3.2.1.3. Composition de Tripeptide (TC)

La composition tri-peptidique, également appelée spectre 3-mère [CAO 13], est une extension de la notion de fréquence des acides aminés adjacents mis en œuvre avec la

composition de dipeptide ajoutant plus de valeur d'ordre local à mesure que la fréquence des trois acides aminés possibles est considérée. Le nombre d'éléments de caractéristiques dans ce cas est de 8000, ce qui est substantiellement important, ce qui est considéré comme l'un des inconvénients car la plupart des algorithmes d'apprentissage automatique ne favorisent pas les données de grande dimension en raison du problème de mauvais conditionnement qui peut survenir lors du calcul. et peut conduire à une mauvaise convergence ou à l'absence d'état de convergence. Malgré la faille, il a été prouvé qu'il convient avec certains algorithmes, qui implémentent la sélection de caractéristiques pour choisir uniquement les colonnes d'entités qui ont une capacité de prédiction et ignorer les autres. La composition tripeptidique peut être calculée comme illustré dans la formule (3.3) :

$$f(a_i a_j a_k) = \frac{N(a_i a_j a_k)}{\sum_{i=1}^{20} \sum_{j=1}^{20} N(a_i a_j a_k)} * 1000 \qquad (3.3)$$

Où :

$N(a_i a_j a_k)$ Représente le nombre de tripeptides $a_i a_j a_k$ présents dans la séquence.

$\sum_{i=1}^{20} \sum_{j=1}^{20} N(a_i a_j a_k)$: est le nombre total des tripeptides dans la séquence = *N-2*.

Telle que la composition de dipeptide et la composition de tripeptide, on pourrait continuer à considérer des AA plus contigus tels que la composition de tétra-peptide et de penta-peptide. Cependant, le nombre de caractéristiques devient considérablement élevé car dans le cas d'une composition tétra-peptidique, le nombre de caractéristiques devient 20^4 = 160 000, ce qui est extrêmement élevé.

3.2.1.4. Composite Amino acid composition

Les compositions d'acides aminés ci-dessus sont ajoutées ensemble pour former une composition d'acides aminés composite [CAO 13].

Le vecteur de caractéristiques sera représenté comme suit:

$$V = \{ f_1 , f_2 , f_3 , ... , f_{8420} \}$$

Où: $f_1 ... f_{20}$ est la résultat de la méthode UAAC.

$f_{21} ... f_{420}$ est la résultat de la méthode DC.

$f_{421} ... f_{8420}$ est la résultat de la méthode TC.

On peut remarquer que la composition en acides aminés ne met en œuvre aucune des propriétés physico-chimiques de l'acide aminé mais elle dépend uniquement des valeurs discrètes des fréquences des acides aminés.

3.2.2. Auto-corrélation

Comme son nom l'indique, l'auto-corrélation est la corrélation entre les valeurs d'une seule variable contrairement à la corrélation conventionnelle qui cherche à trouver la corrélation entre deux variables. L'auto-corrélation fait partie des statistiques spatiales utilisées pour la première fois pour détecter la non-aléatoire dans les données de séries chronologiques [BAR 71], la dépendance spatiale géographique ou la co-variation de propriétés dans l'espace géographique. L'application du concept d'auto-corrélation en tant que descripteur de protéine découle du fait qu'une séquence protéique peut être concevable comme un espace et la corrélation entre deux valeurs d'une propriété d'acide aminé à différentes positions dans la séquence protéique peut être évaluée positivement ou négativement corrélée. Avec cette notion de corrélation classique, l'auto-corrélation peut être exprimée mathématiquement en termes de formule de coefficient de corrélation de Pearson [PEA 95].

Compte tenu des valeurs des propriétés, $\{P_1 , P_2 , P_3 , ... , P_N\}$ pour la séquence $\{aa_1 , aa_2 , aa_3 , ... , aa_N\}$ où aa est un résidu d'acide aminé, la fonction d'auto-corrélation (Equation 3.4) peut être définie comme suit:

$$r_k = \frac{\sum_{i=1}^{n-k}(p_i - P)(p_{i+k} - P)}{\sum_{i}^{n}(p_i - P)^2} \qquad (3.4)$$

Où : *P*est la moyenne des indices de propriété.

Un certain nombre de descripteurs d'auto-corrélation ont été proposés pour les séquences protéiques [CAO 13, REN 10]. Tous s'accordent sur le fait qu'en tant qu'auto-corrélation, ils cherchent à trouver une corrélation entre deux valeurs dans la même séquence. Les valeurs, en fait, sont les valeurs des propriétés physico-chimiques des acides aminés. Dans un cas typique, chaque résidu d'acide aminé est remplacé par une valeur d'une propriété physico-chimique.

Habituellement, différents AA ont des valeurs différentes. La même propriété physicochimique est utilisée pour tous les résidus de séquence à la fois. Avant le calcul de l'auto-corrélation, les valeurs de la propriété *P* pour chaque acide aminé doivent être standardisées en utilisant la moyenne et l'écart type des valeurs de propriété [LI 06, ONG 07].

Les descripteurs de protéines d'auto-corrélation les plus courants comprennent l'auto-corrélation Geary, l'auto-corrélation Moran et l'auto-corrélation Moreau-Broto.

3.2.2.1. Auto-corrélation Moran

L'auto-corrélation Moran a été la première mesure de l'auto-corrélation pour étudier les phénomènes stochastiques distribués dans l'espace [MOR 50]. Comme le coefficient de corrélation de Pearson, ses valeurs vont de +1 (forte auto-corrélation positive), à 0 (un motif aléatoire) à -1 (forte auto-corrélation négative).

3.2.2.2. Auto-corrélation Geary

L'auto-corrélation Geary [GEA 54] est comprise entre 1 et 2. Il n'y a pas d'auto-corrélation si le coefficient est 1, auto-corrélation positive pour les valeurs comprises entre 0 et 1 et auto-corrélation négative si les valeurs comprises entre 1 et 2. Cependant, les valeurs sont parfois supérieures à 2 [HAI 89].

3.2.2.3. Auto-corrélation de Moreau-Broto

Ce descripteur peut être calculé en utilisant la formule (3.5) [BRO 84].

$$r_k = \sum_{i=1}^{n-k} (P_i P_{i+k});\ k = 1,2,3,\ldots \qquad (3.5)$$

Où :

k est le n-lag. P_i est la propriété d'AA standardisée du résidu *i* de la séquence protéique.

P_{i+k} est la propriété normalisée du résidu en position i + k.

La version normalisée de l'auto-corrélation de Moreau-Broto est l'équation ci-dessus (3.5) divisée par (n-k) [LI 06, ONG 07, CAO 13] comme montré dans l'équation (3.6).

$$r_k = x = \frac{\sum_{i=1}^{n-k} (P_i P_{i+k})}{n-k};\ k = 1,2,3,\ldots \qquad (3.6)$$

3.2.3. Les descripteurs locaux

Les descripteurs de composition, transition et distribution (CTD) [GOV 11] consistent en sept propriétés physicochimiques; hydrophobicité; volume de Van der Waals normalisé; polarité; polarisabilité; charge; structures secondaires; et accessibilité aux solvants. Pour chaque propriété physicochimique, les acides aminés sont répartis en trois groupes. Les acides aminés d'un même groupe sont considérés comme ayant la même propriété (tableau 3.1)

3.2.3.1. La Composition

La composition est définie comme le nombre de résidus d'acides aminés avec cette propriété particulière divisé par le nombre total d'acides aminés dans une séquence protéique. Il existe 21 caractéristiques (3 éléments d'attributs pour chacune des sept propriétés physico-chimiques).

3.2.3.2. La Transition

La transition est définie comme le pourcentage de fréquence avec lequel les résidus d'acides aminés ayant une propriété particulière sont suivis par des résidus d'une propriété différente. Le nombre d'éléments caractéristiques est de 21 (3 pour chacune des sept propriétés physico-chimiques).

3.2.3.3. La Distribution

La distribution est définie comme la longueur de la chaîne dans laquelle se situent respectivement les 1%, 25%, 50%, 75% et 100% des acides aminés ayant une propriété particulière.
Le nombre d'éléments caractéristiques pour la distribution est de 105 (15 pour chacune des sept propriétés physico-chimiques).

Propriétés	***Groupe 1***	***Groupe 2***	***Groupe 3***
Hydrophobicité	Polar R K E D Q N	Neutral G A S T P H Y	Hydrophobicité C L V I M F W
Normalize der Waals volume	0 - 2.78 G A S T P D	2.95 - 4.0 N V E Q I L	4.03 - 8.08 M H K F R Y W
Polarité	4.9 - 6.2 L I F W C M V Y	8.0 - 9.2 P A T G S	10.4 - 13.0 H Q R K N E D
Polarisabilité	0 - 1.08 G A S D T	0.128 - 0.186 C P N V E Q I L	0.219 - 0.409 K M H F R Y W
Charge	Positive K R	Neutral A N C Q G H I L M F P S T W Y V	Negative D E
Structure Secondaire	Hélice E A L M Q R H	Brin V I Y C W F T	Bobine G N P S D
Accessibilité aux	Enterré	Exposée	Intermédiaire

solvants	A L F C G I V W	R K Q E N D	M S P T H Y

Tableau 3. 1: Groupement des acides aminés pour CTD.

Par exemple, pour une séquence donnée: "MTEITAAMVKELRESTGAGA", il sera encodé comme "32132223311311222222" selon sa division d'hydrophobicité.

FEPS est une application Web pour l'extraction de caractéristiques protéiques qui calcule les caractéristiques séquentielles les plus courantes des protéines [ISM 16], cet outil calcule différentes saveurs de composition, transition et distribution comprenant les trois ensemble pour les sept propriétés physico-chimiques (147 éléments), uniquement la composition pour les sept propriétés (21 éléments), uniquement la transition pour les sept propriétés (21 éléments), uniquement la distribution pour les sept propriétés (105 éléments), Selon [GOV 11], le CTD s'est avéré efficace dans la prédiction de la localisation subcellulaire de la protéine.

3.2.4. Quasi-sequence-order descriptors

Les descripteurs d'ordre quasi-séquentiel sont proposés par Chou [CHO 00]. Ils sont dérivés de la matrice de distance entre les 20 acides aminés.

3.2.4.1. Sequence-order-coupling numbers (SOCN)

SOCN est un exemple de modèle hybride qui comprend à la fois un modèle discret et séquentiel [LI 06, ONG 07, CAO 13]. Les 20 premières caractéristiques sont la composition des acides aminés tandis que les caractéristiques de 21 et plus reflètent l'ordre de séquence en utilisant quatre propriétés physico-chimiques; hydrophobicité, hydrophilie, polarité et volume de la chaîne latérale. Les caractéristiques sont dérivées d'une matrice de distance créée en calculant la distance entre chaque paire des 20 acides aminés en utilisant la matrice de distance physicochimique de Schneider-Wrede [SCH 94] ou la matrice de distance chimique de Grantham [GRA 74]. Les 20 premières caractéristiques sont données par l'équation (3.7):

$$X_i = \frac{f_i}{\sum_{i=1}^{20} f_i + w \sum_{d=1}^{30} r_d} \quad (3.7)$$

Où: *i = 1,2,...20.* f_i est la fréquence normalisée de l'acide aminé *i, w* est un facteur de pondération (la valeur par défaut est 0,1) et r_d est le *d ème* rang du numéro de couplage de l'ordre de séquence, il est calculé par la formule suivante (3.8).

$$r_d = \sum_{i=1}^{n-d} (d_{i,\ i+d})^2;\ d = 1,2,3,\dots maxlag \quad (3.8)$$

Où *maxlag* est la valeur de décalage maximum (*maxlag* value).

Les caractéristiques de 21 et plus sont données par l'équation (3.9):

$$X_i = \frac{wr_{d-20}}{\sum_{i=1}^{20} f_i + w\sum_{d=1}^{30} r_d}; d = 21,22,\ldots maxlag \quad (3.9)$$

Le nombre d'éléments caractéristiques est déterminé par la valeur de décalage maximale *(maxlag).*

Remarque: *maxlag* est le décalage maximum et la longueur de la protéine ne doit pas être inférieure à *maxlag*.

3.2.5. La Composition en Pseudo Acides Aminés (PseAAC)

Il y a deux variantes de la composition en pseudo acides aminés, le *type 1* qui fournit des vecteurs de caractéristiques de dimension *(20 + λ)* et le *type 2* qui produit des vecteurs de caractéristiques de dimension *(20 + i × λ)* où λ représente le rang de l'ordre de la séquence, et *i* est le nombre des attributs d'AA sélectionnés.

3.2.5.1. PseAAC Type 1

La composition de pseudo acides aminés est proposée par Chou [CHO 01, SHE 08, CHO 11] est un autre exemple des caractéristiques hybrides qui combine à la fois des caractéristiques discrètes et séquentielles. C'est également une amélioration des numéros de couplage d'ordre de séquence (SOCN expliquée au-dessus). Les 20 premières caractéristiques représentent la composition discrète des acides aminés et les autres représentent les caractéristiques d'ordre de séquence calculées à l'aide de trois propriétés physico-chimiques; hydrophobicité *(H1)*, hydrophilie *(H2)* et masse de la chaîne latérale *(M)* des acides aminés. Les trois propriétés *H1, H2 et M* des 20 acides aminés sont standardisées comme illustré dans la formule (3.10) :

$$\dot{P}(i) = \frac{P(i) - \sum_{i=1}^{20} \frac{P(i)}{20}}{\sqrt{\sum_{i=1}^{20} \left[P(i) - \sum_{i=1}^{20} \frac{P(i)}{20}\right]}} \quad (3.10)$$

Où $P(i)$ est la propriété et $P(i)$ est la propriété standardisée d'un acide aminé.

Les caractéristiques sont ensuite calculées par un facteur de corrélation présenté dans l'équation (3.11) :

$$C_\lambda = \frac{1}{N - \lambda} \sum_{i=1}^{L-\lambda} \theta(R_i, R_{i+\lambda}) \quad (3.11)$$

Où : C_λ est le facteur de corrélation de premier niveau, qui indique l'ordre de séquence entre tous les résidus d'acides aminés les plus adjacents λ dans la séquence protéique (λ = *1, 2,…*

m) où *m* est la valeur λ maximale, *N* est le nombre d'acides aminés dans la séquence, et $\theta(R_i, R_{i+\lambda})$ est le facteur de corrélation et est donné par :

$$\theta(R_i, R_j) = \frac{1}{3}\{[H_1(R_j) - H_1(R_i)]^2 + [H_2(R_j) - H_2(R_i)]^2 + [M(R_j) - M(R_i)]^2\} \quad (3.12)$$

Où: $H_1(R_i)$, $H_2(R_i)$, $M(R_i)$ sont : l'hydrophobicité, l'hydrophilie et la masse de la chaîne latérale normalisées de l'acide aminé R_i .

Les 20 premières caractéristiques sont la composition en acides aminés et sont données par la formule (3.13):

$$X_i = \frac{f_i}{\sum_{i=1}^{20} f_i + w\sum_{d=1}^{30} C_j} \quad (3.13)$$

Où: f_i est la fréquence relative du type d'acide aminé *i*, *w* est un facteur de pondération (la valeur par défaut est 0,1) et *C* est le facteur de corrélation de premier niveau.

Les caractéristiques de 21 et plus sont données par l'équation (3.14):

$$X_i = \frac{wC_{d-20}}{\sum_{i=1}^{20} f_i + w\sum_{d=1}^{30} C_j}; d = 21,22,\dots m \quad (3.14)$$

Où: m est la valeur λ maximale [LI 06, CAO 13].

3.2.5.2. Amphiphilic PseAAC (PseAAC Type 2)

La composition en pseudo acides aminés amphiphiles (Am-PseAAC) a été proposée par Chou [CHO 01]. Am-PseAAC est également reconnue comme la composition de pseudo acides aminés de type 2. Les définitions de ces qualités sont similaires aux descripteurs PseAAC. À partir de $H_1(i)$ et $H_2(j)$ définis précédemment. Les fonctions de corrélation d'hydrophobicité et d'hydrophilie sont définies respectivement comme suit (formule 3.15):

$$H_{i,j}^1 = H_1(i)H_1(j)$$
$$H_{i,j}^2 = H_2(i)H_2(j) \quad (3.15)$$

D'après ces qualités, les facteurs d'ordre de séquence peuvent être définis dans les formules suivantes (Equation 3.16).

$$r_1 = \frac{1}{N-1}\sum_{i=1}^{N-1} H_{i,i+1}^1$$

$$r_2 = \frac{1}{N-1}\sum_{i=1}^{N-1} H_{i,i+1}^2$$

$$r_3 = \frac{1}{N-2}\sum_{i=1}^{N-2} H_{i,i+2}^1$$

$$r_4 = \frac{1}{N-2}\sum_{i=1}^{N-2} H_{i,i+2}^2 \qquad (3.16)$$

.....

$$r_{2\lambda-1} = \frac{1}{N-\lambda}\sum_{i=1}^{N-\lambda} H_{i,i+\lambda}^1$$

$$r_{2\lambda} = \frac{1}{N-\lambda}\sum_{i=1}^{N-\lambda} H_{i,i+\lambda}^2$$

Ensuite, un ensemble de descripteurs appelé Amphiphilic Pseudo Amino Acid Composition est défini comme suit (Formules (3.17), (3.18)) :

$$P_c = \frac{f_c}{\sum_{r=1}^{20} f_r + w\sum_{j=1}^{2\lambda} r_j} \qquad 1 < c < 20 \qquad (3.17)$$

$$P_c = \frac{wr_u}{\sum_{r=1}^{20} f_r + w\sum_{j=1}^{2\lambda} r_j} \qquad 21 < u < 20+2\lambda \qquad (3.18)$$

Où : *w* est le facteur de pondération et est pris comme *w = 0, 5* dans le travail de Chou KC [CHO 05].

3.2.6. Z-Values

Les valeurs z [SEC 07, DAV 07], également connues sous le nom de descripteurs Sandberg [SAN 98, LAP 02], sont les principaux composants de 26 propriétés physicochimiques différentes, mesurées et calculée à partir des acides aminés, et représentent essentiellement l'hydrophobicité / hydrophilicité (z1), stérique / propriétés de masse et polarisabilité (z2), polarité (z3) et les effets électronique (z4 et z5) des acides aminés [LAP 02].

Dans [SEC 07] 5 z-values sont utilisées pour représenter chaque acide aminé de la séquence protéique.

Par exemple, l'acide aminé Alanine (A) a des valeurs de 5-z: z1 = 0,24, z2 = -2,32, z3 = 0,60, z4 = -0,14 et z5 = 1,30. Par conséquent, une séquence protéique de longueur *N* serait représentée par *N * 5* caractéristiques. Dans [DAV 07, SEC 07], les auteurs suggèrent que les z-values pour tous les acides aminés de chaque protéine sont moyennées de sorte qu'une protéine soit représentée par seulement 5 z-values, au lieu de *5 * N*. Cela est nécessaire car la plupart des méthodes d'apprentissage automatique ne peuvent pas gérer des instances (dans ce cas, des protéines) qui ont un nombre variable de caractéristiques (dans ce cas, les z-

values). Il convient de noter qu'ils ont essayé des méthodes plus compliquées d'agrégation des z-values, mais ils ont obtenu de meilleurs résultats avec cette méthode plus simple.

À l'origine, dans [SEC 07], les auteurs ont utilisé les z-values moyennes de toute la séquence d'acides aminés. Après quelques recherches expérimentales, ils ont découvert que pour classer les protéines GPCR, il serait préférable d'utiliser 15 z-values [DAV 07]. Ces z-values sont ensuite calculées comme suit:

Les 5-values sont calculées et moyennées sur toute la séquence protéique. Autres 5 z-values sont calculées à partir de l'extrémité N-terminale (les premiers 150 acides aminés de la séquence protéique) de la protéine et autres 5 z-values sont calculées à partir de l'extrémité C-terminus (les 150 derniers acides aminés de la séquence protéique). Le nombre de 150 acides aminés a été trouvé, dans les expériences précédentes, pour donner la plus grande amélioration de la précision [DAV 07].

3.2.7 Descripteurs d'entropie de Shannon

Les séquences de protéines sont représentées par un vecteur de caractéristiques multidimensionnelles basés sur les caractéristiques de la théorie de l'information, ce type d'extraction contient deux catégories.

3.2.7.1 Entropie de Shannon

L'entropie de Shannon mesure la conservation des acides aminés dans les séquences, et elle a été largement utilisée dans la prédiction des modification post-traductionnelle des protéines [CAP 07]. Pour un peptide donné, il est calculé par la formule suivante (3.19)

$$En = -\sum_{i=1}^{20} p_i \log_2(p_i) \qquad (3.19)$$

Où : p_i représente la fréquence d'apparition de l'acide aminé *i* dans la séquence.

3.2.7.2. Entropie relative de Shannon

L'entropie relative de Shannon mesure la conservation des acides aminés par rapport à la distribution de fond. Il est calculé comme suit (formule 3.20). [WEI 16]

$$REn = -\sum_{i=1}^{20} p_i \log_2(\frac{p_i}{p_0}) \qquad (3.20)$$

Où: p_0 est la distribution uniforme de l'apparition d'un type d'acide aminé.

En récapitulation, le tableau suivant résume les méthodes de représentation de protéines utilisées dans notre travail ainsi que tous les travaux connexes.

Methods	**Used in**		**Size**
Amino Acid Composition "AAC"	Bhasin & Raghava [BHA 04] König & al. [KON13] Gao & al. [GAO 13] Gao & al. [GAO 06] Silla & Freitas [SIL 11] Kumar & al. [KUM 17]	Secker & al. [SEC 09] Shrivastava & al. [SHR 10] Naveed & Khan [NAV 11] Kumar & al. [KUM 15] Lin & Li [LIN 07]	20
Pseudo Amino Acid Composition "PseAAC" Chou in 2001	Li & al. [LI 10] Xiao & Qiu [XIA 10] Xiao & al. [XIA 11] Gu & al. [GU 15]. Naveed & Khan [NAV 12] Zia & Khan [ZIA 11]	Maq & Khan [MAQ 13] Ahmad &Al. [AHM 15] Khan & al. [KHA 15] Kumar & al. [KUM 15] Kumar & al. [KUM 17] Cheng &al. [CHE 17]	20+ i* λ
Amphiphilic Pseudo Amino Acid Composition Chou 2005	Zia & Khan [ZIA 11] Chou & Shen [CHO 06] Rehman & al. [REH 13]	Chou [CHO 05] Khan & al [Khan 10] Rehman & Khan [REH11]	20+2λ
Depiptide Composition	Bhasin & Raghava [BHA04] Li & al. [LI 10] Gao & al. [GAO 06] Naveed & Khan [NAV 11] Kumar & al. [KUM 17]	Khan & al. [KHA 17] Ahmad &Al. [AHM 15] Lin & Li [LIN 07] Feng & al. [FEN 14]	400
Local Descriptos	Silla & Freitas [SIL 11] Secker & al [SEC 10]	Tong & Tammi [TON 08] Cui et al. [CUI 07]	210

Tableau 3. 2: Synthèse des travaux utilisant les MRP pour la prédiction de la fonction des RCPG.

3.3. Sélection de caractéristiques

La sélection de caractéristiques (FS) est devenue l'art qui attire l'attention de plusieurs chercheurs durant ces dernières années, cette sélection permet d'identifier et d'éliminer les attributs qui pénalisent les performances d'un modèle complexe dans la mesure où elles peuvent être bruitées, redondantes ou non pertinentes. De plus, la mise en évidence des variables pertinentes facilite l'interprétation et la compréhension des aspects médicaux et biologiques [SET 13], ainsi, elle permet d'améliorer la performance de prédiction des méthodes de classification et de passer outre le fléau de la haute dimensionnalité de ces données (the curse of dimentionalty).

3.3.1. Définition du problème

La sélection de caractéristiques (FS) est généralement définie comme un processus de recherche pour trouver un sous-ensemble d'attributs pertinents parmi ceux de l'ensemble initial. la pertinence d'un sous-ensemble d'attributs dépend toujours des objectifs et des critères du système. En général, la notion du FS peut être définie par :

Soit $V = \{f_1, f_2, f_3, \ldots, f_N\}$ un ensemble d'attributs de taille N où N représente le nombre total de caractéristiques étudiées. Soit Ev une fonction qui permet d'évaluer un sous-ensemble d'attributs sélectionnés (3.21), l'objectif de la sélection est de trouver un sous-ensemble V_s (V_s c V) de taille N_s (N_s c N) tel que :

$$Ev(Vs) = \max_{Z \subset Vs} Ev(Z) \qquad (3.21)$$

Où : $|Z| = N_s$ et N_s est soit un nombre prédéfini par l'utilisateur ou soit généré par la méthode de sélection, comme il est indiqué dans le chapitre 6 dans notre proposition.

La figure suivante (3.1) illustre la procédure générale d'une méthode de sélection de caractéristiques, qui est proposée par Dash et Liu en 1997 [DAS 97]

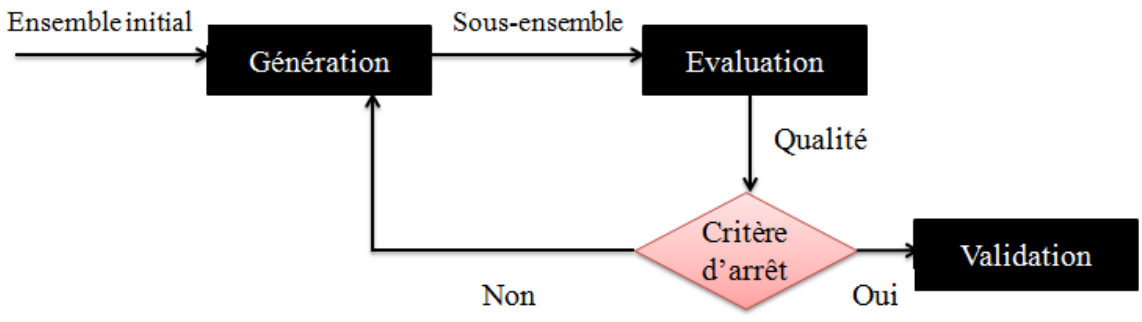


Figure 3. 1: Procédure général de la sélection de caractéristique.

Dans la littérature, il existe trois types des stratégies de FS:

- La première stratégie: La taille du sous ensemble est prédéfinie à priori et l'algorithme de sélection cherche à trouver le sous ensemble optimal de cette taille.
- La deuxième stratégie : consiste à sélectionner le plus petit sous ensemble dont la performance est plus grande ou égale à un seuil prédéfini dans le système.
- La troisième stratégie : cherche à trouver un compromis entre l'amélioration de la performance (par exemple dans notre travail la précision et le taux d'erreur de la classification) et la réduction de la taille du sous ensemble.

  Le but est de choisir le sous ensemble qui optimise les deux objectifs en même temps.

3.3.2. La pertinence d'une caractéristique

La performance d'un algorithme de classification dépend fortement des attributs utilisés dans la tâche d'apprentissage. La présence des attributs bruités ou redondants peut diminuer cette performance. Du coup, il existe plusieurs définitions de la pertinence d'un attribut, la plus connue est celle de John [JOH 94, JOH 97].

Les catégories de la pertinence des caractéristiques sont illustrées dans le tableau suivant (3.3).

Caractéristiques du f_i	Très pertinente	Son absence engendre une détérioration significative de la performance de classification
	Peu pertinente	$\exists V$ un sous ensemble d'attributs tq: $V_s U \{f_i\}$ soit significativement meilleur que la performance de V
	Non pertinente	Se sont les attributs bruités (inutiles), et ils seront supprimés de l'ensemble initial.

Tableau 3.3 Types de pertinence des caractéristiques.

3.3.3 Techniques d'évaluation

Les stratégies utilisées pour évaluer un sous-ensemble d'attributs dans les algorithmes de sélection de caractéristiques peuvent être classées en trois types principales : "Filter", Wrapper" et "Embedded" expliquées par la suite.

3.3.3.1 Type Filter

Ce modèle est le premier qui a été utilisé pour la sélection des attributs, Dans celui-ci, le critère d'évaluation utilisé évalue la pertinence d'une caractéristique selon des mesures qui reposent sur les propriétés des données d'apprentissage. Cette méthode est considérée, davantage comme une étape de prétraitement (filtrage) avant la phase d'apprentissage. En d'autres termes, l'évaluation se fait généralement indépendamment d'un classificateur [JOH 94]. La procédure du modèle "filter" est montrée dans la figure (3.2).

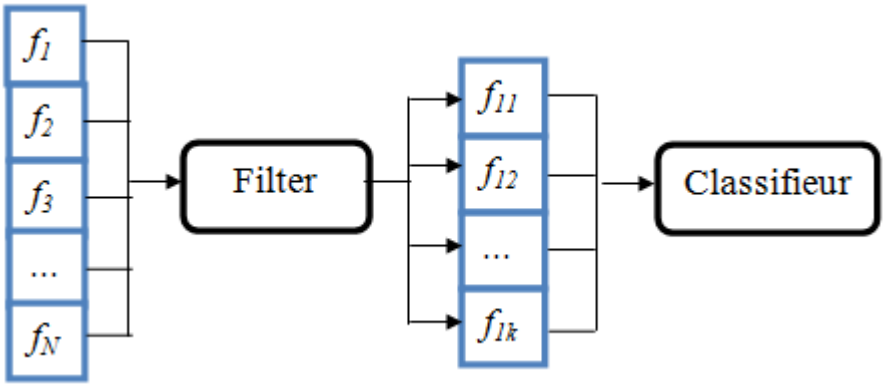


Figure 3. 2: Principe du modèle Filter.

Les caractéristiques sont généralement évaluées par des mesures calculées pour chacune d'elles, alors le but d'une méthode d'évaluation Filter est de calculer un score pour évaluer le degré de pertinence de chaque attribut f_i [GUY 03] . Les mesures les plus utilisées dans la littérature comme critère d'évaluation ou score sont :

- **Critère de corrélation :** Il est utilisé dans le cas d'une classification binaire [GUY 03].
- **Critère de Fisher :** Il permet de mesurer le degré de séparabilité des classe à l'aide d'une caractéristique donnée. [DUD 00, FUR 00].
- **L'information mutuelle :** C'est une mesure de dépendance entre les distributions de deux populations, et elle est définie comme la divergence de Kullback-Leibler (KL) [COV 91].
- **SNR :** Signifie "Signal to Noise Ratio coefficient", est un score qui mesure le pouvoir de discrimination d'une caractéristique entre deux classes [CHO 11 a].
- D'autre critères d'évaluation sont proposés dans [GOL 99, TUS 01, HAS 01].

Le principal avantage de ces méthodes est leur efficacité calculatoire et leur robustesse face au sur-apprentissage. Malheureusement, ces méthodes ne tiennent pas compte des interactions entre attributs et tendent à choisir des attributs comportant de l'information

redondante plutôt que complémentaire [GUY 03]. De plus, ces méthodes ne tiennent absolument pas compte de la performance des algorithmes de classification qui suivent la sélection [KOH 97].

3.3.3.2. Type Wrapper

Ces stratégies sont appelées aussi méthodes enveloppantes, évaluent un sous-ensemble de caractéristiques par sa performance de classification en utilisant un algorithme d'apprentissage. Elles ont été introduit par [KOH 97] pour résoudre le problème majeur des méthodes précédentes, qui est le fait qu'elles ignorent l'influence des caractéristiques sélectionnées sur la performance du classifieur à utiliser par la suite. La figure suivante (3.3) illustre le principe de fonctionnement du modèle Wrapper.

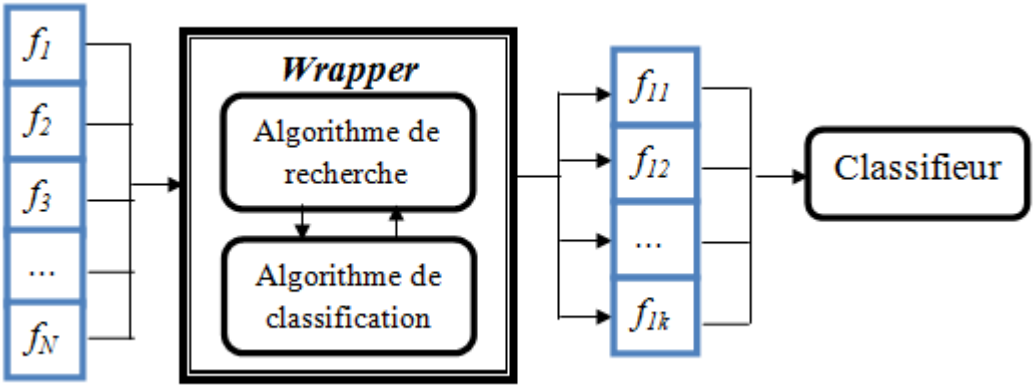


Figure 3. 3: Principe de fonctionnement du modèle Wrapper.

L'évaluation est effectuée à l'aide d'un algorithme de classification qui estime la pertinence d'un sous-ensemble donné de caractéristiques. A ce stade, les sous ensembles d'attributs sélectionnés par cette stratégie sont bien adaptés au classifieur utilisé, et ils ne sont pas forcement valides si on utilise un autre algorithme de classification. La complexité de l'algorithme d'apprentissage rend ces méthodes très coûteuses en terme temps d'exécution et espace mémoire. En général, pour diminuer le temps de calcul et pour éviter les problèmes de sur-apprentissage, le mécanisme de validation croisée (Cross Validation) est fréquemment utilisée. De plus, la complexité de cette technique peut être remédier par l'utilisation des méthodes de recherche heuristiques ou stochastiques. Une meilleure performance des méthodes "wrapper" par rapport à certaines méthodes de filtrage a été démontrée par Kohavi et John [KOH 97].

Les modèle "wrapper" est considéré comme étant meilleures que celui de filtrage selon [LI 08, HUA 08]. Elles sont capables de sélectionner des sous-ensembles d'attributs de petite

taille qui sont performants pour l'algorithme de classification utilisé, mais il existe deux inconvénients majeurs qui limitent ces méthodes :

- La complexité en terme espace mémoire et temps d'exécution nécessaire pour effectuer la sélection surtout pour les données massives.
- La dépendance des caractéristiques pertinentes sélectionnées par rapport au algorithme de classification utilisé.

3.3.3.3. Type Embedded

Contrairement aux méthodes précédentes, les stratégies Embedded (Méthodes intégrées) incorporent la sélection d'attributs lors du processus d'apprentissage. L'avantage principal de ces techniques est leur plus grande rapidité par rapport aux approches Wrapper puisqu'elles évitent que l'algorithme de classification recommence de zéro pour chaque sous-ensemble de caractéristiques. Les arbres de décisions sont l'illustration la plus emblématique [SET 13]. La procédure générale du principe de fonctionnement des méthodes Embedded est montrée dans la figure (3.4).

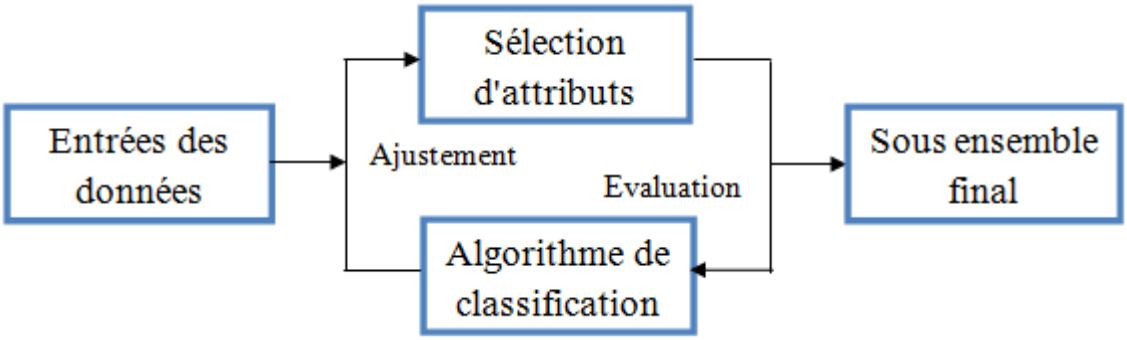


Figure 3. 4: Principe de l'approche Embedded.

3.3.4. Sélection de caractéristique dans la bioinformatique

Il existe diverses méthodes d'extraction de caractéristiques qui ont été rapportées dans la littérature sur l'analyse de séquence. Les travaux de King et al. [KIN 00] a été parmi les premiers à dériver des propriétés physiques et chimiques directement à partir de séquences pour prédire la fonction des protéines avec une précision raisonnable en utilisant les classificateurs ILP et C4.5. King et ses collègues [KIN 01] ont étudié plus en détail la représentation la plus appropriée d'une séquence protéique et leurs résultats ont montré que les attributs basés sur la phylogénie sont la représentation la plus précise d'une séquence protéique pour la prédiction fonctionnelle que la fréquence des résidus et la structure prédite.

Cependant, les problèmes liés à l'obtention des caractéristiques discriminatoires restent non résolus [RAH 09].

Parmi les travaux les plus cités dans la littérature, citons Jensen et al. [JEN 02] pour prédire la fonction des protéines en utilisant 14 attributs fonctionnels qui incluent des caractéristiques associées aux modifications post-traductionnelles (PTM) et au tri des protéines à l'aide de classificateurs de réseaux neuronaux. Les caractéristiques significatives utilisées par eux comprennent les sites O-ß-GlcNAc, la glycosylation N-liée, la structure secondaire, les propriétés physico-chimiques, etc. dérivées de la séquence protéique. Leur étude suggère que la séquence des acides aminés est plus bénéfique et directement liée aux caractéristiques fonctionnelles que la structure des protéines.

Les travaux de Cai et al. [CAI 03] et Han et al. [HAN 04] divisent 20 acides aminés en trois groupes fonctionnels différents, à savoir hydrophobe, neutre et polaire, basé sur différentes propriétés physiochimiques et la fonction enzymatique a été prédite à l'aide de classificateurs SVM. Ces études ont suggéré différentes combinaisons de caractéristiques et pour sélectionner un ensemble plus optimal de caractéristiques qui peuvent être réalisées en utilisant des méthodes de sélection de caractéristiques.

Dans ce contexte, Al-Shahib [AL-S 04, AL-S 05] a enquêté sur divers méthodes de FS sur un ensemble de caractéristiques dérivées de séquences d'acides aminés. Les caractéristiques comprennent les propriétés locales (par exemple, la fréquence et le nombre total de chaque acide aminé, hydrophobe, chargé, polaire, etc.) et les caractéristiques globales (par exemple le point isoélectrique et le poids moléculaire). Ils ont montré que la sélection des fonctionnalités est vitale et qu'une sélection rigoureuse améliore les performances du classificateur. Un autre travail de [UMA 07] en utilisant diverses caractéristiques de protéines, des propriétés de composition de base aux caractéristiques physico-chimiques, structure secondaire, motif, etc. Dans cette étude, ils ont utilisé un modèle de mélange d'arbres de décision stochastiques pour la prédiction de la fonction enzymatique.

L'un des travaux les plus récents dans ce domaine est celui de Lee et al. [LEE 08] qui a proposé une méthode pour générer de nouvelles caractéristiques pour la classification de la fonction ligase. Les caractéristiques présentent diverses informations locales de séquence protéique basées sur des résidus chargés positivement et négativement. Pour produire le sous-ensemble de caractéristiques optimal, ils ont utilisé une méthode FS basée sur la corrélation qui évalue un sous-ensemble de caractéristiques plutôt que des caractéristiques individuelles.

En général, ces études rapportées suggèrent que la FS doit être exécutée sur les caractéristiques et ce sont des caractéristiques qui sont discriminantes dans la classification de la fonction enzymatique. Parmi les méthodes FS, l'utilisation d'algorithmes d'ensembles approximatifs est encore rare et offre de nombreuses opportunités aux chercheurs intéressés.

3.3.5. Revue de quelques méthodes de FS

Dans cette section, nous présentons quelques méthodes de FS utilisées par plusieurs chercheurs et donnent de bons résultats. Nous avons choisi de présenter de méthodes fondées sur les différentes techniques d'évaluation expliquées précédemment.

3.3.5.1. Relief

Une des méthodes de filtrage les plus connues pour la sélection de caractéristiques est la méthode relief. Cette méthode fut proposée en 1992 par Kira et Rendell [KIR 92]. Son principe est de calculer une mesure globale de la pertinence des caractéristiques en accumulant la différence des distances entre des exemples d'apprentissage choisis aléatoirement et leurs plus proches voisins de la même classe et de l'autre classe. Le tableau 3.4 montre le pseudo code de cette méthode. La simplicité, la facilité de la mise en œuvre ainsi que la précision même sur des données bruitées, représentent les avantages de cette méthode. En revanche, sa technique aléatoire ne peut pas garantir la cohérence des résultats lorsqu'on applique plusieurs fois la méthode sur les mêmes données. Par ailleurs, cette méthode ne prend pas en compte la corrélation éventuelle entre les caractéristiques. Afin d'éviter le caractère aléatoire de l'algorithme, John et al. [JOH 94] ont proposé une version déterministe appelée ReliefD. D'autres variantes de cet algorithme, pour améliorer sa performance, sa vitesse ou les deux, ont été proposées dans [KOL 96, LIU 02].

Algorithme 1 : Sélection de caractéristiques par la méthode Relief
Entrées:
Une base d'apprentissage $A = \{X_1, X_2, ..., X_M\}$ où chaque exemple $Xi = \{x_{i1}, x_{i2}, ..., x_{iN}\}$
Nombre d'itérations T
Sorties:
$W[N]$: vecteur de poids des caractéristiques (f_i), $-1 \leq W[i] \leq 1$
$\forall i$; $W[i] = 0$;
Pour $t = 1$ *à* T **Faire**

Choisir aléatoirement un exemple X_k

Chercher deux plus proches voisins (un dans sa classe (X_a) et un deuxième dans l'autre classe (X_b))

Pour *i = 1 à N* **Faire**

$$W[i] = W[i] + \frac{|x_{ki} - x_{bi}|}{M * T} - \frac{|x_{ki} - x_{ai}|}{M * T}$$

Fin Pour

Fin Pour

Retourner *W*

Tableau 3.4 : Algorithme de Relief.

3.3.5.2. SAC (Sélective Adaptative de Caractéristiques)

C'est une méthode de sélection de descripteurs proposée par Kachouri et al. en 2010 [KAC 10]. Cette méthode, développée dans le cadre d'un ensemble de descripteurs a plusieurs dimensions, peut être adaptée pour une sélection de caractéristiques. L'idée générale de la méthode est de construire un ensemble de classificateurs SVM appris sur chacun des descripteurs et de sélectionner les meilleurs par discrimination linéaire de Fisher (FLD). Ils proposent de considérer la performance d'apprentissage des modèles correspondant à ces descripteurs pour l'identification d'une meilleure discrimination de Fisher. Le tableau 3.5 donne le pseudo code de cette méthode.

Algorithme 2 : Sélection de caractéristiques par la méthode SAC

Entrées:

Une base d'apprentissage $A = \{X_1, X_2, ..., X_M\}$ où chaque exemple $X_k = \{desc_{k1}, desc_{k2}, ..., desc_{kN}\}$, $k = 1...m$ et $X^i = \{desc_{1i}, desc_{2i}, ..., desc_{Mi}\}$, $i = 1...N$

Sorties:

M_s: les classifieurs retenus.

Pour *i = 1 à N* **Faire**

M_i = Apprentissage SV $M(X^i)$

$Pr(M_i)$ = taux de classification en utilisant le modèle M_i

Fin Pour

L = Trier $(Pr(M_i))$ par ordre décroissant $\forall i \in \{1,2,...,N\}$

$k = F\ LD(L)$

Retourner $M_s = (M_{s1}, M_{s2}, ..., M_{sk})$

Tableau 3.5: Algorithme de la méthode *SAC*.

3.3.5.3. Branch and bound

Ce type de méthode est lié à la modélisation du problème de recherche du meilleur sous-ensemble sous forme de graphe. Alors les algorithmes développés sur les graphes sont applicables, par exemple la méthode "Branch and Bound" (BB) . Cette méthode consiste à énumérer un ensemble de solutions d'une manière intelligente en ce sens que, en utilisant certaines propriétés du problème en question, cette technique arrive à éliminer des solutions partielles qui ne mènent pas à la solution que l'on recherche. Pour ce faire, cette méthode se dote d'une fonction qui permet de mettre une borne sur certaines solutions pour soit les exclure, soit les maintenir comme des solutions potentielles. Bien entendu, la performance de cette méthode dépend de la qualité de cette fonction d'évaluation partielle. Cette technique a été appliquée pour résoudre des problèmes de sélection de caractéristiques en 1977 par Narendra et Fukunaga [NAR 77]. Son principe est de construire un arbre de recherche où la racine représente l'ensemble des caractéristiques et les autres nœuds représentent des sous-ensembles de caractéristiques.

En parcourant l'arbre de la racine jusqu'aux feuilles, l'algorithme enlève successivement la plus mauvaise caractéristique du sous ensemble courant (nœud courant) qui ne satisfait pas le critère de sélection. Une fois que la valeur attribuée à un nœud est plus petite qu'un seuil (bound), les sous-arbres de ce nœud sont supprimés. Cette technique garantit de trouver un sous-ensemble optimal de caractéristiques à condition d'utiliser une fonction d'évaluation monotone. L'inconvénient de cette méthode est son temps de calcul qui croit vite avec l'augmentation du nombre de caractéristiques et qui devient impraticable à partir d'un certain nombre (30 caractéristiques). Une amélioration de cette méthode en utilisant d'autres techniques de recherche dans l'arbre afin d'accélérer le processus de sélection a été proposée dans [CHE 03, SOM 2004].

3.3.5.4. Les Algorithmes Génétiques

Les AG ont été utilisés dans le domaine de la sélection de caractéristiques afin d'accélérer la recherche et d'éviter les optima locaux. De nombreuses études rapportées dans la littérature ont montré que les méthodes qui utilisent les AG comme technique de recherche ont donné de meilleurs résultats que les résultats obtenus par les autres méthodes de sélection [JAI 97, KUN 99, ISH 00]. Le chapitre 2 était consacré à la présentation d'une description détaillée des AG ainsi que des méthodes de sélection qui utilisent ces techniques.

Le tableau 3.6 résume les inconvénients de toutes les méthodes de sélection de caractéristiques présentées ci-dessus

Méthode	**Type**	**Recherche**	**Non élimination de redondance**	**Non prise en compte ses interactions**	**Complexité**	**Dépendance à la fonction d'évaluation**
BB	Filter ou Wrapper	Heuristique			X	X
Relief	Filter	Aléatoire	X	X		
SAC	Hybride	Heuristique	X	X		X
AG	Filter ou Wrapper	Aléatoire			X	X

Tableau 3. 6: Résumé des méthodes de FS et leurs limites.

3.4. Conclusion

Dans ce chapitre, nous avons présenté les concepts d'extraction et de sélection de caractéristiques ainsi qu'un état de l'art des méthodes et algorithmes usuels pour résoudre ces problèmes. La littérature abondante depuis plusieurs décennies sur le problème de sélection de variables (FS) témoigne non seulement sur son importance mais aussi sur ces difficultés, de choisir a priori les caractéristiques pertinentes pour une application donnée n'est pas aisé et plus spécifiquement dans le domaine biologique.

L'optimisation de la prédiction des fonctions de protéines se concentre essentiellement sur le bon choix de la méthode d'extraction de caractéristiques et l'algorithme de classification approprié ainsi que l'amélioration des vecteurs d'attributs (Echantillons) à classer en utilisant l'approche de FS.

Ce domaine de recherche restera toujours actif tant qu'il est motivé par l'évolution des systèmes de collecte et de stockage des données d'une part et par les exigences des experts d'autre part.

Deuxième partie

Problèmes et propositions

Chapitre 4
Contribution 1 : Etude analytique des stratégies d'extraction de caractéristiques

4.1. Introduction

Les séquences de protéines sont des chaines alphabétiques de différentes tailles, leur classification nécessite une étape de prétraitement. Cette dernière assure la transformation de cette chaine alphabétique en un vecteur d'attributs numériques, en utilisant l'une des stratégies existantes.

Dans le chapitre précédent, section (3.2), nous avons détaillé les MRP les plus utilisées dans la littérature, chacune a ses avantages et ses lacunes ce qui mis l'amélioration de la précision de la classification toujours en question.

Ce chapitre est consacré à étudier et analyser les stratégies les plus utilisées dans les travaux de la classification des RCPG, cette analyse dépend étroitement à l'utilisation de plusieurs algorithmes d'apprentissage automatique afin d'effectuer à chaque fois une classification des vecteurs numériques et évaluer le taux de précision obtenu.

4.2. Contributions et proposition

Notre principale contribution est de clarifier l'utilité et la tâche primordiale des MRP dans l'amélioration de la qualité de la classification. Nous cherchons aussi à réaliser une étude critique inter-domaine afin de créer une taxonomie des stratégies d'extraction de caractéristiques qui sont actuellement dispersés dans différentes applications protéomique. Cette taxonomie se base essentiellement sur l'identification des similitudes et les importantes différences entre ses stratégies en utilisant différents classifieurs.

Étant donné que la plupart des travaux existants utilisent une méthode de représentation des protéines pour la classification des RCPG, il y en a d'autres qui font la comparaison entre deux ou trois méthodes. De nombreux auteurs assertent que certaines MRP sont meilleurs que d'autres, mais ils utilisent souvent un nombre réduit d'algorithmes de classification pour décider une bonne évaluation.

Notre proposition est de faire une comparaison analytique basée apprentissage. Dans cette étude, nous nous sommes intéressés aux méthodes de représentation de protéines utilisées dans la prédiction de la fonction des RCPG qui sont : AAC, PseAAC, Am-PseAAC, DC et LD. Le choix de ces méthodes est justifié par le fait qu'elles sont diverses en terme de taille et de contenu (informations gardées dans la séquence numériques). Tandis que les algorithmes de fouille de données (AFD) sélectionnés pour effectuer la tâche de classification des chaines protéiques sont : Bayesien Network (BN), Naive Bayes (NB),

SVM, MLP, IBK, J48, Tables de Décision (DT) et ZeroR (ZR), tous ces classifieurs sont implémentés dans l'environnement Weka que nous allons expliquer par la suite.

Le choix de ces algorithmes est justifié par le fait qu'ils sont implémentés dans le même environnement et ils ont un principe de fonctionnement différent pour bien tester la variation de la précision et le taux d'erreur de la classification.

Ces classifieurs sont connus pour leurs capacités d'apprentissage, d'optimisation et de mémorisation, des caractéristiques qui peuvent être d'une grande efficacité pour une évaluation crédible et pertinente.

La figure suivante (4.1) présente les étapes nécessaires de notre proposition.

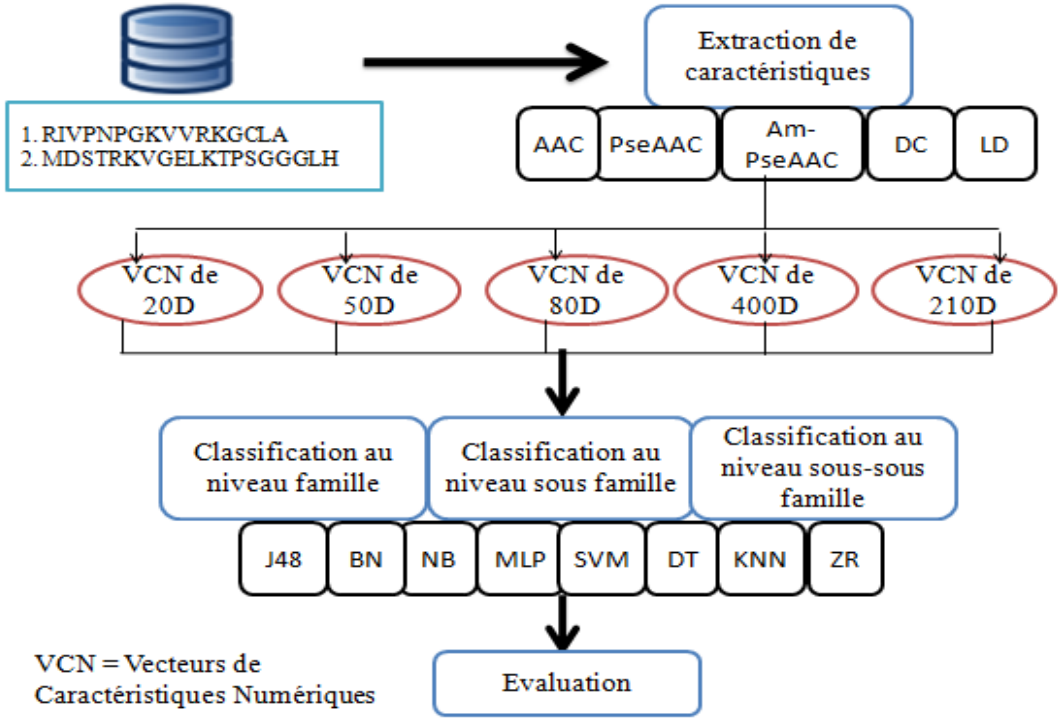


Figure 4. 1: Etapes du système proposé.

4.2.1. Prétraitement de données

Nous avons téléchargé les séquences de protéines à partir d'un fichier source, ce dernier peut être organisé de différentes manières mais généralement, il a un format non structuré. L'étape de prétraitement sert à transformer ce fichier source en un fichier cible. Ce dernier est représenté par une base de données, où nous pouvons sauvegarder toutes les chaines protéiques ainsi toutes leurs informations nécessaires, afin d'obtenir une représentation adéquate et exploitable par notre système.

4.2.2. Extraction de caractéristiques

Typiquement, n'il y a aucun algorithme de fouille de données capable de traiter des entrées alphabétiques (non numériques), donc il faut effectuer une étape de représentation de protéines en utilisant l'une des stratégies d'extraction de caractéristiques. Dans notre travail, nous proposant d'utiliser cinq MRP afin de bien localiser les différences dans les résultats de la prédiction et le taux d'erreur obtenus.
Dans les prochaines sections, nous allons détailler ces méthodes ainsi que l'outil utilisé.

4.2.3. Classification

Une fois le prétraitement est terminé et les données sont prêtes à être utilisées, on entame une classification hiérarchique à trois niveaux (famille, sous famille et sous sous-famille) en utilisant plusieurs classifieurs tel qu'il est présenté dans la figure (4.1).

4.2.4. Evaluation

La dernière étape dans notre système est l'évaluation de chaque stratégie (MRP). Elle se fait en calculant la précision et le taux d'erreur de la classification dans chaque niveau séparément par les formules (07) et (10) respectivement. Le résultat obtenu nous permettra de distinguer la méthode de représentation de protéine et l'algorithme d'apprentissage automatique les plus pertinents.

4.3. Etude expérimentale

4.3.1. Outils et Méthodes

4.3.1.1. Base de données

La base de données que nous avons principalement utilisée pour l'apprentissage et l'évaluation de notre expérimentation a été téléchargée à partir du site Web [9]. Dans ce dernier, nous avons trouvé un fichier XML contenant 31500 séquences, dont seulement 10200 contient une famille, une sous-famille et une sous-sous-famille. L'utilisation d'un fichier pour manipuler et explorer ces informations nous semblait coûteuse en terme d'efforts de programmation, pour cela nous avons préféré de transformer ce fichier en une base de données pour organiser les séquences dans une table contenant plusieurs colonnes.

4.3.1.2. L'environnement Weka

Waikato Environment for Knowledge Analysis "Weka" est un ensemble d'outils pour manipuler et analyser des fichiers de données, implémentant la plupart des algorithmes d'intelligence artificielle, des arbres de décision et des réseaux de neurones. Il est écrit en java, disponible sur le site [10].

Nous avons utilisé Weka pour les raisons suivantes:

- Facile à utiliser à cause de la disponibilité de son guide avec toutes les mises à jour.
- Efficace dans le traitement des données bruitées ayant une nature difficile (mauvaise représentation, données manquantes, ...etc).
- Répondre à nos besoins pour évaluer la classification;
- Intégration de plusieurs modules tels que : la sélection d'attributs, la classification supervisée et clustering.

4.3.1.3. Protr

Le package protr offre une boîte à outils unique et complète pour générer divers schémas de représentation numérique de séquences de protéines. Il est disponible gratuitement sur le Réseau d'archives R complet « Comprehensive R Archive Network » [11]. Cette vignette correspond à la version protr 1.1-0 et a été composée le 29/12/2015. Généralement, chaque type de descripteurs (caractéristiques) peut être calculé avec une fonction nommée extractX () dans le package protr, où X représente l'abréviation du nom du descripteur.

Dans notre travail, Nous avons utilisé l'outil « protr » pour générer les différents vecteurs d'attributs numériques des méthodes suivantes :

a. La méthode AAC

La fonction extractAAC() permet d'extraire les valeurs numériques de chaque séquence alphabétique stockées dans la base de données en utilisant la formule (3.1). Ce vecteur contient 20 éléments qui réfèrent la fréquence des 20 acides aminés dans la séquence, comme le montre le tableau 4.1.

Attributs	A	R	N	D	C	E	Q	G	H	I
Valeurs	0.04	0.05	0.03	0.02	0.04	0.03	0.02	0.03	0.03	0.05
Attributs	L	K	M	F	P	S	T	W	Y	V

Valeurs	0.16	0.02	0.01	0.09	0.05	0.11	0.06	0.02	0.04	0.09

Tableau 4. 1: vecteur d'attributs de 20D.

b. La méthode PseAAC

En utilisant la fonction extractPAAC(), un vecteur de 50 éléments sera calculé. Le tableau suivant montre uniquement les 30 attributs obtenus de la première séquence dans notre base de données.

Xc1.A	Xc1.R	Xc1.N	Xc1.D	Xc1.C	Xc1.E	Xc1.Q	Xc1.G	Xc1.H	Xc1.I
3.83	4.38	2.74	1.92	3.28	2.74	1.92	3.01	4.65	14.78
Xc1.L	Xc1.K	Xc1.M	Xc1.F	Xc1.P	Xc1.S	Xc1.T	Xc1.W	Xc1.Y	Xc1.V
2.19	0.82	8.48	4.10	9.58	5.20	2.19	3.83	8.48	0.02
Xc2.λ.1	Xc2.λ.2	Xc2.λ.3	Xc2.λ.4	Xc2.λ.5	Xc2.λ.6	Xc2.λ.7	Xc2.λ.8	Xc2.λ.9	Xc2.λ.10
0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02

Tableau 4. 2: Les 30 premiers attributs du vecteur produit de PseAAC.

c. La méthode Am-PseAAC

La fonction extracAPAAC() sert à calculer un vecteur ayant 80 attributs selon les deux propriétés hydrophobiques et hydrophiliques des acides aminés consécutifs. Le tableau ci-dessous donne les 40 premiers attributs de la séquence précédente.

Pc1.A	Pc1.R	Pc1.N	Pc1.D	Pc1.C	Pc1.E	Pc1.Q	Pc1.G	Pc1.H	Pc1.I
11.89	13.59	8.50	5.95	10.19	8.50	5.95	9.34	9.34	14.44
Pc1.L	Pc1.K	Pc1.M	Pc1.F	Pc1.P	Pc1.S	Pc1.T	Pc1.W	Pc1.Y	Pc1.V
45.87	6.80	2.55	26.33	12.74	29.73	16.14	6.80	11.89	26.33
Pc2.Hyd.1	Pc2.Hyd.1.	Pc2.Hyd.2	Pc2.Hyd.2.	Pc2.Hyd.3	Pc2.Hyd.3.	Pc2.Hyd.4	Pc2.Hyd.4.	Pc2.Hyd.5	Pc2.Hyd.5.
0.01	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.01	0.00
Pc2.Hyd.6	Pc2.Hyd.6.	Pc2.Hyd.7	Pc2.Hyd.7.	Pc2.Hyd.8	Pc2.Hyd.8.	Pc2.Hyd.9	Pc2.Hyd.9.	Pc2.Hyd.10	Pc2.Hyd.10.
0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00

Tableau 4. 3: Les 40 premiers attributs du vecteur produit de Am-PseAAC.

d. La méthode DC

En utilisant la fonction extractDC() qui est similaire à la fonction extractAAC(), un vecteur de 400 attributs numériques sera obtenu. Les 50 premiers attributs de ce vecteur sont montrés dans le tableau suivant.

AA	RA	NA	DA	CA	EA	QA	GA	HA	IA	LA	KA	MA	FA	PA	SA	TA
0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
WA	**YA**	**VA**	**AR**	**RR**	**NR**	**DR**	**CR**	**ER**	**QR**	**GR**	**HR**	**IR**	**LR**	**KR**	**MR**	**FR**
0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.01
PR	**SR**	**TR**	**WR**	**YR**	**VR**	**AN**	**RN**	**NN**	**DN**	**CN**	**EN**	**QN**	**GN**	**HN**	**IN**	
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	

Tableau 4. 4: Les 50 premiers attributs du vecteur produit de DC.

Nous avons également implémenté la méthode de descripteurs locaux pour produire un vecteur d'attributs numériques de dimension 210D comme nous le détaillerons par la suite.

4.3.1.4. La méthode des Descripteurs Locaux (LD)

Cette méthode transforme une séquence alphabétique en un vecteur de dimension 210D. Elle se décompose en deux étapes:

- La première étape consiste à transformer la chaine de protéine P_i en une autre chaine secondaire *P'i* formée uniquement des symboles *H, N* et *P* selon l'appartenance de chaque acide aminé à son groupe approprié.
- La deuxième étape comporte des traitements de calcul sur cette dernière chaine pour obtenir les valeurs des descripteurs *C, T* et *D.* Nous allons détailler ces deux étapes par la suite.

a. Codage de la séquence

Les 20 naïfs acides aminés sont divisés en trois groupes fonctionnels, chaque AA est codé par un des indices suivants : *H= Hydrophobic*, *P= Polar*, *N= Neutral* selon le groupe auquel il appartient. La classification des AA correspondante se présente dans le tableau suivant.

Groupe	Nom	Acides aminés
H	Hydrophobic	C V L I M F W
P	Polar	R K E D Q N
N	Neutral	G A S T P H Y

Tableau 4. 5: Groupes des Acides Aminés.

Exemple : Pour une séquence *Pi* donnée*:*

M	**T**	**E**	**I**	**T**	**A**	**A**	**M**	**V**	**K**	**E**	**L**	**R**	**E**	**S**	**T**	**G**	**A**	**G**	**A**

Nous allons substituer chaque AA par l'indice de son groupe auquel il l'appartient. La séquence secondaire P'_i sera codée comme suit :

H	***N***	***P***	***H***	***N***	***N***	***N***	***H***	***H***	***P***	***P***	***H***	***P***	***P***	***N***	***N***	***N***	***N***	***N***	***N***

La position ou la variation de ces groupes dans la séquence représente la base de calcul des trois descripteurs locaux Composition (C), Transition (T) et Distribution (D).

b. Calcul des descripteurs

Cette étape sert à trouver les valeurs des descripteurs. L'entrée de cette étape est la chaine secondaire P'_i, à partir de cette chaine nous pouvons calculer les trois descripteurs *C, T* et *D* de chaque groupe d'acides aminés *H, N* et *P*. Dès que ce traitement est effectué, un vecteur d'attributs numériques sera obtenu. Comme les acides aminés sont divisés en trois groupes, le calcul des descripteurs *C, T* et *D* génère au total 21 attributs (3 pour *C*, 3 pour *T* et 15 pour *D*). Bien que cette technique soit valide si elle est appliquée tout au long de la séquence

d'acides aminés, nous avons divisé les séquences d'acides aminés en 10 régions. Chaque descripteur *C, T* et *D* est calculé sur les 10 sous-séquences, ce qui donne 210 attributs décrivant la protéine.

- ✓ **Le descripteur *"Composition C"***

Il représente le pourcentage global de chaque groupe dans la séquence, nous pouvons le définir par la formule (4.1):

$$C(x) = \frac{\mathrm{n(x)}}{\mathrm{N}} \qquad (4.1)$$

Tel que :

x : H, N, P.

n(x) : Le nombre d'acides aminés de type x.

N : La taille de la séquence.

Exemple: Nous voulons calculer la composition C(x) des trois groupes H, N et P de la séquence P'_i. Le résultat sera présenté dans le tableau suivant:

Group	Number of occurrences in sequence	Composition
H	05	5/20 = 0.25
P	05	5/20 = 0.25
N	10	10/20 = 0.5

Tableau 4. 6: Le résultat de calcul de descripteur C.

Donc pour la composition, nous avons extrait les trois premiers attributs composant notre vecteur qui sont: *C (H), C (P)* et *C (N).*

- ✓ **Le descripteur *"Transition T"***

La transition du groupe *H* à *P* est le pourcentage de fréquence à laquelle *H* est suivi de *P* ou *P* est suivi de *H* dans la séquence codée. Autrement dit, c'est la fréquence à laquelle les acides aminés appartenant à un groupe sont suivis par les acides aminés appartenant à un groupe différent. Nous pouvons le calculer en utilisant la fonction suivante (4.2):

$$\mathrm{T}_x = \frac{N_{rx} + N_{xr}}{\mathrm{N} - 1} \qquad (4.2)$$

x € (S = { H, P, N }); r = S - x;

N =La taille de la séquence; N_{xr} , N_{rx} = Le nombre des dipeptides (deux AA) encodés.

Exemple: Nous voulons calculer la transition T_x des trois groupes *H, N* et *P* de la séquence P'_i. Le tableau suivant montre les résultats obtenus:

Groupe	N_{rx}	N_{xr}	Transition
H	3	4	7/19 = 0.36
P	3	3	6/19 = 0.31
N	3	2	5/19 = 0.26

Tableau 4. 7: Le résultat de calcul du descripteur T.

Donc pour la transition, nous avons extrait les trois autres attributs composant notre vecteur qui sont: *T (H), T (P)* et *T (N).*

- ✓ **Le descripteur *"Distribution D"***

Le descripteur de distribution décrit la répartition de chaque groupe dans la séquence. Il existe cinq descripteurs de «distribution» pour chaque groupe et ce sont les pourcentages de position dans la séquence complète pour les premières, 25, 50, 75 et 100% des occurrences d'un groupe spécifié.

En résumé, pour chaque séquence donnée, nous l'avons initialement divisée en 10 régions après avoir calculé les descripteurs locaux C, T et D pour chaque groupe d'acides aminés. Donc, pour chaque région, nous avons 21 attributs qui sont:

C(H)	C(P)	C(N)	T(H)	T(P)	T(N)	D1(H)	D25(H)	D50(H)	D75(H)	D100(H)
D1(P)	D25(P)	D50(P)	D75(P)	D100(P)	D1(N)	D25(N)	D50(N)	D75(N)	D100(N)	

4.3.1.5. Mesures de performance

Pour mesurer les performances de notre méthode, le teste de jackknife est effectué sur l'ensemble de données. Dans le teste de jackknife, chaque séquence de protéines de l'ensemble de données est choisie à son tour comme échantillon de test et les séquences de protéines restantes sont utilisées comme ensemble de données d'apprentissage pour prédire l'étiquette de l'échantillon de test. Ainsi, ce processus est répété N fois pour un ensemble de données de N protéines. Comparé à d'autres méthodes de validation croisée, telles que le sous-échantillonnage et le test de jeu de données indépendant, le test jackknife est considéré comme le moyen le plus efficace et fiable.

Le test jackknife a été utilisé pour mesurer les performances de divers prédicateurs. Les mesures de performance utilisées pour l'évaluation des classificateurs sont: la précision globale, le coefficient de corrélation de Matthews (MCC) et F-measure. Ces mesures sont calculées à partir des paramètres suivants : VP (vrai positif) et VN (vrai négatif) sont les nombres d'échantillons positifs et négatifs correctement prédits, respectivement. FP (faux positif) et FN (faux négatif) sont les nombres d'échantillons positifs et négatifs mal prédits, respectivement.

Nous avons ajouté une autre mesure qui est le taux d'erreur pour calculer le taux d'erreur de classification pour contrôler sa variation dans les trois niveaux hiérarchiques. Quant à la classification du RCPG au niveau de la famille, sous-famille et sous-sous-famille, le taux

d'erreur, la précision, F-measure et MCC sont utilisés pour évaluer les performances de la classification.

Accuracy: Cette métrique mesure l'efficacité globale de l'algorithme, nous pouvons la calculer à l'aide de l'équation (4.7).

F-measure: C'est une mesure de la précision d'un test qui prend en compte à la fois la précision et le rappel du test pour calculer le score, cette valeur peut être interprétée comme une moyenne pondérée de la précision et du rappel, où une F-measure atteint sa meilleure valeur à 1 et son pire score à 0:

Précision: C'est le rapport entre le nombre de vrais positifs (VP) et la somme des vrais positifs et des faux positifs (FP). Une valeur de 1 exprime le fait que tous les exemples énumérés étaient vraiment positifs.

Précision = VP / (VP + FP) (4.3)

Rappel: Un rappel de 1 signifie que tous les exemples positifs ont été trouvés.

Rappel = VP / (VP + FN) (4.4)

Fmeasure = 2 * (Précision * Rappel) / (Précision + Rappel) (4.5)

MCC: Le coefficient de corrélation de Matthews prend des valeurs dans l'intervalle de {-1, 1}, une valeur de 1 signifie que le classifieur ne fait jamais aucune erreur, et une valeur de -1 signifie que le classifieur fait toujours des erreurs. Il se calcule par l'équation (4.6):

$$\boldsymbol{MCC} = \frac{(\mathrm{TP} * \mathrm{TN}) - (\mathrm{FP} * \mathrm{FN})}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (4.6)$$

ER: Le taux d'erreur de classification peut être calculé à l'aide de l'équation (4.10).

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}} \quad (4.7)$$

$$Sensitivity = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} * 100 \quad (4.8)$$

$$Specificity = \frac{\mathrm{TN}}{\mathrm{FP} + \mathrm{TN}} * 100 \quad (4.9)$$

$$Taux\ d'erreur = \frac{\text{Nombre d}'\text{exemples mal classés}}{\text{Nombre total d}'\text{exemples}} \quad (4.10)$$

4.3.2. Résultats expérimentaux

Dans cette section, nous aborderons d'abord l'impact des différentes MRP «moins étudié dans la littérature» sur la tâche de la prédiction de la fonction des protéines et nous discuterons également de l'impact des différents AFD. En outre, toutes les expérimentations

ont été réalisées en utilisant une validation croisée de 10 fois. Ici, nous visons à déterminer le changement des valeurs de performances lors de l'utilisation de différentes MRP et différents classifieurs. Les tableaux suivants montrent les résultats de la classification des RCPG à trois niveaux pour chaque MRP utilisée dans ce travail.

4.3.2.1. La méthode AAC

AFD	Niveau	C-C-Ins	Inc-C-Ins	ER	ACC	F-m	MCC
BayesNet	Famille	99.29%	0.71 %	0.007	0.99	0.99	0.98
	Sousfamille	95.9%	4.1%	0.04	0.97	0.96	0.95
	S-S-famille	88.5%	11.5%	0.11	0.9	0.88	0.88
NB	Famille	98.91%	1.09%	0.01	0.99	0.99	0.97
	Sousfamille	93.66%	6.34%	0.06	0.96	0.94	0.94
	S-S-famille	86.5%	13.5%	0.13	0.88	0.87	0.86
SVM	Famille	99.4%	0.6%	0.006	0.99	0.99	0.98
	Sousfamille	97.72%	2.28%	0.02	0.97	0.97	0.96
	S-S-famille	66.5%	33.5%	0.33	0.52	0.56	0.56
MLP	Famille	100%	0%	0	1.00	1.00	1.00
	Sous famille	99.98%	0.02%	0.0002	0.99	0.99	0.99
	S-S-famille	94.050%	5.95%	0.06	0.94	0.94	0.93
IBK	Famille	100%	0%	0	1.00	1.00	1.00
	Sous famille	99.98%	0.02%	0.0002	0.99	0.99	0.99
	S-S-famille	95.23%	4.77%	0.04	0.95	0.95	0.95
J48	Famille	99.6%	0.4%	0.004	0.99	0.99	0.99
	Sous famille	99.98%	0.02%	0.0002	0.99	0.99	0.99
	S-S-famille	90.5%	9.5%	0.09	0.9	0.9	0.9
Decision Table	Famille	99.96%	0.04%	0.0004	0.99	0.99	0.99
	Sous famille	99.98%	0.02%	0.0002	0.99	0.99	0.99
	S-S-famille	79.6%	20.4%	0.2	0.8	0.79	0.78
ZeroR	Famille	78.15%	21.85%	0.2	0.61	0.68	0.00
	Sous famille	30.12%	69.88%	0.69	0.09	0.14	0.00
	S-S-famille	23.5%	76.5%	0.76	0.05	0.09	0.00

Tableau 4. 8: Évaluation de la classification des RCPG en utilisant la méthode AAC.

À partir de ce tableau, nous remarquons que la classification au niveau de la famille en utilisant la méthode AAC donne de bons résultats tels que les valeurs de précision ont été incluses dans l'intervalle [0,97, 1] et que la valeur du taux d'erreur était comprise entre [0, 0,01], à l'exception du classificateur ZR qui donne les pires résultats pour toutes les stratégies et ce avec tous les classificateurs. Notons que les algorithmes MLP et IBK atteignent le résultat optimal.

Dans la classification au niveau de sous-famille, nous remarquons une légère diminution des valeurs de mesures de performance de sorte que les valeurs de l'ACC, ER soient comprises entre [0,96, 0,99] et [0,0002, 0,06] respectivement. Notons que les classificateurs MLP, J48 et DT ont donné les meilleurs résultats.

Puisqu'il y a une augmentation remarquable du nombre de classes, dans la classification au niveau des sous-sous-familles, on mentionne une diminution significative des valeurs de mesures de performance, Les valeurs ACC et ER deviennent comprises entre [0,52, 0,95] et [0,04, 0,33] respectivement. Cependant, le classificateur SVM a donné un mauvais résultat en utilisant la méthode AAC, les classificateurs IBK et MLP restent toujours les meilleurs.

4.3.2.2. La méthode PseAAC

AFD	**Niveau**	**C_C_In**	**Inc_Cl_In**	**ER**	**ACC**	**F-m**	**MCC**
BN	Famille	96.65%	3.35%	0.03	0.98	0.97	0.92
	Sous famille	87.8%	12.2%	0.12	0.92	0.88	0.87
	S-S- famille	83.77%	16.22%	0.16	0.88	0.84	0.84
NB	Famille	90.72%	9.28%	0.09	0.91	0.9	0.74
	Sous famille	81.9%	18.1%	0.18	0.89	0.83	0.82
	S-S-famille	77.98%	22.02%	0.22	0.83	0.77	0.78
SVM	Famille	99.09%	0.91%	0.009	0.99	0.99	0.97
	Sous famille	95.22%	4.77%	0.04	0.94	0.94	0.94
	S-S-famille	91.89%	8.11%	0.08	0.91	0.91	0.91
MLP	Famille	100%	0%	0	1.00	1.00	1.00
	Sous famille	99.98%	0.02%	0.0002	0.99	0.99	0.99
	S-S-famille	92.86%	7.13%	0.07	0.92	0.92	0.92
IBK	Famille	98.94%	1.06%	0.01	0.98	0.98	0.97
	Sous famille	92.06%	7.94%	0.08	0.92	0.92	0.9
	S-S-famille	88.04%	11.96%	0.12	0.88	0.87	0.87
J48	Famille	99.55%	0.45%	0.004	0.99	0.99	0.98
	Sous famille	99.98%	0.02%	0.0002	0.99	0.99	0.99
	S-S-famille	90.57%	9.43%	0.09	0.9	0.9	0.9
DT	Famille	99.96%	0.04%	0.0004	0.99	0.99	0.99
	Sous famille	99.98%	0.02%	0.0002	0.99	0.99	0.99
	S-S-famille	77.56%	22.43%	0.22	0.79	0.77	0.76
ZR	Famille	78.14%	21.86%	0.21	0.61	0.68	0
	Sous famille	30.12%	69.88%	0.69	0.09	0.14	0
	S-S-famille	23.53%	76.47%	0.76	0.05	0.09	0

Tableau 4. 9: Évaluation de la classification des RCPG en utilisant la méthode PseAAC.

Cette méthode produit un vecteur d'attributs de dimension 50D, du fait du choix de la valeur λ égale à 30. La classification des séquences RCPG au niveau de la famille en utilisant la méthode PseAAC donne un bon résultat pour la plupart des algorithmes d'apprentissage automatique utilisés dans ce travail. De plus les classificateurs BN, SVM, MLP, IBK, J48 et DT ont réalisé des valeurs de précision entre [0,98, 1] et les valeurs de taux d'erreur entre [0, 0,03]. En ce qui concerne l'algorithme NB, il donne une valeur de précision plus petite et un taux d'erreur plus élevé par rapport aux algorithmes précédents.

Dans la classification au niveau des sous-familles, les valeurs de précision et de taux d'erreur variaient respectivement entre [0,89, 0,99] et [0,0002, 0,18], mais nous mentionnons que le meilleur résultat est obtenu en utilisant les classificateurs MLP, J48 et DT. En ce qui concerne la classification au niveau de la sous-sous-famille, les mesures de performance sont diminuées de sorte que les valeurs de précision et de taux d'erreur deviennent respectivement entre les deux intervalles [0,79, 0,92] et [0,07, 0,22]. Notons que les classificateurs SVM et MLP donne les meilleurs résultats.

4.3.2.3.La méthode Am-PseAAC

AFD	Niveau	C-C-In	Inc-C- In	ER	ACC	F-m	MCC
BN	Famille	91.24%	8.76%	0.08	0.95	0.92	0.8
	Sous famille	81%	19%	0.19	0.86	0.82	0.79
	S-S-famille	72.56%	27.44%	0.27	0.77	0.73	0.72
NB	Famille	88.23%	11.77%	0.11	0.91	0.89	0.69
	Sous famille	70.57%	29.43%	0.29	0.82	0.71	0.7
	S-S-famille	66.48%	33.52%	0.33	0.74	0.65	0.66
SVM	Famille	91.44%	8.56%	0.08	0.91	0.91	0.75
	Sous famille	77.07%	22.93%	0.23	0.77	0.76	0.72
	S-S-famille	68.66%	31.34%	0.31	0.71	0.67	0.65
MLP	Famille	100%	0%	0	1	1	1
	Sous famille	99.98%	0.02%	0.0002	0.99	0.99	0.99
	S-S-famille	86.36%	23.64%	0.23	0.86	0.86	0.85
IBK	Famille	99%	1%	0.01	0.99	0.99	0.97
	Sous famille	92.78%	7.22%	0.07	0.92	0.92	0.91
	S-S-famille	80.33%	19.67%	0.19	0.8	0.8	0.78
J48	Famille	100%	0%	0	1.00	1.00	1.00
	Sous famille	99.98%	0.02%	0.0002	0.99	0.99	0.99
	S-S-famille	85.19%	14.81%	0.14	0.85	0.85	0.84
DT	Famille	99.96%	0.04%	0.0004	0.99	0.99	0.99
	Sous famille	99.98%	0.02%	0.0002	0.99	0.99	0.99
	S-S-famille	73.24%	26.76%	0.26	0.73	0.72	0.71
ZeroR	Famille	78.14%	21.86%	0.21	0.61	0.68	0
	Sous famille	30.13%	69.87%	0.69	0.09	0.14	0
	S-S-famille	23.53%	76.47%	0.76	0.05	0.09	0

Tableau 4. 10 Évaluation de la classification des RCPG en utilisant la méthode Am-PseAAC.

La classification au niveau de la famille à l'aide de la stratégie Am-PseAAC via les classificateurs MLP et J48 atteint une valeur de précision égale à 1 avec une valeur de taux d'erreur nulle, mais les algorithmes IBK et DT produisent une valeur de précision égale à 0,99 avec des valeurs de taux d'erreur minimales. Pour les autres classificateurs, les valeurs de précision et de taux d'erreur variaient entre [0,91, 0,95] et [0,08, 0,11] respectivement. Dans la classification au niveau de la sous-famille, les algorithmes MLP, J48 et DT donnent les mêmes et les meilleurs résultats, mais le classificateur IBK prend la deuxième place avec une petite diminution de la valeur de précision et une légère augmentation de la valeur du taux d'erreur. Les algorithmes restants ont une précision moins efficace et des valeurs de taux d'erreur comprises entre [0,77, 0,86] et [0,19, 0,29] respectivement. En ce qui concerne la classification au niveau de la sous-sous-famille, les valeurs de précision et de taux d'erreur variaient entre [0,71, 0,86] et [0,14, 0,33] respectivement. Nous pouvons mentionner que tous les classificateurs ont donné des résultats proches.

4.3.2.4. La méthode DC

AFD	Niveau	C-Cl-Ins	Inc-Cl-In	ER	ACC	F-m	MCC
BayesNet	Famille	98.13%	1.87%	0.01	0.98	0.98	0.95
	Sous famille	86.92%	13.08%	0.13	0.89	0.87	0.85
	S-S-famille	88.32%	11.68%	0.11	0.92	0.89	0.89
NaiveBayes	Famille	96.05%	3.95%	0.04	0.97	0.96	0.9
	Sous famille	83.75%	16.25%	0.16	0.86	0.84	0.81
	S-S-famille	86.7%	13.3%	0.13	0.91	0.88	0.87
SVM	Famille	95.63%	4.37%	0.04	0.92	0.93	0.85
	Sous class	86.13%	13.87%	0.13	0.81	0.82	0.81
	S-S-famille	64.28%	35.72%	0.35	0.49	0.54	0.54
MLP	Famille	99.93%	0.07%	0.0007	0.99	0.99	0.99
	Sous class	99.65%	0.35%	0.0035	0.99	0.99	0.99
	S-S-famille	96.74%	3.26%	0.03	0.96	0.96	0.96
IBK	Famille	97.79%	2.21%	0.02	0.97	0.97	0.94
	Sous class	94.41%	5.59%	0.05	0.94	0.94	0.93
	S-S-famille	92.05%	7.95%	0.08	0.92	0.92	0.92
J48	Famille	99.22%	0.77%	0.007	0.99	0.99	0.97
	Sous class	99.98%	0.02%	0.0002	0.99	0.99	0.99
	S-S-famille	90.82%	9.18%	0.09	0.9	0.9	0.9
D T	Famille	99.96%	0.04%	0.0004	0.99	0.99	0.99
	Sous famille	99.98%	0.02%	0.0002	0.99	0.99	0.99
	S-S-famille	81.26%	18.74%	0.18	0.83	0.81	0.8
ZeroR	Famille	78.14%	21.86%	0.21	0.61	0.68	0
	Sous famille	30.13%	69.87%	0.69	0.09	0.14	0
	S-S-famille	23.53%	76.47%	0.76	0.05	0.09	0

Tableau 4. 11: Évaluation de la classification des RCPG en utilisant la méthode DC.

Le calcul des fréquences de dipeptides dans une chaîne protéique produit un vecteur de 400 attributs numériques. La classification au niveau de la famille à l'aide de cette méthode via les algorithmes BN, NB, MLP, SVM, IBK, J48 et DT donne de bons résultats. Les valeurs

de précision et de taux d'erreur variaient entre [0,92, 0,99] et [0,0007, 0,04] respectivement. Dans la classification au niveau de la sous-famille, les classificateurs MLP, J48 et DT restent très efficaces avec une valeur de précision égale à 0,99 et un taux d'erreur maximal égal à 0,003, mais il y a une légère diminution des valeurs de mesure de performance pour le reste des algorithmes de sorte que la précision et la valeur du taux d'erreur se situaient respectivement dans [0,81, 0,94] et [0,05, 0,16]. La classification au niveau de la sous-sous-famille nécessite un algorithme puissant en raison du grand nombre de données ainsi qu'au grand nombre de classes. Cela justifie le fait que les performances des classificateurs SVM sont considérablement réduites où les valeurs de la précision et du taux d'erreur deviennent égales à 0,49, 0,35 respectivement. Idem pour le classificateur DT où les valeurs de précision et de taux d'erreur deviennent respectivement 0,83, 0,18. Notons qu'il y a une légère diminution des mesures de performance à l'aide des classificateurs MLP, IBK et J48. Contrairement aux algorithmes BN et NB où l'on constate une amélioration des mesures de performance par rapport à la classification au niveau de la sous-famille, de sorte que les valeurs de précision augmentent respectivement à 0,92, 0,91 et les valeurs de taux d'erreur deviennent respectivement égales à 0,11, 0,13.

4.3.2.5. La méthode LD

AFD	Niveau	Co-Cl-In	In-Cl-Ins	ER	ACC	F-m	MCC
BayesNet	Famille	83.42%	16.58%	0.16	0.94	0.87	0.7
	Sous famille	67.48%	32.52%	0.32	0.75	0.69	0.65
	S-S-famille	67.28%	32.72%	0.32	0.78	0.67	0.68
NaiveBayes	Famille	74.85%	25.15%	0.25	0.72	0.72	0.43
	Sous famille	48.8%	51.2%	0.51	0.6	0.44	0.4
	S-S-famille	51.54%	48.46%	0.48	0.56	0.49	0.49
SVM	Famille	81.84%	18.16%	0.18	0.85	0.76	0.37
	Sous famille	57.86%	42.14%	0.42	0.82	0.57	0.55
	S-S-famille	49.2%	50.8%	0.5	0.8	0.48	0.51
MLP	Famille	100%	0%	0	1	1	1
	Sous famille	99.98%	0.02%	0.0002	0.99	0.99	0.99
	S-S-famille	94.38%	5.62%	0.05	0.94	0.94	0.94
IBK	Famille	99.74%	0.26%	0.002	0.99	0.99	0.99
	Sous famille	98.9%	1.1%	0.01	0.98	0.98	0.98
	S-S-famille	94.2%	5.8%	0.05	0.94	0.94	0.94
	Famille	99.28%	0.7%	0.007	0.99	0.99	0.99

J48	Sous famille	100%	0%	0	1	1	1
	S-S-famille	90.65%	9.35%	0.09	0.9	0.9	0.9
DT	Famille	100%	0%	0	1	1	1
	Sous famille	99.96%	0.04%	0.0004	0.99	0.99	0.99
	S-S-famille	79.71%	20.29%	0.2	0.8	0.79	0.78
ZR	Famille	78.15%	21.85%	0.21	0.61	0.68	0
	Sous famille	30.13%	69.87%	0.69	0.09	0.14	0
	S-S-famille	23.53%	76.47%	0.76	0.05	0.09	0

Tableau 4. 12 Évaluation de la classification des RCPG en utilisant la méthode LD.

La classification au niveau de la famille à l'aide de la méthode des descripteurs locaux via les classificateurs MLP, DT, IBK et J48 donne un résultat optimal, avec des valeurs de précision comprises entre [0,99, 1] et des taux d'erreur compris entre [0, 0,007]. Les algorithmes restants s'avèrent être moins efficaces avec des valeurs de précision et de taux d'erreur comprises respectivement dans les intervalles [0,72, 0,94], [0,16, 0,25], en particulier les classificateurs NB et SVM. Similairement à la classification au niveau de la famille, les algorithmes MLP, IBK, J48, DT donnent toujours les meilleurs résultats dans la classification au niveau de la sous-famille, où J48 atteint les valeurs maximales, au niveau des classificateurs restants. Une réduction importante des métriques de performance est noté avec des valeurs de précision et de taux d'erreur se situant respectivement entre [0,6, 0,82], [0,32, 0,51]. Dans la classification au niveau de la sous-sous-famille, les algorithmes MLP et IBK donnent des résultats identiques et meilleurs, le classificateur J48 marque également un bon résultat avec une précision égale à 0,9. Cependant l'algorithme SVM atteint la même précision du classificateur DT qui est égale à 0,8, mais avec un taux d'erreur élevé, égal à 0,5. L'algorithme BN produit une valeur de précision égale à 0,78 avec un taux d'erreur égal à 0,32, tandis que l'algorithme NB marque une plus mauvaise précision, égale à 0,56.

Dans la section suivante, nous analyserons et comparerons la variation des performances de ces algorithmes de classification en utilisant chaque stratégie d'extraction de caractéristiques.

4.3.3. Discussion

Dans cette section, nous allons analyser chaque niveau de classification séparément pour poursuivre la variation des taux des mesures de performance. Les figures suivantes présentent l'évaluation de la classification au niveau famille, sous famille et sous sous-famille en utilisant les stratégies d'extraction de caractéristiques utilisées dans ce travail.

4.3.3.1.Classification au niveau famille

Dans ce niveau le nombre de classe est très petit, il existe réellement sept classes dont trois d'entre elles contiennent peu d'échantillons (exemples), ce qui nous conduit à les éliminer. Donc nous avons classé les 10200 séquences en 4 classes fonctionnelles comme suit : La classe A appelée Rhodopsin, la classe B1 appelée Secretin, la classe B2 appelée Adhesion, la classe C appelée Glutamate. La classe Frizzled (Classe F), Taste 2 (Classe T) et Other (Classe O), sont ignorés car la base de données de protéines actuelle ne contient pas suffisamment de séquences appartenant à ces classes.

Les figures (Figure 4.2 et Figure 4.3) montrent la variation des valeurs de Accuracy et de taux d'erreur au niveau famille en utilisant les cinq MRP et neuf AFD.

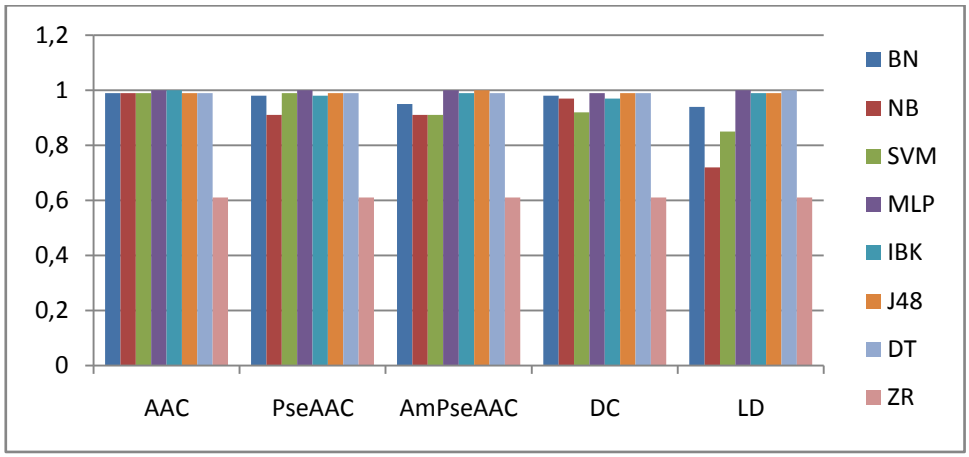


Figure 4. 2 Evaluation des valeurs de accuracy au niveau famille.

Selon la figure (4.2) toutes les barres qui représentent les classifieurs dans la stratégie AAC sont très proches et donnent des valeurs de Accuracy entre 0.99 et 1 sauf l'algorithme ZR qui produit un mauvais résultat (ACC=0.61). Ce dernier est stable pour toutes les MRP. Pour la méthode PseAAC, il est évident qu'il existe une petite diminution au niveau de la valeur de Accuracy en utilisant le classifieur NB par rapport à la stratégie précédente. Mais généralement tous les AFD utilisés donnent aussi un bon résultat. Concernant la stratégie Am-PseAAC, les algorithmes BN, NB, SVM marquent une baisse légère de la valeur de accuracy par rapport aux méthodes précédentes contrairement aux autres classifieurs qui restent stables. La représentation des protéines en utilisant la méthode DC donne des bonnes valeurs de accuracy pour la classification au niveau famille sauf l'algorithme SVM qui marque une petite diminution de 0.07. Quant à la méthode LD les classifieurs MLP, IBK, J48 et DT restent performants comme les méthodes précédentes, par contre SVM, BN et surtout NB marquent une différenciation (diminution) statistiquement significative.

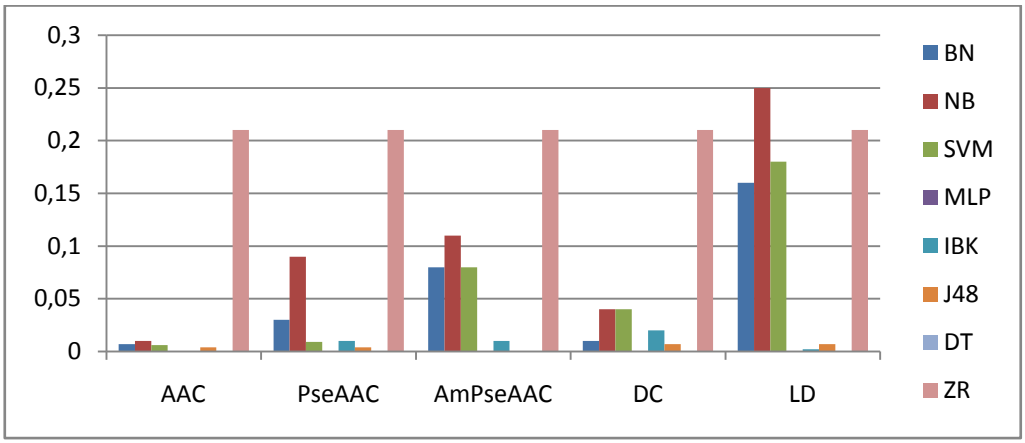


Figure 4.3: Evaluation des valeurs de taux d'errer au niveau famille.

Pour une bonne classification, le taux d'erreur doit être aussi minimal que possible. Les deux stratégies AAC et DC donnent de meilleurs taux d'erreurs par rapport aux méthodes restantes de telle sorte que les valeurs les plus mauvaises obtenues par l'algorithme NB sont 0.01 et 0.04 respectivement.

BN, NB et SVM produisent des taux d'erreurs plus élevés que les autres classifieurs en utilisant les méthodes Am-PseAAC, LD.

Nous avons un taux d'erreur tendant vers 0 obtenu à partir de classifieur MLP en utilisant toutes les stratégies d'extraction de caractéristiques, il est très performant pour la classification des RCPG mais il est couteux en terme de ressources (temps d'exécution et mémoire de stockage). L'algorithme ZR fournit un taux d'erreur égale à 0.21 qui est fixe quelque soit la méthode de représentation des séquences.

Tout cela illustre que toutes les stratégies d'extraction de caractéristiques peuvent être utilisées pour classer les RCPG au niveau de la famille avec des précisions élevées. Cependant, il serait toujours possible d'améliorer la précision de la classification en choisissant le bon AFD.

4.3.3.2. Classification au niveau sous famille

Nous avons classé les RCPG en 17 sous-familles. La classe A contenait 11 sous-familles qui sont: Protein_receptors, peptide_receptors, aminergic_receptors, nucleotide_receptors, lipid_receptors, orphan_receptors, alicarboxylic_acid_receptors, Other, melatonin_receptors et Steroid_receptors. Chacune des familles Secretin (B1) et Adhesion (B2) contenait une seule sous-famille qui sont : peptide receptors et adhesion receptors respectivement, et la

dernière classe Glutamate (C) comprend 4 sous-familles comme suit: Ion_receptors, Amino_acid_receptors, Sensory_receptors and orphan_receptors.

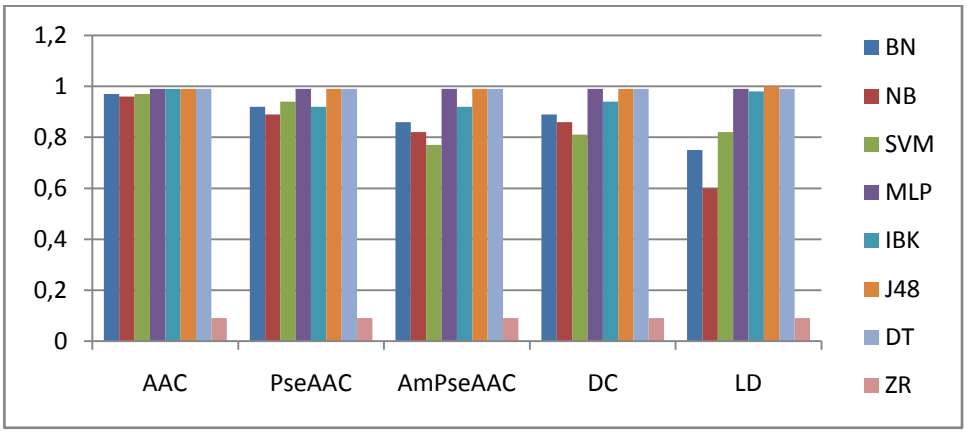


Figure 4. 4: Evaluation des valeurs de accuracy au niveau sous famille.

La classification du RCPG au niveau de la sous-famille par les algorithmes MLP, J48 et DT s'avère très efficace en raison des bonnes valeurs de ACC qui sont comprises entre 0.99 et 1 en utilisant toutes les stratégies, de telle sorte que le classifieur J48 atteint la valeur 1 en utilisant la méthode LD.

L'algorithme ZR donne une mesure fixe et la plus mauvaise qui est égale à 0,09 avec toutes les stratégies. Quant aux AFD restants, les valeurs de ACC sont en variation continue par rapport à la stratégie d'extraction de caractéristiques, tel que pour le classifieur IBK où nous avons trouvé un bon ACC en utilisant la méthode AAC ou LD. Mais pour BN et NB, la méthode AAC a donné un bon ACC qui est égal à 0,96 et LD a produit le pire qui est égal à 0,6.

Notons que la stratégie AAC donne toujours un bon résultat même au niveau sous famille quelque soit l'algorithme utilisé. L'algorithme SVM reste moins efficace en utilisant les méthodes Am-PseAAC, DC et LD.

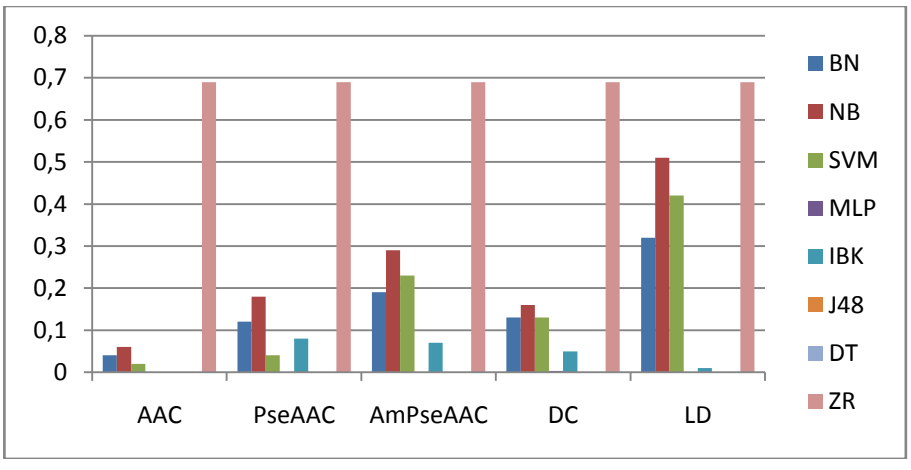


Figure 4. 5: Evaluation des taux d'erreurs au niveau sous famille

Idem à l'analyse précédente, la classification du RCPG au niveau de la sous-famille à travers les algorithmes MLP, J48 et DT a donné un taux d'erreur minimal compris entre [0.0002, 0,0004], de sorte que le classifieur J48 atteint la valeur zéro en utilisant la méthode LD. Un taux d'erreur plus élevé est obtenu en utilisant le classifieur ZR pour toutes les stratégies. Ce qui signifie que ce dernier a fait plusieurs erreurs au moment de la classification des séquences. Nous avons remarqué que l'algorithme IBK a donné un taux d'erreur minimum en utilisant les méthodes AAC et LD, mais les classifieurs BN, NB marquaient une valeur supérieure et mauvaise égale à 0.51, 0.32 respectivement en utilisant la méthode LD et une bonne valeur égale à 0.04, 0.06 respectivement en utilisant la méthode AAC.

En comparant ces résultats avec ceux de la classification au niveau de la famille, nous avons remarqué une légère augmentation du taux d'erreur et une légère diminution des mesures de performance. Cela est du à l'utilisation d'une énorme base de séquences avec l'augmentation du nombre de classes. Nous pouvons également constater que la méthode AAC, LD peut être utilisée pour la classification au niveau de la sous-famille en utilisant les classificateurs J48, DT et IBK grâce aux résultats optimaux.

4.3.3.3. Classification au niveau sous sous-famille

Les 10200 séquences de RCPG ont été classées en 69 sous sous-familles. En raison de ce grand nombre de familles, la classification à ce niveau nécessite beaucoup de temps de traitement. D'après les tableaux précédents, il est évident que toutes les mesures de performance sont considérablement diminuées par rapport aux niveaux supérieurs.

A partir des figures suivantes qui concernent les valeurs de ACC et du taux d'erreurs de la classification des RCPG en utilisant huit AFD pour les différentes stratégies d'extraction de caractéristiques utilisées, nous pouvons soulever les points suivants:

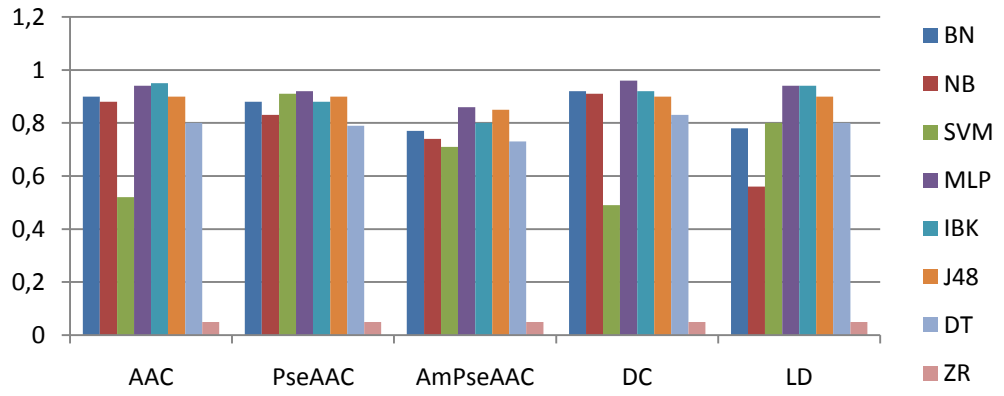


Figure 4. 6: Evaluation des valeurs de accuracy au niveau sous sous-famille

- Les algorithmes de classification au niveau sous sous-famille fournissent des résultats moins efficaces par rapport aux niveaux supérieurs à cause de l'augmentation du nombre de classes, ce qui conduit à un grand nombre de Faux Positifs (FP) et Faux Négatifs (FN) obtenus par le système.

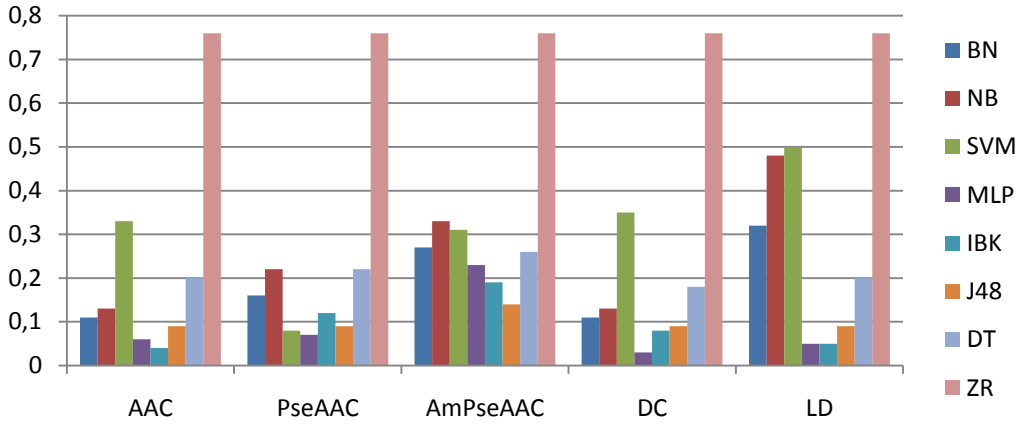


Figure 4. 7: Evaluation des taux d'erreurs au niveau sous sous-famille.

- L'algorithme IBK est plus efficace lors de l'utilisation des stratégies AAC, DC et LD.
- Le classifieur MLP fournit un bon résultat avec toutes les stratégies et dans les trois niveaux de classification.
- On peut remarquer une diminution significative dans les performances (ACC, ER) en utilisant l'algorithme SVM dans les méthodes AAC et DC.
- Ce dernier algorithme perd son efficacité dans la classification au niveau sous sous-famille par rapport aux niveaux supérieurs en utilisant les stratégies AAC et DC.
- Les algorithmes J48, DT ont été adoptés pour l'utilisation des méthodes AAC, PseAAC, DC et LD et ils sont moins efficaces dans la méthode Am-PseAAC à cause de leur mauvais taux d'erreur.
- Les classificateurs NB et BN produisent de mauvais résultats en utilisant les méthodes Am-PseAAC, LD.
- Le classificateur ZR est le pire pour toutes les stratégies quelque soit le niveau de classification.

Une stratégie d'extraction de caractéristiques donne un bon résultat avec certains algorithmes mais pas avec les autres. Donc à partir de ces résultats nous pouvons conclure que le choix d'une méthode de représentation variait selon le contexte de travail et surtout l'algorithme de classification utilisé car certains algorithmes sont sensibles au bruit, ambiguïté, flou des données.

4.4.Conclusion

Dans ce chapitre, nous avons présenté une étude empirique où différentes approches d'extraction de caractéristiques sont comparées à l'aide d'un pool précis et diversifié de classificateurs pour la tâche de prédiction de la fonction des RCPG. Nous obtenons un certain nombre d'observations statistiquement robustes concernant le comportement des différentes mesures de performance testées. Nous avons utilisé cinq méthodes de représentation de protéines avec huit algorithmes d'apprentissage automatique. Nous avons réalisé les expérimentations sur une base de RCPG contenant 10200 séquences connues. Les résultats obtenus montrent que l'inférence automatique de la fonction de protéines orphelines est une tâche difficile en protéomique, et qu'il y a deux problèmes majeurs dans la tâche de prédiction computationnelle de la fonction des protéines, qui sont le choix de la stratégie de représentation et le choix de l'algorithme de classification. Il existe plusieurs façons d'extraire des caractéristiques d'une protéine, et le choix de la représentation des caractéristiques peut être aussi important que le choix de l'algorithme de classification.

Nous visons à pallier avec ces deux problèmes en utilisant les méthodes bio-inspirées pour choisir le meilleur couple (MRP / AFD) qui optimise la classification et améliore les performances.

Chapitre 5
Contribution 2: Un algorithme évolutionnaire pour la sélection de meilleur couple (AFD\MRP)

5.1. Introduction

La section précédente se concentrait sur l'analyse de l'influence des MRP et les AFD sur la précision de la classification des RCPG. Nous avons pu conclure que le choix de la stratégie de représentation de caractéristiques peut être aussi important que le choix de l'algorithme de classification. De ce fait, pour optimiser la classification et obtenir une prédiction précise de la fonction des protéines, il faut garantir une bonne sélection du couple (MRP/ AFD) pour chaque ensemble de données utilisé.

Les algorithmes génétiques représentent l'une des plus grandes et des plus efficaces familles d'approches bio-inspirées de l'intelligence artificielle. Ils s'inspirent des processus de l'évolution naturels. Nous avons été attirés par cette approche à cause de ses résultats prometteurs dans différents domaines et surtout dans les problèmes d'intelligence computationnelle expliqués dans le chapitre précédent.

Ce chapitre est dédié à notre deuxième contribution principale, qui consiste en la proposition d'un algorithme génétique pour le choix du meilleur couple (MRP / AFD) qui nous garantit une classification pertinente.

Nous nous intéressons à traiter la partie de la classification au niveau sous-sous-famille en choisissant la meilleur représentation des séquences et l'algorithme convenable, qui nous produisent une classification optimale.

Ce chapitre est organisé comme suit : la Section 2 comporte notre proposition et l'explication de tous les détails de fonctionnement de l'algorithme génétique proposé pour la sélection (MRP/ AFD). La section 3 contient l'étude expérimentale, les outils utilisés pour sa réalisation et une discussion détaillée des résultats obtenus. La dernière section présente une conclusion et les futures recherches.

5.2. Choix du couple (MRP/ AFD) par l'algorithme génétique

Les approches des algorithmes génétiques ont connu un développement croissant au cours des dernières années avec des applications dans les domaines de l'intelligence artificielle, la reconnaissance de formes, la classification et la sélection de caractéristiques … etc. Plus récemment, ces méta-heuristiques ont été largement appliqués aux problèmes de fouille de données en bioinformatique afin de résoudre divers problèmes tels que : l'alignement des séquences multiple, la découverte de la structure de protéines, la recherche des motifs protéiques, et également la tâche la plus principale en bioinformatique, la prédiction de la fonction de protéine.

5.2.1. Architecture générale du système

La fonction de protéine se définit par rapport à un niveau structurel, et une protéine peut avoir plusieurs fonctions, au sein d'un même niveau et/ou entre niveaux différents. Pour faire face à cette nature hiérarchique des données biologiques, nous avons proposé un système de classification au niveau sous sous-famille, qui contient 69 classes.

L'objectif du système proposé est la sélection de la meilleure stratégie pour la représentation numérique des protéines ainsi que le bon algorithme de classification, à l'aide de l'algorithme génétique pour réaliser une classification précise et performante.

En général, les algorithmes évolutionnaires ont alors commencé à toucher de nombreux domaines. Pour qu'ils soient efficaces, il suffit de répondre à quelques contraintes :

- Le nombre de solutions potentielles doit être très grand.
- Il n'y a pas de méthode exacte permettant d'obtenir une solution.
- Une méthode presque optimale est acceptable.
- On peut évaluer la qualité d'une solution potentielle.

Si ces quatre contraintes sont vérifiées, alors un algorithme génétique peut s'avérer être une bonne solution pour trouver une réponse au problème qui, bien qu'elle ne puisse être garantie comme la meilleure, sera en tout cas acceptable, et ce dans un temps raisonnable.

Dans cette section, nous détaillerons notre proposition et les étapes nécessaires à sa réalisation. La figure suivante montre le schéma général du système proposé.

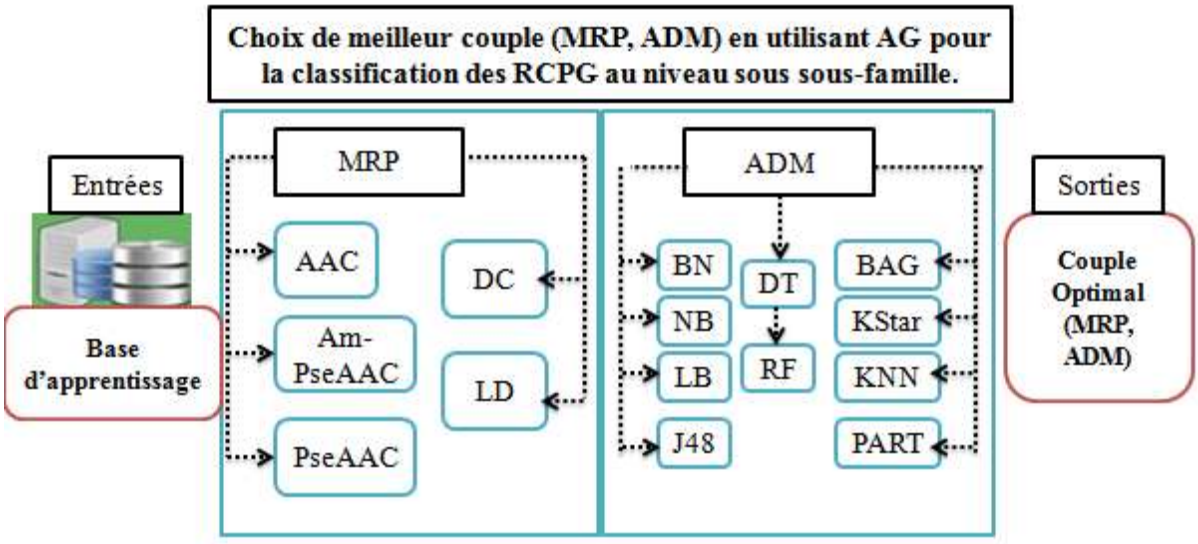


Figure 5. 1: L'architecture générale du système proposé.

Les algorithmes génétiques opèrent sur une population d'individus pour produire de meilleures approximations. A chaque génération, une nouvelle population est créée par le

processus de sélection d'individus en fonction de leur niveau d'aptitude dans le domaine problématique, et de leur recombinaison à l'aide d'opérateurs inspirés de la génétique naturelle. La progéniture pourrait également subir une mutation. Ce processus conduit à l'évolution de populations d'individus mieux adaptées à leur environnement que les individus à partir desquels ils ont été créés, tout comme dans l'adaptation naturelle. Un diagramme de flux pour le processus d'apprentissage avec l'algorithme génétique est représenté dans la figure suivante.

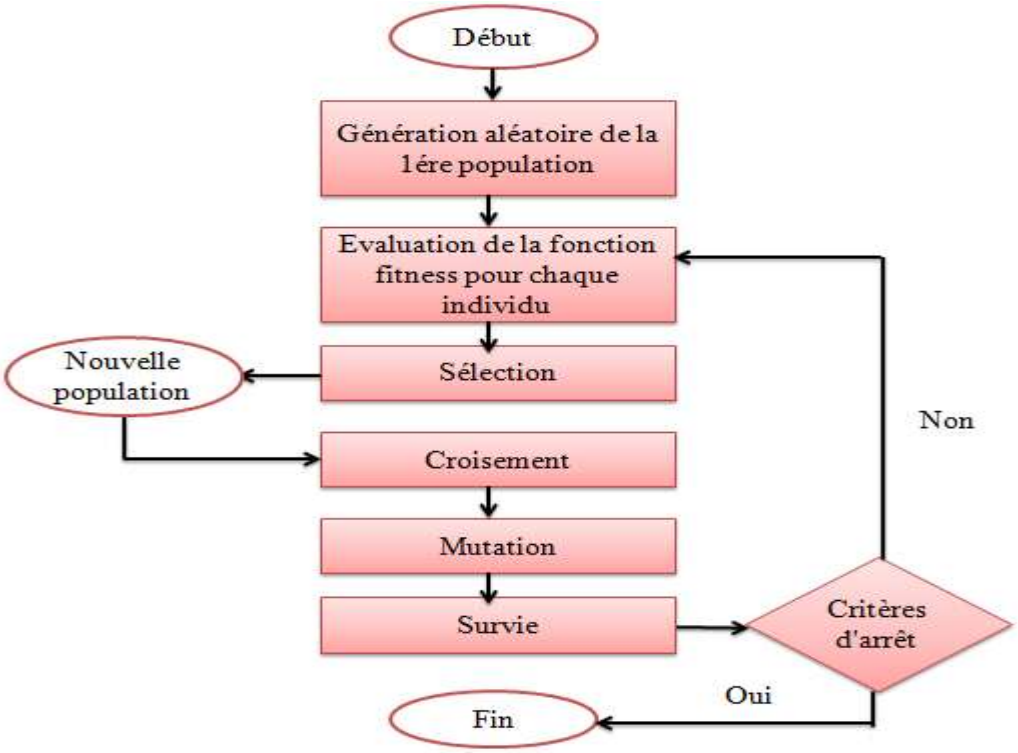


Figure 5. 2: Organigramme général de l'AG.

Les algorithmes évolutionnaires vont donc partir d'une population de solutions potentielles à un problème. Chacune est évaluée, pour lui attribuer une note, appelée fitness. Plus la fitness d'une solution est forte et plus celle-ci est prometteuse [MAT 14].

Les meilleurs individus sont ensuite sélectionnés, et se reproduisent. Deux opérateurs artificiels sont alors utilisés : le croisement entre deux individus, appelé crossover, et des mutations aléatoires. Une étape de survie s'applique alors pour créer la nouvelle génération d'individus.

Dans notre cas, chaque individu de la population représente un modèle prédictif. Le nombre de gènes est le nombre total de MRP et d'AFD utilisés. Les gènes ici sont des valeurs binaires, et représentent l'activation ou non de MRP, AFD particuliers dans le modèle. Le nombre d'individus, ou la taille de la population, doit être choisi pour chaque application. Habituellement, il est défini sur N * M, tel que:

N : le Nombre de méthodes de représentation de protéines utilisées;

M : le nombre d'algorithmes de classification.

"GA search" pour la sélection du couple (MRP, AFD) pertinent
Entrées:
MRP: AAC, PseAAC, Am- PseAAC, DC, LD.
DMA: BN, NB, J48, Bag, KStar, KNN, DT, LB, PART, RF.
MaxIteration: Itération maximale utilisée comme critère d'arrêt
ProbCross : Probabilité de croisement
ProbMut : Probabilité de mutation
Sortie:
Le meilleur ensemble (MRP, AFD)
Begin
Initialiser la Population
While not *MaxIteration* **do**
Evaluer la fonction fitness de chaque individu en utilisant l'équation N° 5.1.
Sélection des meilleurs individus
Appliquer l'opérateur de Croisement
Appliquer l'opérateur de Mutation
Survivre (créer la nouvelle population)
EndWhile.
End.

Tableau 5. 1: Algorithme Génétique pour la sélection du couple (MRP/AFD).

L'algorithme se termine généralement soit en atteignant un nombre de générations maximum ou en atteignant la stabilité de la meilleure fitness. Cette dernière peut être causée par un optimum local. Il y a plusieurs paramètres à choisir avant de commencer cet algorithme: en commençant par la représentation des chromosomes, puis en passant par le processus de choix du nombre de chromosomes dans la population initiale, en effectuant un calcul de fitness, en initialisant une sélection parentale, en définissant l'opérateur et le taux de croisement, l'opérateur et le taux de mutation.

Nous allons maintenant décrire en détail les opérateurs et les paramètres correspondants utilisés par l'algorithme génétique présenté dans le tableau précédent.

5.2.2. Codage des individus

La représentation des chromosomes est principalement choisie pour:

- S'adapter au problème d'optimisation (ici sélection (représentation / classificateur)) aux opérateurs de l'AG.
- Simplifier les expérimentations pour que la plupart des opérations se fassent sur la représentation et non sur l'ensemble des séquences.
- Optimiser le traitement et réduire la complexité.

Pour pouvoir atteindre ces objectifs, nous pensons que la solution proposée doit respecter les règles suivantes:

- **Proximité:** toute opération doit être proche de l'espace de représentation de telle sorte qu'après l'opération est effectuée sur un chromosome, le résultat doit rester dans les critères définis pour la représentation. Sinon, le processus peut être laissé avec certains chromosomes qui peuvent être tronqués ou ne pas être représentés du tout, ce qui s'éloigne de l'objectif de l'algorithme.
- **Représentativité:** Une représentation doit être capable de représenter n'importe quelle solution même dans un espace réduit (c'est-à-dire en fixant le nombre maximum d'espaces).
- **Invariance de séquence:** N'importe quelle opération effectuée sur une représentation ne doit endommager aucune solution d'origine.

Outre ces règles, le principal problème des algorithmes génétiques est le temps considérable d'exécution. La condition pour réduire le temps d'exécution est d'avoir des opérateurs d'algorithmes génétiques qui conduisent à une convergence, ce qui est difficile à réaliser compte tenu de la nature de ce type de fonctions. Les chercheurs tentent d'avoir cette convergence en sélectionnant les individus ayant la meilleure fitness; et le déclencheur pour arrêter l'exécution est soit un nombre maximum d'itérations, soit un niveau de seuil de valeur de fitness défini comme paramètres. Ces deux choix peuvent ne pas conduire à une solution acceptée.

Dans notre travail, nous avons codé chaque chromosome par 15 gènes binaires. Les cinq premiers gènes sont consacrés à la stratégie de représentation des séquences, et les gènes restants sont dédiés aux algorithmes de classification utilisés, la figure suivante (5.3) illustre le codage des individus effectué. Chaque gène positif signifie que la stratégie / l'algorithme correspondant est inclus dans le modèle.

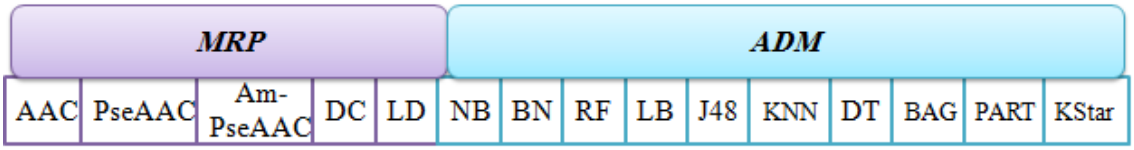


Figure 5. 3: Codage des chromosomes.

5.2.3. Initialisation

Cette étape consiste à créer et à initialiser les individus de la population. L'algorithme génétique étant une méthode d'optimisation stochastique, les gènes des individus sont généralement initialisés aléatoirement. Afin d'illustrer cet opérateur, considérons un modèle prédictif représenté par la figure précédente (figure 5.3). Si nous générons une population de 4 individus, alors nous aurons quatre couples (MRP, AFD) aléatoires. La figure (5.4) montre cette population.

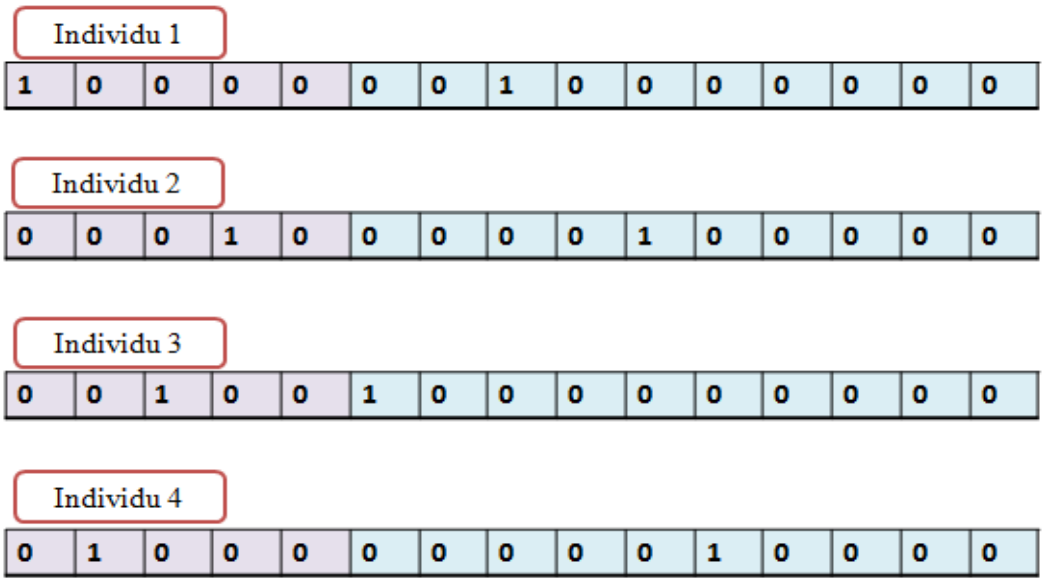


Figure 5. 4: Initialisation de 4 individus.

Dans chaque individu, nous exigeons l'activation d'un seul algorithme et d'une seule méthode à la fois. Tel qu'il est présenté dans la figure précédente, dans le premier individu, la méthode de représentation sélectionnée est AAC et le classificateur utilisé est RF, dans le deuxième individu la stratégie choisie est DC et l'algorithme de classification est J48.

5.2.4. Calcul de la Fonction Fitness

Une fois que nous avons généré et initialisé la population, nous devons attribuer à chaque individu sa fonction fitness. Pour évaluer cette fonction, nous devons construire le modèle prédictif (présenté dans la figure 5.3) avec les données d'apprentissage, puis évaluer son taux d'erreur et sa précision prédictive avec les individus sélectionnés. Evidemment, un taux

d'erreur de classification élevé signifie une faible fitness. Les meilleurs individus auront une plus grande probabilité d'être sélectionnés pour le croisement. La valeur de fitness attribuée à chaque individu sera calculée à l'aide d'Algorithme suivant :

Algorithme 2 : Calcul de la fonction Fitness
Input : MRP = Une stratégie aléatoirement choisie AFD = Un classifieur aléatoirement choisi; **Output :** MinER = Taux d'erreur : Calculé en utilisant l'équation 4.7 MaxAcc = Accuracy : Calculé en utilisant l'équation 4.10 Fit = Fitness : Calculé en utilisant l'équation 5.1 **Begin** Faire une étape de classification en fonction des individus sélectionnés. Calculer ER, ACC en utilisant AFD; Calculez la Fit de chaque individu sélectionné à l'aide de l'équation 5.1. **End.**

Tableau 5. 2: Calcul de la fonction fitness.

Le tableau suivant montre les taux d'erreur de classification (ER) et les précisions (ACC) de chaque individu initialisé précédemment. Notez que la Fonction fitness correspondante à chaque individu est la différence entre les mesures de ACC et ER.

	ER	**ACC**	**Fit**
Individu 1	0.05	0.92	0.87
Individu 2	0.08	0.98	0.9
Individu 3	0.33	0.73	0.4
Individu 4	0.09	0.79	0.7

Tableau 5. 3 Exemple de calcul de la fonction fitness.

5.2.5. Sélection

Une fois le calcul de la fonction fitness est effectué, l'opérateur de sélection choisit les individus qui se recombineront pour la génération suivante. Les individus les plus susceptibles de survivre sont ceux qui sont mieux adaptés à l'environnement. Par conséquent, l'opérateur de sélection sélectionne les individus en fonction de leur niveau de fitness. Le nombre d'individus sélectionnés est (S / 2), tel que S est la taille de la population.

La sélection élitiste permet aux individus les plus aptes de survivre directement pour la prochaine génération. La taille d'élitisme contrôle le nombre d'individus directement sélectionnés. L'une des méthodes de sélection les plus utilisées est la roue de roulette, qui est tournée et les individus sont sélectionnés au hasard. L'individu correspondant est sélectionné pour la recombinaison. La figure suivante illustre le processus de sélection pour notre

exemple. Dans ce cas, l'individu 2 a été sélectionné par élitisme, et l'individu 4 a été sélectionné par roulette. Notez que, bien que l'individu 1 ait plus de fitness que l'individu 4, il n'a pas été sélectionné en raison de la nature stochastique de l'algorithme génétique.

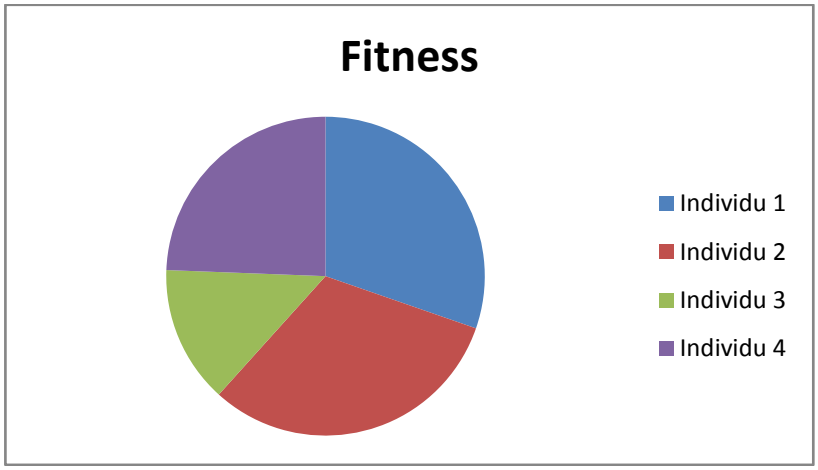


Figure 5. 5: Le processus de sélection.

5.2.6. Croisement

Lors de la reproduction, on choisit pour chaque enfant de un à N parents. Les informations génétiques des différents parents sont mixées avec l'opérateur de croisement.

Le croisement le plus courant consiste à prendre un point de coupure dans le génome. Tous les gènes situés avant ce point viennent du premier parent, et ceux situés après, du deuxième. C'est aussi l'opérateur le plus proche de la réalité biologique. Il est dit discret, car il garde les valeurs telles quelles.

Cet opérateur n'est pas forcément utilisé pour chaque reproduction : il est possible de créer des descendants avec un seul parent, sans avoir besoin de croisement. Il faut donc déterminer le taux de croisement de l'algorithme, généralement supérieur à 50 %. Là encore, c'est l'expérience et le problème qui guideront les choix.

Dans notre cas, après la terminaison de l'opérateur de sélection, l'opérateur de croisement recombine les individus sélectionnés pour générer une nouvelle population. Cet opérateur sélectionne deux individus au hasard et combine leurs caractéristiques pour obtenir quatre descendants pour la nouvelle population, jusqu'à ce que la nouvelle population ait la même taille que l'ancienne.

Nous choisissons la méthode de croisement à un point de coupure, fixée au point 5 de chaque individu. La figure suivante illustre l'étape de croisement pour notre exemple. Ici, nous avons généré deux enfants de deux parents sélectionnés précédemment (individu 2 et individu 4).

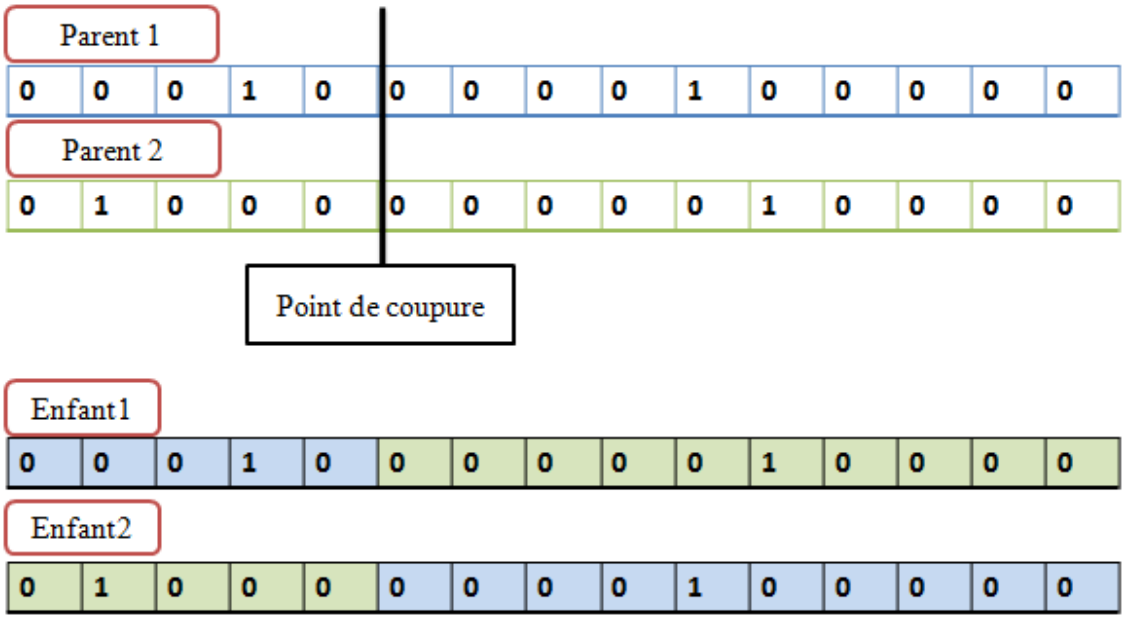


Figure 5.6: Processus du croisement.

5.2.7. Mutation

L'opérateur de croisement peut générer des enfants très similaires aux parents. Cela pourrait provoquer une nouvelle génération avec une faible diversité. L'opérateur de mutation résout ce problème en changeant la valeur de certains gènes chez les enfants au hasard pour garantir la diversification afin d'éviter les optima locaux.

L'opérateur de mutation modifie les caractéristiques d'une solution de manière complètement aléatoire, ce qui nous permet d'introduire et de maintenir la diversité au sein de notre population de solutions. Cet opérateur joue le rôle d'un "élément perturbateur", il introduit du "bruit" au sein de la population. Il consiste donc à choisir aléatoirement certains gènes. La probabilité qu'un gène soit touché par une mutation s'appelle le taux de mutation qui est une très faible probabilité *Pm*, généralement comprise entre 0,001 et 0,01. S'il est trop élevé, les bonnes solutions risquent fort de disparaître. Trop faible, il ne permet pas de trouver de nouvelles solutions rapidement. Il faut donc trouver le bon compromis. Là encore, en fonction de la taille de la population, du nombre de gènes ou du problème, on choisira des taux différents. Un bon départ consiste cependant à partir d'un taux de 5%, et à l'adapter ensuite selon les besoins.

Ce type de mutation permet, surtout si l'espace de recherche est grand, de se déplacer graduellement dans celui-ci. De plus, dans certains cas, les mutations peuvent consister à :

- Ajouter un gène ou en supprimer un, à la fin ou en milieu de chromosome.

- Dupliquer un gène (en particulier dans le cas d'un nombre variable de gènes).
- Échanger la place de deux gènes (C'est le type utilisé dans ce travail).

L'image suivante montre la mutation de l'un des enfants de la nouvelle génération. Comme nous pouvons le voir, les sixième et dixième gènes de l'enfant qui a été muté.

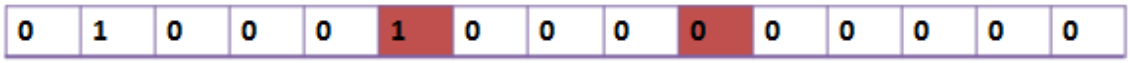


Figure 5. 7: Exemple de mutation de l'enfant 2.

Une étape de survie s'applique alors pour créer la nouvelle génération d'individus.

5.2.8. Survie

Les enfants étant créés, il faut maintenant obtenir une nouvelle génération d'adultes qui peuvent ou non se reproduire. Si la solution la plus simple consiste à remplacer toute la génération des parents par la génération des enfants, il existe cependant plusieurs autres stratégies de survie.

On choisit un simple remplacement : à chaque génération, tous les enfants deviennent les adultes, qui, eux, disparaissent.

On obtient donc à la fin de la survie une nouvelle population, et on peut reboucler tout le processus.

5.3. Etude Expérimentale

5.3.1. Base de données

Il est important de choisir une base de données appropriée, car l'utilisation d'une base de données adéquate, volumineuse et complète est nécessaire à l'évaluation et à la comparaison des performances des algorithmes et des méthodes de représentation du système proposés. Les séquences RCPGs obtenues sont sous forme de séquences primaires d'AA qui varient en longueur et en contenu, d'une séquence à une autre (de 250 à 1200 acides aminés) et qui se composent de différents alphabets représentant les 20 acides aminés natifs.

Il est donc important d'effectuer une étape de prétraitement pour pouvoir exploiter nos séquences protéiques. L'étape de prétraitement est composée de deux parties, la première consiste en la transformation du fichier contenant les séquences en une base de données afin de faciliter la manipulation et l'accès par notre système. La seconde partie consiste en une normalisation des données de sorte de pouvoir les sauvegarder dans un fichier .arff. Ce dernier se décompose de deux parties : la première contient les noms d'attributs et la

deuxième contient les séquences, leur identifiants, leur noms, leur famille, leur sous famille et sous sous-famille dans le but d'adapter nos données au système proposé et de les rendre exploitables par l'environnement Weka.

Nous avons discuté dans le chapitre précédent de la base de données GPCRdb, utilisée pour les expérimentations. Dans cette partie des expérimentations, nous avons gardé la même BDD GPCRdb, puisqu'elle contient un grand nombre de séquences.

5.3.2. Prétraitement de la BDD

La base de données utilisée dans notre étude a plus de 35000 chaînes d'acides aminés de quatre classes différentes. Nous avons effectué un prétraitement sur l'ensemble des séquences pour améliorer la qualité des informations disponibles, comme le montre la figure suivante (5.8).

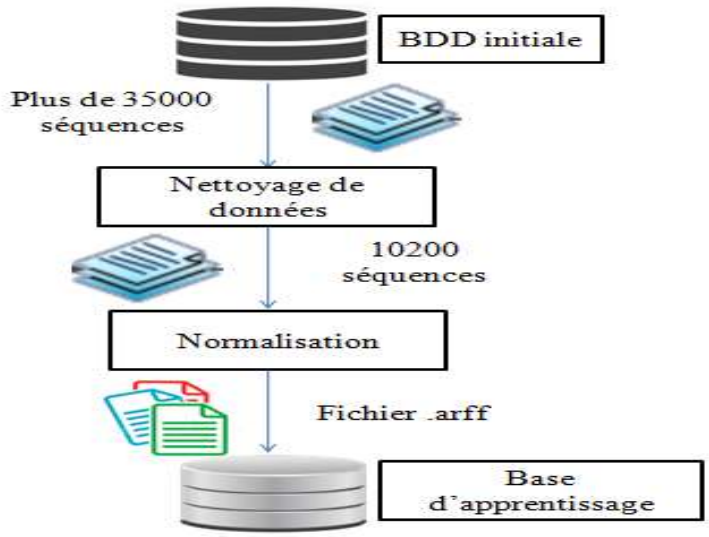


Figure 5. 8: Prétraitement sur la BDD.

Au départ, nous avons analysé les séquences et vérifié les redondances des données et les séquences orphelines, en les supprimant et en réduisant la base de données à un total de 10200 chaines protéiques. Une analyse préliminaire a montré le coût de calcul élevé de travailler avec la BDD complète. De plus les séquences ayant une classe inconnue ne sont pas utiles dans notre système qui a effectué une classification supervisée nécessitant la connaissance de la classe de chaque séquence à priori.

Après avoir obtenu les vecteurs d'attributs de chaque stratégie de représentation de protéine utilisée dans ce travail, nous avons construit un fichier .arff pour sauvegarder les vecteurs numériques de toutes les séquences ainsi que les attributs utilisés et la classe, sous classe et sous sous-classe de chacune. Cette étape de normalisation nous produit cinq bases

d'apprentissages, une pour chaque MRP pour effectuer la classification dans l'environnement Weka.

5.3.3. Outils de l'expérimentation

Comme nous l'avons détaillé dans le chapitre précédent, les outils utilisés pour effectuer les expérimentations sont :

- ✓ **L'environnement Weka :** pour effectuer la classification en utilisant différents algorithmes de fouille de données qui sont : NB, BN, KNN, RF, BAG, LB, PART, KStar, DT, J48.
- ✓ **Protr :** pour obtenir une représentation numérique des séquences protéiques par les méthodes : AAC, PseAAC, Am-PseAAC, DC. Rappelons que nous avons implémenté la méthode LD (expliquée dans le chapitre précédent) pour produire un vecteur d'attributs de dimension 210D.
- ✓ **Mesure de performance :** Pour mesurer la performance de notre système proposé, nous avons focalisé sur l'accuracy et le taux d'erreur de la classification. Ces mesures sont utilisées pour trouver la valeur de la fonction de fitness de chaque solution.

Le tableau suivant récapitule toutes les mesures utilisées dans ce travail ainsi leurs formules de calcul:

Abb	Description	Equation	N°
ACC	Accuracy	$Eq.4.7$	
ER	Error Rate	Eq. 4.10	
Fit	Fonction Fitness	$Fitness = \text{Acc} - ER$	5.1

Tableau 5. 4: Mesures de performance.

5.3.4. Paramètres de l'AG

Nous avons effectué des tests préliminaires pour définir quels intervalles de paramètres seraient les plus adéquats pour nos expériences, et définir un critère d'arrêt basé sur le nombre de générations, afin d'avoir un contrôle total de nos expériences concernant l'effort de calcul requis.

Le tableau suivant récapitule et illustre les paramètres de l'AG et leurs valeurs.

Description	Nature / Valeur
Initialisation de la population	Aléatoire
Représentation	Binaire
Opérateur de croisement	Point unique
Probabilité de croisement *« Pc »*	70%

Opérateur de mutation	Point unique
Probabilité de mutation « *Pm* »	0.01 (1%)
Taille de la population	75 individus
Taille du chromosome	15 gènes
Nombre de génération	50
Méthode de sélection	Sélection par roulette + sélection élitiste
Phase de survie	Remplacement simple
Critère d'arrêt	Nombre de génération

Tableau 5. 5: Paramètres des expérimentations.

5.3.5. Recherche de la meilleure solution

Pour rechercher la meilleure solution, nous avons effectué un total de 20 expérimentations représentant les différentes combinaisons des plages de paramètres indiquées dans le tableau précédent. La figure (5.9) montre le processus expérimental exécuté pour choisir la meilleure solution.

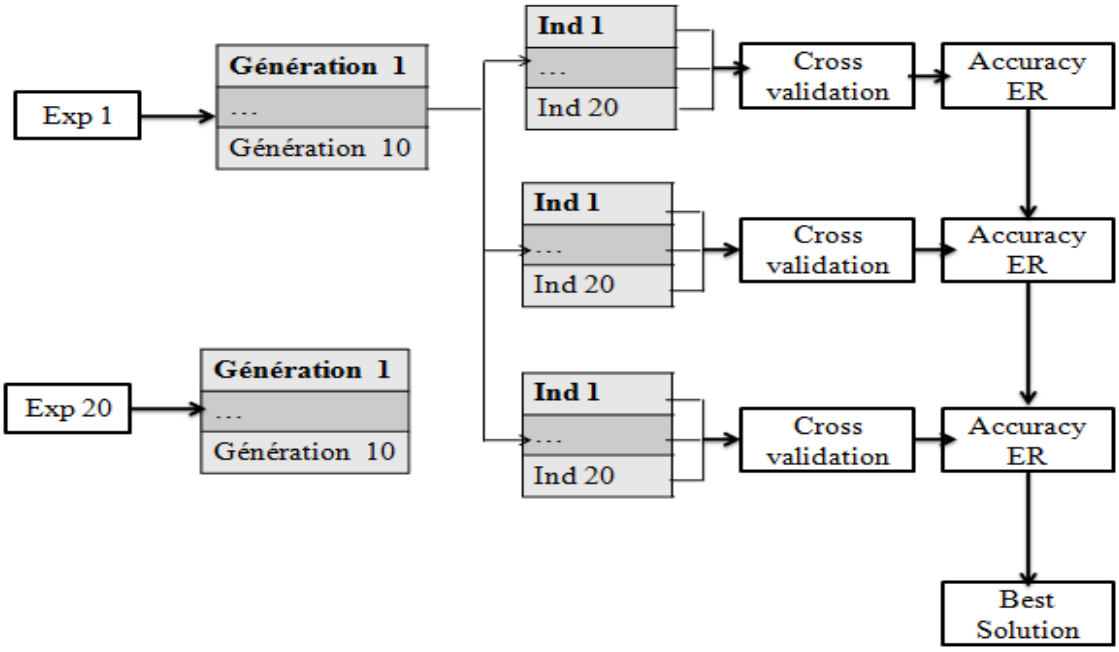


Figure 5. 9: Critères pour choisir la meilleure solution de l'AG.

Après l'exécution de l'AG, nous avons obtenu un ensemble d'expériences, chacune avec 10 générations, pour une validation statistique. Pour chacune de ces générations, il existe un ensemble de 20 candidats possibles non dominés (individus sélectionnés aléatoirement). Avec ces solutions candidates, nous avons effectué un test de validation croisée par 10 dossiers de chacune d'elles et obtenu les valeurs d'ER et de l'accuracy.

L'expérience qui a obtenu les meilleurs résultats était l'ensemble de paramètres choisi (taille de la population, nombre de générations, Pc et Pm).

Ensuite, après avoir choisi le meilleur ensemble de paramètres de notre solution grâce à des tests approfondis, nous avons également plusieurs solutions candidates possibles dans cet ensemble de paramètres, qui correspondent aux meilleures solutions trouvées.

5.3.6. Discussion

Afin d'optimiser les performances de classification, nous avons effectué une validation statistique de la base de données utilisée, en appliquant un processus de validation croisée par 10 dossiers. Cela signifie que nous avons divisé l'ensemble de données de chaque solution candidate en 10 sous-ensembles.

La figure suivante montre la progression de la variation de la valeur de fitness pour chaque algorithme. Tous les résultats sont produits avec un nombre d'itérations maximales égal à 50, la probabilité de croisement = 0,7 et la probabilité de mutation = 0,01.

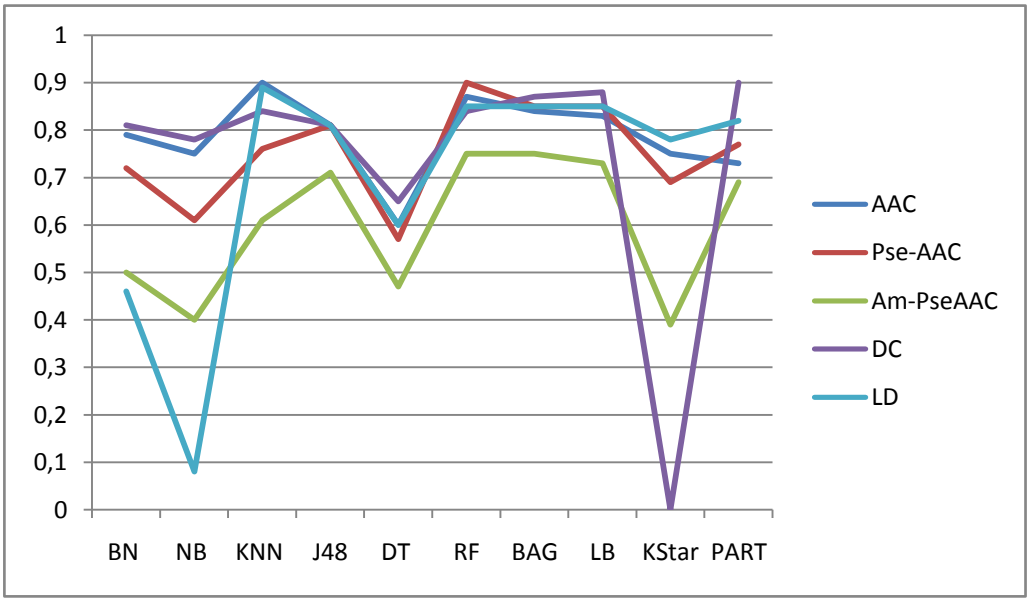


Figure 5. 10: Progression des valeurs de fitness selon les MRP et les AFD.

Les valeurs de la fonction fitness de chaque classificateur en utilisant les cinq stratégies de représentation physico-chimique des protéines se calculent en fonction des mesures de l'accuracy et du taux d'erreur. Ces valeurs sont incluses dans l'intervalle [0, 1] tel que la meilleure valeur obtenue est égale à 0.9. Statistiquement, ceux ayant les fitness les plus élevées auront le plus d'enfants, mais tous ont au moins une chance de se reproduire, même si elle reste faible, tel que l'individu: 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0.

Les performances de l'AG varient considérablement selon la sélection des stratégies et des classificateurs. Ceci est particulièrement important en ce qui concerne l'amélioration de la qualité de la classification en termes de précision et de taux d'erreur.

Analysons d'abord les résultats de la figure précédente, les couples (LD, NB), (LD, BN), (DC, KStar), (Am-PseAAC, NB), (Am-PseAAC, BN), (Am-PseAAC, DT), (Am-PseAAC, KStar) représentent les plus mauvais individus à cause de leur valeur de fitness qui ne dépasse pas 0.5.

L'impact négatif de l'utilisation d'un mauvais classificateur avec une bonne représentation (par exemple DC avec KStar et LD avec NB ont une valeur de fitness de 0 / 0,08 respectivement) semble être plus grand que l'impact de l'utilisation d'une mauvaise représentation même avec un bon classificateur (par exemple Am-PseAAC avec DT et KNN ont des valeurs de fitness de 0,47 / 0,61). Il est intéressant de noter que la plupart des articles sur la prédiction de la fonction des protéines sont plus préoccupés par l'essai de différents algorithmes de classification que de différentes stratégies d'extraction de caractéristiques et leur impact sur la précision prédictive.

Le système proposé nous fournit facilement le meilleur algorithme à utiliser avec la meilleure représentation pour classer les séquences des RCPG comme illustré dans la figure ci-dessous (Figure 5.11).

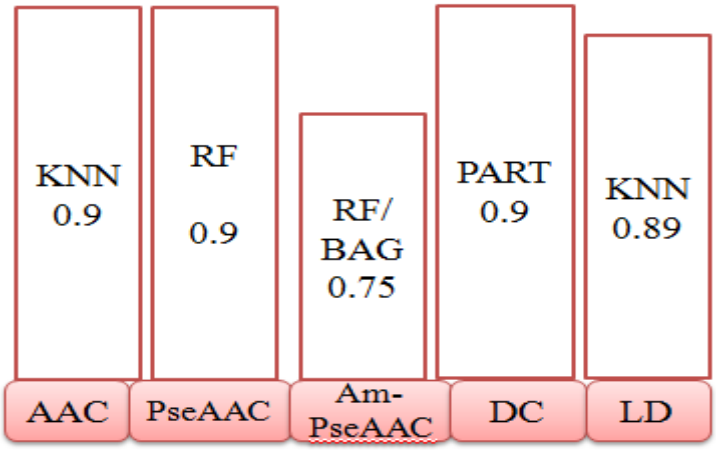


Figure 5. 11: La meilleure fitness de chaque représentation.

En se basant sur les valeurs de fitness de l'algorithme génétique, malgré la nature étroite (très proche) de mesures, nous avons remarqué que les meilleurs couples (MRP, AFD) obtenus pour la base de données utilisée sont (AAC, KNN), (PseAAC, RF) et (DC, PART).

5.4. Conclusion

Dans ce travail, nous avons présenté une proposition d'un système pour la sélection du meilleur couple (MRP, AFD), en analysant l'impact de différentes représentations de protéines et de différents types d'algorithmes de classification pour la tâche de la prédiction de la fonction des RCPG.

Nous avons utilisé cinq stratégies de représentation protéique qui sont des représentations indépendantes de l'alignement calculées à partir de la séquence protéique: AAC, PseAAC et Am-PseAAC, DC et LD. Pour effectuer la classification, nous avons utilisé dix algorithmes de fouille de données: KNN, BN, NB, RF, BAG, PART, KStar, J48, DT et LB et ceux pour sélectionner la meilleure correspondance de représentation et de classifieurs. La fonction fitness est calculée à partir des mesures de l'accuracy et du taux d'erreur, et elle converge vers la solution optimale en un nombre réduit d'itérations.

Nous avons réalisé les expérimentations sur la base de données GPCRdb, qui contient plus de 35000 séquences. Nos résultats expérimentaux montrent qu'en général, quel que soit le type de protéine: AAC est un très bon descripteur avec KNN car il est simple et fournit de meilleurs résultats que les autres algorithmes. RF a fonctionné nettement mieux que les classificateurs restants avec les méthodes PseAAC et Am-PsAAC. Le DC fournit de très bons résultats (sauf pour le KStar dans lequel les résultats sont très mauvais et sont égaux à 0). En ce qui concerne la stratégie LD, l'algorithme NB produit une fitness très mauvaise à cause de l'augmentation de son taux d'erreur.

Compte tenu des résultats spécifiques à la classification des RCPG au niveau sous sous-famille, notre recommandation basée sur l'AG représente les couples (AAC, KNN), (PseAAC, RF), (Am-PseAAC, RF), (Am-PseAAC, BAG), (DC, PART) et (LD, KNN) qui sont les meilleures solutions fournit par notre système.

Les recherches futures incluraient la réalisation des expérimentations additionnelles avec d'autres types de méthodes de représentations de protéines, un nombre plus grand d’algorithmes de classification et avec d'autres types de protéines (tel que : les Enzymes, les facteurs de transcription). Une autre direction pour les recherches futures consiste à effectuer des expérimentations plus contrôlées pour voir si le nombre de caractéristiques a une influence significative sur l'efficacité d'un type particulier de caractéristique: par exemple la représentation par AAC a un très petit nombre d'attributs contrairement à la représentation par LD ou DC qui produisent des vecteurs de plus grande taille.

Chapitre 6
Contribution 3 : Sélection de caractéristiques par l'algorithme Bat pour la prédiction de la fonction des RCPG

6.1 Introduction

Il existe plusieurs méthodes de représentation de protéines, certaines parmi elles se calculent à partir de la séquence protéique, et d'autres se basent sur les motifs biologiques. Toutes ces approches produisent des vecteurs numériques différents en terme de taille et de contenu.
Par exemple, dans notre travail, nous avons utilisé cinq stratégies qui sont: AAC qui fournit un vecteur de 20 attributs, PseAAC résulte un vecteur de 50 attributs, Am-PseAAC donne un vecteur d'attributs de 80D, LD produit un vecteur de 210 attributs et finalement DC qui donne un vecteur de 400 éléments. Quelque soit la taille des vecteurs produits, ils peuvent contenir de bruit (informations inutiles, redondances, ambigüités...etc). Ce dernier influence certainement la qualité de la classification. Est-il nécessaire d'éliminer le bruit pour optimiser la classification et améliorer les performances?, et quelles sont les moyens pertinents pour effectuer cette opération?
La taille du vecteur affecte directement les résultats de la classification et le temps d'exécution, surtout pour des ensembles de données complexes avec des centaines et des milliers d'attributs (tel que les protéines), qui peuvent causer la malédiction de la dimensionnalité et des problèmes d'explosion combinatoire. En outre, certains des algorithmes de classification et de clustering ne peuvent pas fonctionner correctement.
L'une des techniques les plus réalisables pour faire face à ce problème est la sélection de caractéristiques qui permet d'optimiser la classification et améliorer les mesures de performance. Les méthodes de sélection de caractéristiques peuvent être divisées en deux approches: Les méthodes Filter et Wrapper en fonction de leur dépendance ou de leur indépendance vis-à-vis de l'algorithme d'induction (Section 3.3), mais malgré leur facilité et leur simplicité dues à des implémentations basées sur des fonctions mathématiques, ces méthodes présentent un inconvénient majeur qui est le faux optimum donné. Pour cette raison, nous avons décidé d'utiliser l'algorithme bio-inspiré chauves-souris qui est une méta-heuristique récente proposée en 2010 par Yang. Malgré ses résultats prometteurs il n'est jamais utilisé, ni dans le domaine de la bioinformatique en général et, ni pour la classification des RCPG en particulier. Il a été largement utilisé pour résoudre les tâches d'optimisation d'ingénierie et les tâches d'optimisation des contraintes.
Nos principales contributions résident dans l'utilisation de l'algorithme de chauves-souris (BAT) qui n'avait jamais été utilisé dans ce domaine malgré ses points forts et ses capacités de trouver les solutions optimales, ainsi leur efficacité pour la résolution des différents problèmes.

Nous organisons ce chapitre comme suit: La section 2 explique le système proposé pour la sélection des caractéristiques ainsi que toutes les étapes de réalisation. Dans la section 3, des expérimentations ont été réalisées pour évaluer l'algorithme. Nous faisons également des comparaisons avec des méthodes supplémentaires et analysons les résultats des expériences. Enfin, la conclusion et les travaux futurs sont présentés dans la section 4.

6.2. Un framework bio-inspiré pour la classification des RCPG

La sélection des caractéristiques fait référence à la recherche de méthodes dont les dimensions seront réduites, et les attributs choisis présentent les données originales.

Au cours de la dernière décennie, l'application des techniques de FS en bioinformatique est devenue une véritable condition préalable à la construction de modèles de classification (Figure 6.1); Saeys et al. [SAE 07] ont passé en revue les techniques de FS en bioinformatique en fournissant leur taxonomie de base. Dans [KAV 17], les auteurs ont utilisé la sélection de caractéristiques pour découvrir des bio-marqueurs pour la prédiction du diabète. Dans ce contexte Bagherzadeh et al. [BAG 16] ont comparé plusieurs algorithmes de sélection de caractéristiques communes pour prédire le diabète sucré.

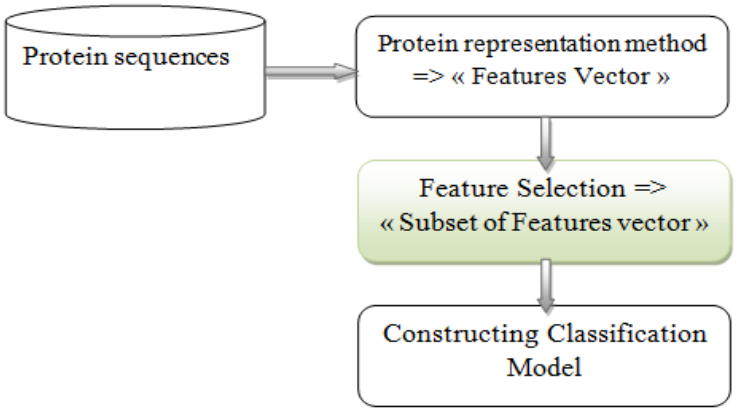


Figure 6. 1:Modèle basé sur la sélection de caractéristiques pour la classification de RCPG.

Pour garantir la bonne performance, nous choisissons d'adapter l'algorithme de chauves-souris pour résoudre le problème de sélection de caractéristiques dans la prédiction de la fonction des RCPG. Les principales étapes de notre architecture générale du système proposé sont illustrées dans la Figure 6.2.

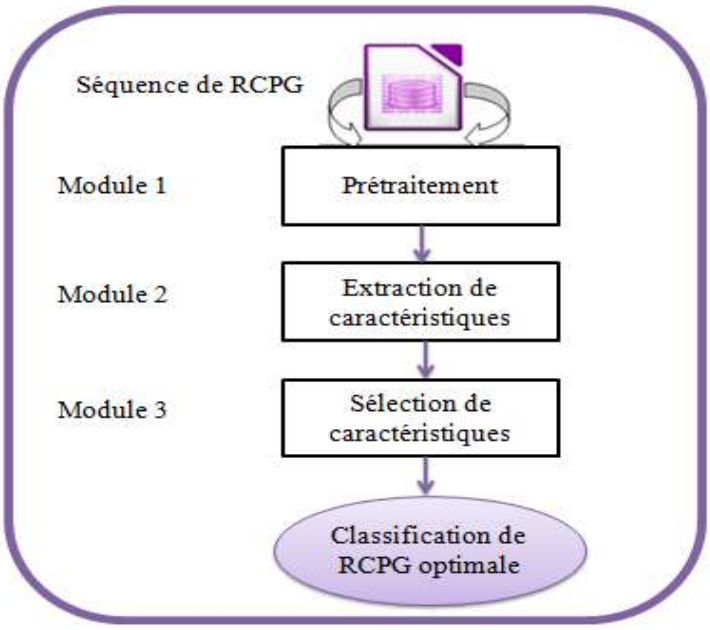


Figure 6. 2: Etapes du système proposé.

En commençant tout d'abord par les données d'entrée, c'est-à-dire les séquences protéiques stockées dans le fichier source et avant de faire tout traitement sur ces séquences, nous avons affaire à une étape de prétraitement très nécessaire pour standardiser et structurer le format des données.

L'étape d'extraction de caractéristiques consiste à transformer les vecteurs alphabétiques en des vecteurs numériques de chaque séquence protéique. Il existe plusieurs méthodes de représentation des protéines pour effectuer cette transformation, le choix d'une stratégie d'extraction de caractéristiques a une forte influence sur les résultats de la classification en termes de précision et de taux d'erreur, c'est pourquoi nous devons utiliser les stratégies d'extraction de caractéristiques les plus couramment utilisées dans la littérature pour effectuer une étude analytique sur les résultats obtenus. Comme le choix d'une MRP peut tendre vers une méthode produisant un vecteur de taille énorme, la troisième étape de notre système sera primordiale et inévitable pour éviter l'explosion combinatoire des algorithmes de classification, en réduisant le temps et l'espace mémoire.

Dans cette section, nous avons présenté l'architecture d'un système basé sur l'algorithme BAT pour résoudre le problème de la sélection des attributs en raison de la simplicité de sa mise en œuvre et de sa rigidité aux données bruyantes. Nous expliquerons et présenterons les détails de ces étapes les prochaines sections.

6.2.1. Module 1: Prétraitement de données

Cette partie du système est chargée d'automatiser la transformation des données stockées aléatoirement dans un fichier source en données structurées enregistrées dans une base de données (Figure 6.3), où chaque séquence a ses informations nécessaires: famille, sous-famille, sous sous-famille également son type, son identifiant et sa description.

Cette étape est d'abord réalisée pour faciliter les traitements ultérieurs, en utilisant une représentation performante qui peut être exploitable par notre système.

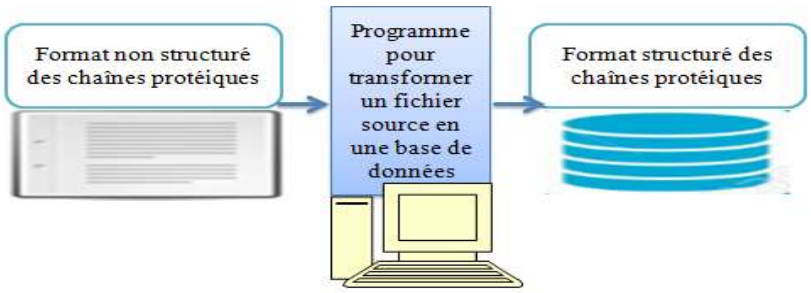


Figure 6. 3: Prétraitement de données.

6.2.2. *Module 2 : Extraction de caractéristiques*

Ce module permet de transformer les chaînes alphabétiques de séquences protéiques en vecteurs numériques. Ces vecteurs sont les entrées pour construire le modèle de classification qui va évaluer les attributs sélectionnés dans la troisième étape de notre système.

D'après la figure 6.4, nous pouvons distinguer trois traitements nécessaires expliqués plus tard.

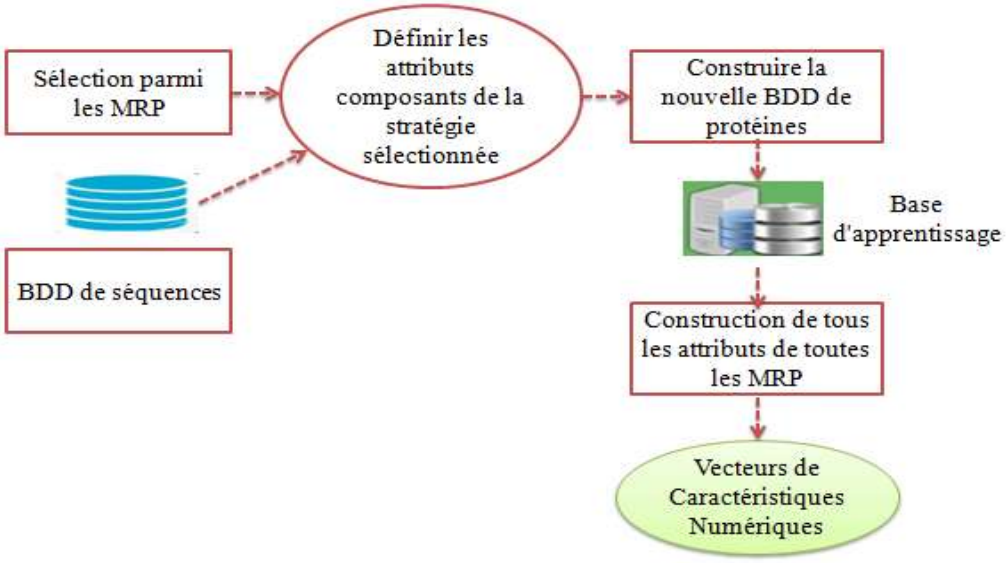


Figure 6. 4: Les étapes requises pour l'extraction de caractéristiques.

6.2.3. *Sélection parmi MRP*

Plusieurs Méthodes de Représentation des Protéines existent dans la littérature, chacune a ses avantages et ses lacunes, le choix d'une méthode ou d'une autre n'est pas évident car il dépend de divers facteurs comme le nombre de caractéristiques produites, et les informations existantes dans les vecteurs numériques (Ordre des acides aminés, longueur de la séquence…).

Dans ce travail, nous avons fait une étude analytique des méthodes de représentation des protéines, pour observer la variation des résultats de la classification à l'aide de différents algorithmes d'apprentissage automatique. Dans les chapitres 4 et 5 nous avons décrit chaque stratégie utilisée ainsi que la nécessité de sélectionner la méthode pertinente pour un résultat optimal.

Dans ce chapitre, nous utilisons les mêmes MRP implémentées précédemment pour constater l'impact de la sélection des caractéristiques sur le résultat de la classification.

Dans cette étape, une stratégie d'extraction de caractéristiques doit être choisie parmi les MRP utilisées pour la représentation numérique des séquences des GPCR.

6.2.4. *Construction de la nouvelle base de données de protéines*

Après avoir sélectionné une MRP, nous obtenons un vecteur de composants numériques; la taille de ce vecteur est relative à la méthode choisie. À partir d'une base de données de séquences alphabétiques des acides aminés, nous allons construire une nouvelle base de données d'apprentissage qui ne contient que les vecteurs d'attributs numériques.

Cette base se présente sous la forme d'un fichier .arff qui commence par une déclaration de tous les attributs puis la représentation numérique de chaque séquence protéique, ainsi que sa famille, sous-famille et sous-sous-famille. Ce fichier sera l'entrée de notre système pour effectuer l'étape de la classification à l'aide de l'environnement Weka.

6.2.5. *Construction de tous les attributs des MRP*

Dans notre travail, nous utilisons cinq MRP. Pour accomplir notre tâche préliminaire qui est l'analyse comparative des stratégies d'extraction de caractéristiques pour la classification des RCPG, nous devons construire pour chaque méthode de représentation protéique, la base d'apprentissage correspondante sous la forme d'un fichier .arff cité précédemment. Par conséquent, le module d'extraction de caractéristiques résulte cinq bases d'apprentissage.

La taille du vecteur de caractéristiques obtenu, varie d'une stratégie à une autre et peut atteindre 400D pour la stratégie DC, ce qui rend la tâche de la classification plus difficile. Un sous-ensemble de caractéristiques approprié peut apporter de nombreux avantages tels que l'amélioration de la précision des prédictions, éviter le sur-apprentissage, distinguer les attributs clés des attributs sans importance et fournir une compréhension concise des données.

Dans la section suivante, nous détaillerons l'étape de la sélection des attributs à l'aide de la méta-heuristique bio-inspirée: l'algorithme de chauve-souris (Bat).

6.2.6. *Sélection de caractéristiques (FS)*

La dernière étape de notre système consiste à rechercher les attributs pertinents pour chacune des MRP utilisées, en éliminant ceux qui sont inutiles. Un algorithme bio inspiré pour la sélection de caractéristiques est proposé. Il permet de générer un sous-ensemble optimal d'attributs à partir d'un vecteur numérique d'entrée et il produira un vecteur de plus petite taille avec les meilleurs attributs en sortie.

Le schéma existant dans la figure 6.5 récapitule les étapes nécessaires pour compléter ce processus. Nous expliquerons toutes ces étapes ultérieurement.

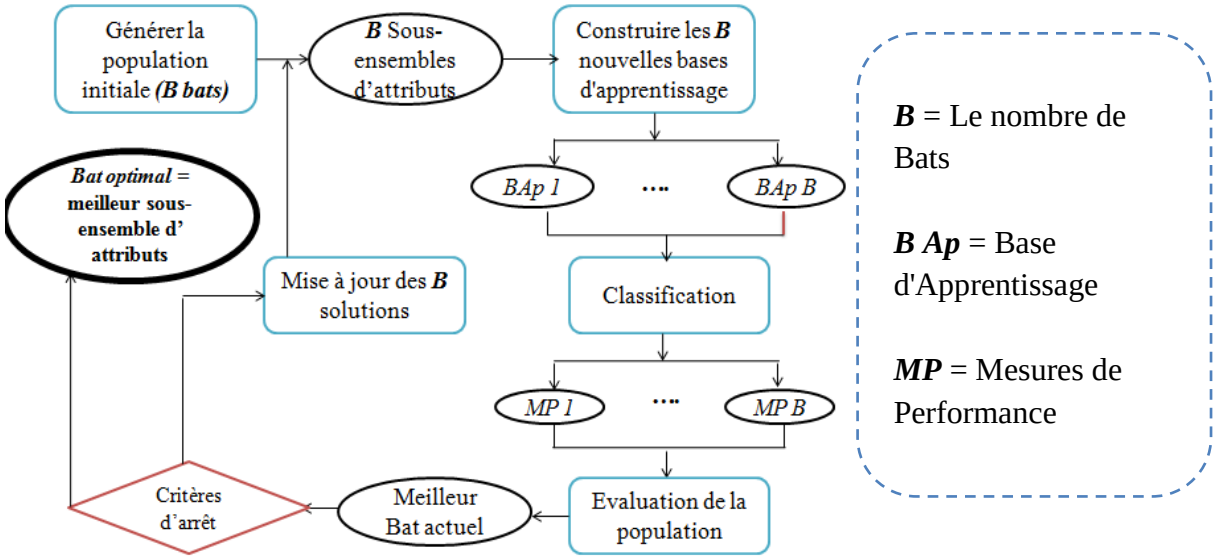


Figure 6. 5: Processus de sélection de caractéristiques.

- **Créer la population initiale:** Cette phase d'initialisation sert à créer une population initiale de *B* Bats, où chaque individu représente une solution au problème qui sera initialisée aléatoirement par des vecteurs *(VB)* de valeurs binaires où certains attributs sont activés et les autres sont inactivés.

Par exemple: *V* est un vecteur de 20 attributs. *VB* est la représentation des chauves-souris sous forme binaire, dans cette étape toutes les séquences des RCPG seront représentées sous forme binaire qui fournit des bats contenant les sous-ensembles d'attributs.

V	*0.3*	*0,6*	*0.3*	*0.2*	*0.5*	*0.18*	*0.2*	*0.1*	*0.4*	*0.3*	*0.2*	*0.1*	*0*	*0.7*	*0.8*	*0.9*	*0.7*	*0.3*	*0.1*	*0*
VB	*0*	*1*	*0*	*0*	*1*	*1*	*0*	*1*	*1*	*1*	*0*	*0*	*1*	*0*	*1*	*1*	*0*	*1*	*0*	*1*

- **Construire les *B* nouvelles bases d'apprentissages *(B Ap)*:** Avant la classification, il est nécessaire de transformer la base de données en une base d'apprentissage avec une représentation binaire pour chaque séquence protéique; chaque attribut ayant la valeur «1» sera pris en considération.

Le nombre de vecteurs dans chaque base d'apprentissage *(B Ap)* est égal au nombre de séquences des RCPG dans la BDD. Dans notre cas, concernant la méthode LD, nous avons des vecteurs d'attributs de 210D, et une base de données qui contient 10200 séquences des RCPG, la figure 6.6 présente la construction de la population ayant des chauves-souris *B*.

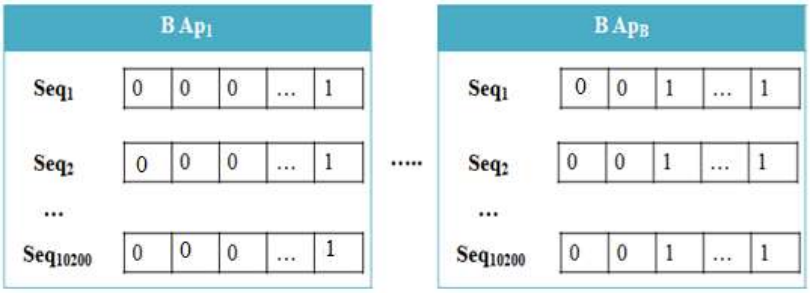


Figure 6. 6: Initialisation aléatoire des B Bats.

Après avoir représenté les chauves-souris et les bases d'apprentissage, nous devons initialiser les paramètres correspondant à chacun selon les formules suivantes:

Frequency: $f_i = f_{min} + (f_{max}-f_{min})\,\beta,$ (6.1)

Vélocité: $v_i^t = v_i^{t-1} + (x_i^t - x_*)\,fi,$ (6.2)

Position: $x_i^t = x_i^{t-1} - v_i^t$ (6.3)

Tel que:

- *fi* est la fréquence d'émission des impulsions de la chauve-souris, et appartient à la gamme $[f_{min}, f_{max}]$ correspondant à la gamme de longueurs d'onde $[\lambda_{min}, \lambda_{max}]$. Par exemple, la gamme de fréquences *[20 KHz, 500 KHz]* correspond à la gamme de longueurs d'onde *[0,7 mm, 17 mm]* vers les plus petites distances. Afin de simplifier la mise en œuvre, on a supposé que $f \in [0, f_{max}]$ sachant que les hautes fréquences

correspondent à des longueurs d'onde courtes. Pour les chauves-souris, les portées typiques sont de quelques mètres. Par conséquent, le taux d'émission d'impulsions peut être dans la plage *[0, 1]* où 0 signifie qu'il n'y a pas de pulsation, et 1 signifie le taux maximum d'émission d'impulsions.

- $\beta \epsilon [0, 1]$ est un vecteur aléatoire extrait d'une distribution uniforme.
- x^* est la meilleure position globale (solution), qui sera calculée en comparant toutes les solutions obtenues par chaque chauve-souris.
- Pour la recherche locale, une fois qu'une solution est sélectionnée parmi les meilleures solutions actuelles, une nouvelle solution pour chaque chauve-souris est générée localement en utilisant le chemin aléatoire.

- **Classification:** Pour évaluer la pertinence des attributs sélectionnés, nous avons effectué une étape de classification pour tester sa précision prédictive et son taux d'erreur. Autant le taux d'erreur est minimal et la précision est maximale, les attributs choisis sont optimaux.

  Dans notre travail, la classification est effectuée à l'aide de dix algorithmes de fouille de données: NB, BN, KNN, J48, RF, BAG, LB, DT, ZR implémentés dans le package Weka. Pour toutes les chauves-souris de chaque itération, nous effectuons des calculs des mesures de performances afin de pouvoir les comparer et en déduire la meilleure chauve-souris (meilleure solution globale).

- **Évaluation de la population:** Pour vérifier la crédibilité de la solution actuelle, il est nécessaire de faire une évaluation locale pour extraire la meilleure solution dans l'itération actuelle et une évaluation globale pour obtenir la meilleure solution fournie à partir de toutes les itérations, ce qui se fait en utilisant la fonction objectif ou fitness basée sur l'étape de la classification. Cette fitness est calculée à l'aide de l'algorithme 1.

Algorithme 1 : La Fonction Fitness

Entrées:

Position = Un Vecteur de caractéristiques;

Algo = Un classifieur choisi;

Sorties:

MinER **=** Error Rate : Calculated using Eq. (4.10)

MaxAcc = Accuracy : Calculated using Eq.(4.7)

MaxMCC = Calculated using Eq. (4.6)

MaxFm = Calculated using Eq. (4.5)

Begin

Générer curFile un fichier .arff, en utilisant les attributs activés;

Calculer *ER, ACC, MCC et Fm* en utilisant Algo sur curFile;

End.

Tableau 6. 1: Calcul de la fonction fitness.

- **Mise à jour des Bat solutions:** Cette étape est la plus cruciale; en effet, le mouvement des chauves-souris est responsable à l'efficacité de l'algorithme. Les équations : Eq. (4.5), Eq. (4.6) et Eq. (4.7) précités définissent la nouvelle solution et mettent à jour la position et la vitesse de chaque chauve-souris dans un espace de grande dimension. Nous adaptons l'algorithme de chauve-souris (Bat) pour la sélection des attributs dans un processus en deux phases. La première phase, appelée ***initialisation***, est utilisée pour construire une solution initiale afin de lancer la recherche. De plus, nous ajustons les paramètres de cet algorithme afin d'obtenir une solution initiale pas trop mauvaise. Dans la deuxième phase, que nous appelons phase d'***amélioration***, nous introduirons la recherche locale afin d'améliorer la qualité de la solution retournée lors de l'initialisation comme mentionné dans l'Algorithme 2 (Tableau 6.2).

Algorithme 2 : Algorithme de chauve souris adapté à la sélection de caractéristiques

Entrées:

nb = Nombre de Bat (Taille de la population).
na = Nombre de caractéristiques sélectionnées
MaxIter= Itération maximale
N = Taille initiale du vecteur
Algo = Choix d'un algorithme de classification.

Output:

S = Sous-ensemble d'attributs sélectionnés *(A1.....A_{na})* ayant la meilleure fonction fitness

Begin

Initialisation

Initialiser le nombre des chauves-souris au ***nb***;
Initialiser le nombre d'attributs à ***na***;
Génération de chauves-souris;
Initialiser aléatoirement la variable ***FGlob*** de fitness global;
For chaque chauve-souris **do**
Déclarer un vecteur ***vect*** de ***N*** attributs initialisé avec la valeur false;
Activer au hasard un sous-ensemble d'attributs de ***vect*** de taille ***na*** en définissant leurs valeurs à true;
Position actuelle = *vect;*
Initialiser aléatoirement la fréquence, la vitesse *(f, V)* en utilisant Eq. (6.1), Eq. (6.2) respectivement;
Initialiser ***r*** = 1, ***A*** = 1;
Calculer la fitness locale ***FLoc*** "en utilisant l'Algorithme 1(Tableau 6.1)";
If FLoc > FGlob then
Mettre à Jour ***FGlob****;*
End If; End For;

Amélioration

Repeat;

For chaque chauve-souris ***do***
Mettre à jour la vitesse de la chauve-souris;
f = Calculer ***FLoc*** de la chauve-Souris "en utilisant l'Algorithme 1(Tableau 6.1)";

If f > FLoc then
Mettre à jour ***FLoc****;*
Mettre à jour la position locale du ***vect*** par Eq. (6.3);
End If;

If f > FGlob then
Mettre à jour ***FGlob****;*
Mettre à jour la position globale;
Mettre à jour ***r****;*
End If; End For;

Until (Max iteration)
End.

Tableau 6. 2: Algorithme Bat pour la sélection des attributs.

6.3. Résultats expérimentaux

Pour prouver l'efficacité de notre approche, nous avons effectué plusieurs expériences de l'algorithme proposé pour la sélection des caractéristiques. Cette section décrit les outils et les méthodes utilisés. Après cela, nous présentons une étude analytique sur les méthodes de représentation des protéines en utilisant plusieurs classifieurs pour la sélection des caractéristiques par l'algorithme de chauve-souris. Une étude comparative avec PSO-Search et EA-Search, qui sont les deux algorithmes bio-inspirés les plus utilisés pour résoudre ce problème, est également effectuée.

6.3.1. *Outils et méthodes*

Les tests ont eu lieu sur un ordinateur portable équipé d'un processeur Intel (R) Core (TM) i5 fonctionnant à 2,3 Ghz et 4 Go de RAM. Les programmes ont été développés en utilisant l'environnement eclipse neon avec jdk 1.8.

- ***Base de données:*** La base de données que nous avons principalement utilisée pour la formation et l'évaluation de notre approche de classification a été téléchargée à partir du site Web [1] mentionné dans le chapitre 4.
- ***Weka:*** Weka (Waikato Environment for Knowledge Analysis) Nous avons utilisé weka pour deux raisons: Premièrement, pour faire l'évaluation de notre système proposé de FS en effectuant à chaque itération une étape de classification du RCPG via neuf AFD qui sont: NB, BN, IBK, J48, RF, LB, BAG, DT et ZR et deuxièmement, nous avons effectué une étape de sélection d'attributs en utilisant EA et PSOSearch afin de les comparer avec l'algorithme que nous avons proposé et implémenté.
- ***Mesures de performance:*** La métrique standard utilisée dans la classification standard pour mesurer et évaluer les algorithmes de fouille de données est le test jackknife. Les mesures de performance utilisées pour l'évaluation des classifieurs sont: la précision (ACC), le coefficient de corrélation de Matthews (MCC) et F-measure (Fm). Une mesure supplémentaire est utilisée dans notre travail qui est le taux d'erreur de classification pour contrôler sa variation avec et sans sélection d'attributs.

6.3.2. *Résultats et discussion*

Dans cette section, nous évaluons l'efficacité de notre approche avec plusieurs expérimentations sur la base de données décrite dans la section précédente. Nous analysons le comportement de notre méthode en raison de la stochastique des algorithmes évolutionnaires. Ici, nous visons à déterminer le meilleur sous-ensemble de caractéristiques et à étudier le changement des mesures utilisées dans la fonction objective.

Le tableau 6.3 présente les résultats de la classification des GPCR au niveau des sous-sous-familles avec FS en utilisant l'algorithme BAT. Après plusieurs tests, nous avons constaté que la modification des paramètres: nombre de chauves-souris et nombre d'itérations, peut améliorer les résultats de classification, et la meilleure valeur obtenue était égale à 10 pour chacun de ces paramètres. Le nombre d'attributs a été choisi pour chaque classifieur après plusieurs expériences afin d'obtenir le meilleur. En outre, toutes les expérimentations ont été réalisées en utilisant une validation croisée de 10 dossiers.

	Classifier	**Att-num**	**ER**	**ACC**	**Fm**	**MCC**
AAC	BayesNet	9	0,11	1	1	1
	NaiveBayes	18	0,12	1	1	1
	IBK	13	0,05	1	0,92	0,92
	J48	18	0,09	1	1	1
	DecisionTable	18	0,09	1	1	1
	RF	13	0.04	0.99	0.98	0.98
	BAG	15	0.07	0.99	0.98	0.98
	LB	17	0.07	1	1	1
	ZeroR	/	0,76	0	0	0
PseAAC	BayesNet	9	0.14	1	0.97	0.97
	NaiveBayes	20	0.15	1	0.93	0.93
	IBK	44	0.06	1	0.97	0.97
	J48	9	0	1	1	1
	DecisionTable	18	0.22	1	0.99	0.99
	RF	24	0.04	0.99	0.98	0.98
	BAG	39	0.07	0.93	0.9	0.9
	LB	28	0.07	1	0.98	0.98
	ZeroR	/	0.76	0	0	0
Am-PseAAC	BayesNet	26	0.22	0.9	0.88	0.88
	NaiveBayes	63	0.25	0.92	0.82	0.83
	IBK	18/32	0.15	1	1	1
	J48	18/43	0.14	1	1	1
	DecisionTable	14	0.25	1	0.96	0.96
	RF	41	0.12	0.9	0.84	0.84

	BAG	43	0.11	1	1	1
	LB	50	0.13	0.82	0.75	0.76
	ZeroR	/	0.76	0	0	0
DC	BayesNet	199	0.08	0.93	0.93	0.9
	NaiveBayes	209	0.11	0.93	0.93	0.88
	IBK	79	0.07	0.92	0.92	0.89
	J48	250	0.08	0.91	0.91	0.85
	DecisionTable	235	0.18	0.85	0.85	0.74
	RF	96	0.06	0.98	0.97	0.97
	BAG	62	0.06	0.96	0.95	0.95
	LB	143	0.05	0.96	0.94	0.93
	ZeroR	/	0.76	0	0	0
LD	BayesNet	25	0,21	0,83	0,84	0,84
	NaiveBayes	36	0,32	0,72	0,77	0,77
	IBK	169	0,05	0,95	0,96	0,96
	J48	160	0,09	0,95	0,94	0,94
	DecisionTable	180	0,19	0,88	0,87	0,87
	RF	141	0.06	0.95	0.91	0.91
	BAG	32	0.06	0.95	0.94	0.94
	LB	69	0.06	0.93	0.93	0.93
	ZeroR	/	0,76	0	0	0

Tableau 6. 3: Evaluation des sous-ensemble d'attributs obtenus par l'algorithme Bat.

La qualité de la classification et la taille du sous-ensemble de caractéristiques sont deux critères considérés pour évaluer les performances des algorithmes. En comparant le premier critère, la précision prédictive et le taux d'erreur, nous avons remarqué que la sélection de caractéristiques à l'aide de l'algorithme de chauve-souris offre de meilleures précisions et des taux d'erreur par rapport à la classification sans sélection de caractéristiques (en utilisant tous les attributs) et ceci pour tous les AFD.

Selon le tableau 6.3:

- Pour les stratégies AAC et PseAAC, la précision prédictive atteint la valeur 1 dans la plupart des classifieurs: BN, NB, IBK, J48, DT et LB, tandis que la différence de MCC, Fm entre les algorithmes d'exploration de données n'est pas statistiquement significative. Néanmoins, les valeurs du taux d'erreur sont en variation continue, où la meilleure valeur est atteinte par le classificateur J48 et la plus mauvaise est égale à 0,22, elle est obtenue par l'algorithme DT en utilisant la méthode PseAAC. Notez que tous les exemples (séquences) sont mal classés en utilisant le classifieur ZR avec toutes les stratégies.
- Pour la méthode Am-PseAAC, les algorithmes IBK, BAG et J48 donnent une précision maximale, des valeurs MCC et Fm égales à 1, avec un taux d'erreur

minimal. Egalement le classifieur DT donne la même précision en utilisant le plus petit sous-ensemble d'attributs mais il y a une légère diminution dans les valeurs de MCC, Fm et la différence entre leurs taux d'erreur est statistiquement significative. Malgré la différence de la taille des sous-ensemble de caractéristiques, les classifieurs RF, NB et BN donnent presque des résultats proches, LB fournit les valeurs les plus mauvaises pour cette stratégie.

- Dans la méthode DC, tous les algorithmes réalisent des mesures de performance quasiment proches, où les classifieurs RF, BAG et LB donnent les meilleurs résultats et DT marque les valeurs les plus faibles.
- Contrairement aux méthodes précédentes, dans la stratégie LD, les valeurs les plus basses sont atteintes en utilisant le classificateur NB; généralement, IBK, J48, RF, BAG, LB et DT donnent de bons résultats.

En comparant le deuxième critère, le nombre d'attributs sélectionnées. Comme on peut le voir dans le tableau 6.3, l'algorithme BAT fonctionne bien dans la sélection d'un plus petit sous-ensemble d'attributs dans tous les classifieurs, et ce nombre change d'un classifieur à l'autre, cette différence du nombre de caractéristiques sélectionnées est statistiquement significative dans toutes les stratégies.

Les figures (Figure 6.7, Figure 6.8) illustrent respectivement les valeurs de accuracy et de taux d'erreur dans la classification standard pour les cinq stratégies utilisées. Ces figures sont présentées pour comparer les résultats de la classification avec FS et sans FS pour constater l'amélioration de la qualité de la classification. Nous voulons optimiser la classification des GPCR en maximisant sa précision et en minimisant son taux d'erreur, pour cela nous avons choisit une technique de FS en utilisant l'algorithme des chauves-souris (BAT).

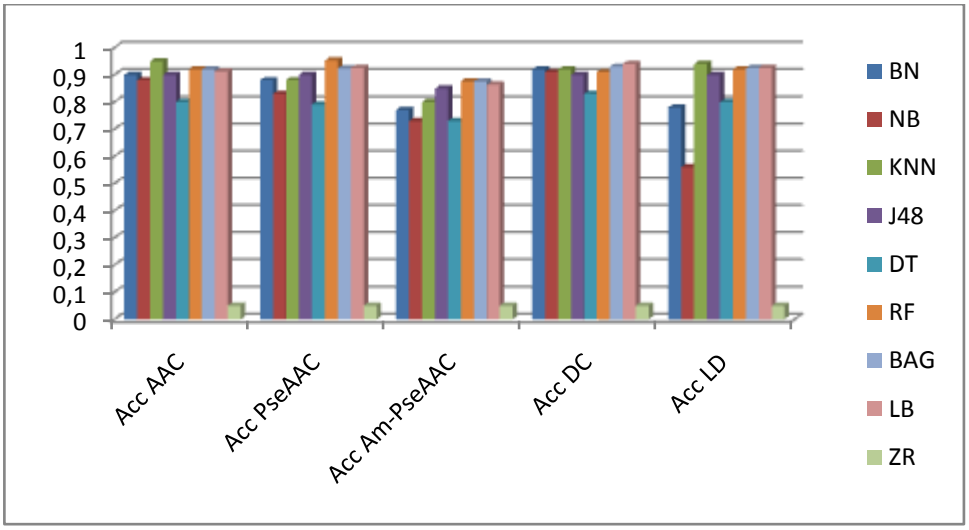


Figure 6. 7: Evaluation de l'accuracy de la classification des exemples sans FS.

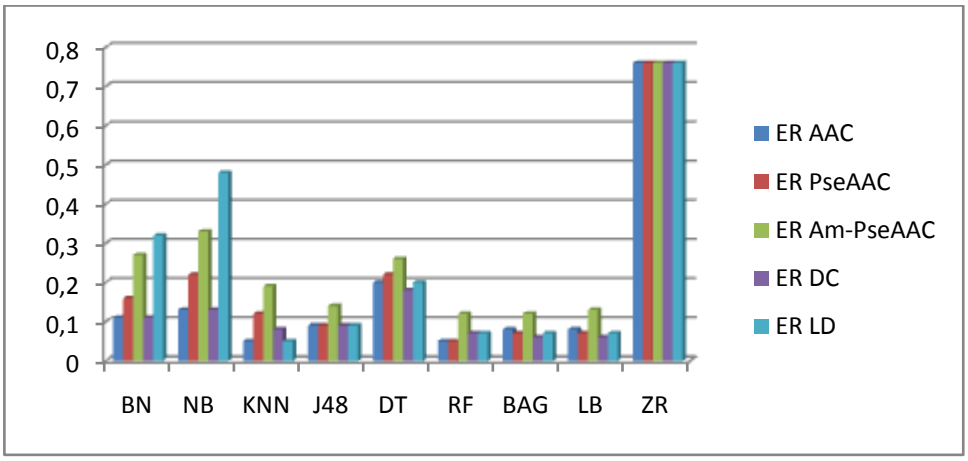


Figure 6. 8: Evaluation de taux d'erreur de la classification des exemples sans FS.

Pour plus d'illustration, les résultats obtenus dans le tableau précédent sont présentés dans les figures suivantes.

Les figures (Figure 6.9, Figure 6.10, Figure 6.11, Figure 6.12, Figure 6.13) présentent la variation des valeurs de mesures de performances (ACC, Fm, MCC, ER) en utilisant tous les algorithmes de classification pour chaque stratégie d'extraction de caractéristiques utilisées dans ce travail. La classification des séquences protéiques est effectuée en utilisant le meilleur sous-ensemble de caractéristiques sélectionnées par l'algorithme de chauve-souris.

Comparons ces figures toujours avec celles présentées au dessus (Figure 6.7, Figure 6.8) qui montrent l'évaluation de la classification standard des séquences avec tous les attributs (sans FS).

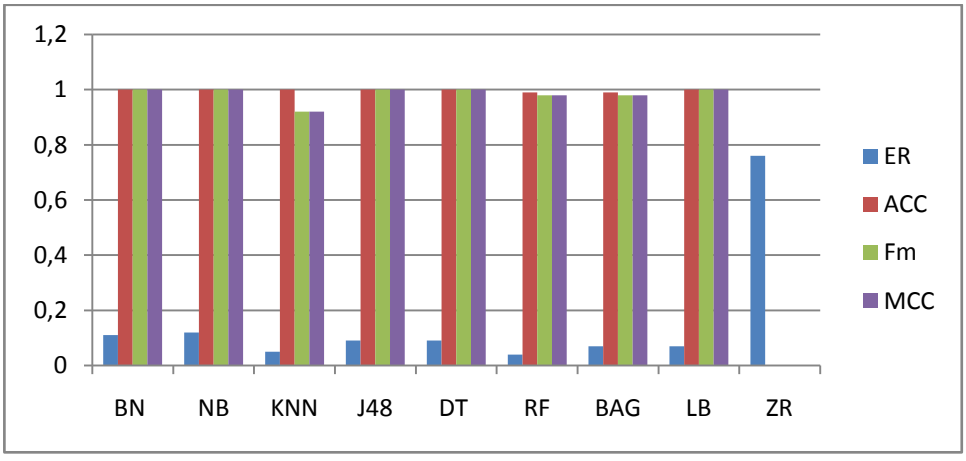


Figure 6. 9: Evaluation de FS par l'algorithme Bat pour la méthode AAC.

• Pour la méthode AAC (Figure 6.9), les valeurs de précision sont 1 pour six classificateurs (BN, NB, KNN, J48, DT, LB) en appliquant la sélection de caractéristiques (FS) par l'algorithme Bat. Cependant, cette valeur est incluse dans l'intervalle [0.8, 0,95] dans une classification standard par des algorithmes cités ci-dessus (Figure 6.7).

Contrairement à l'algorithme BAG qui marque une légère diminution de la valeur de précision en utilisant la technique FS. En ce qui concerne les valeurs du taux d'erreur, on peut remarquer une différenciation statistiquement non significative pour tous les AFD sauf le classifieur DT qui améliore significativement le taux d'erreur en utilisant la FS. Concernant le classifieur ZR, il donne de très mauvais résultats soit en utilisant la FS soit dans la classification standard pour tous les MRP.

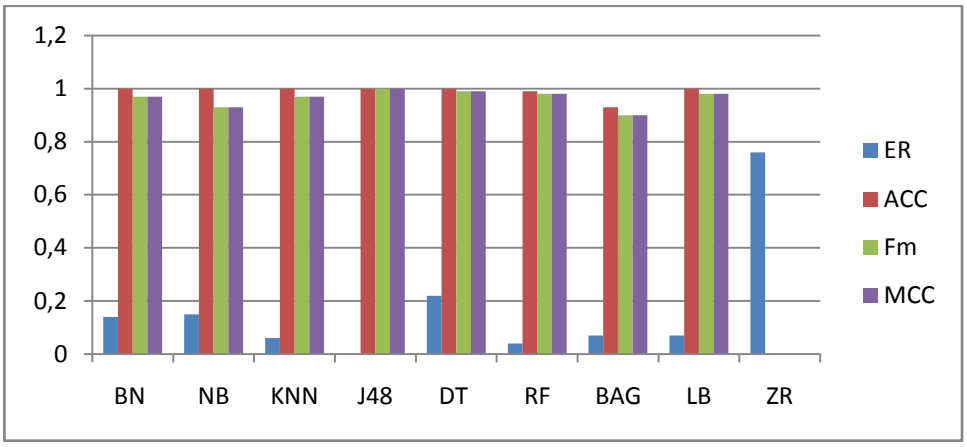


Figure 6. 10: Evaluation de FS par l'algorithme Bat pour la méthode PseAAC.

Les résultats obtenus en utilisant la stratégie PseAAC (Figure 6.10) sont plus ou moins efficaces par rapport à la méthode AAC dans la classification standard, mais la technique de FS par l'algorithme de chauve-souris (Bat) a amélioré efficacement les valeurs de mesures de performance, de telle sorte que la plupart des AFD ont atteints la valeur 1 de précision (BN, NB, KNN, J48, DT et LB), notez que les valeurs du taux d'erreur sont toujours en amélioration continue.

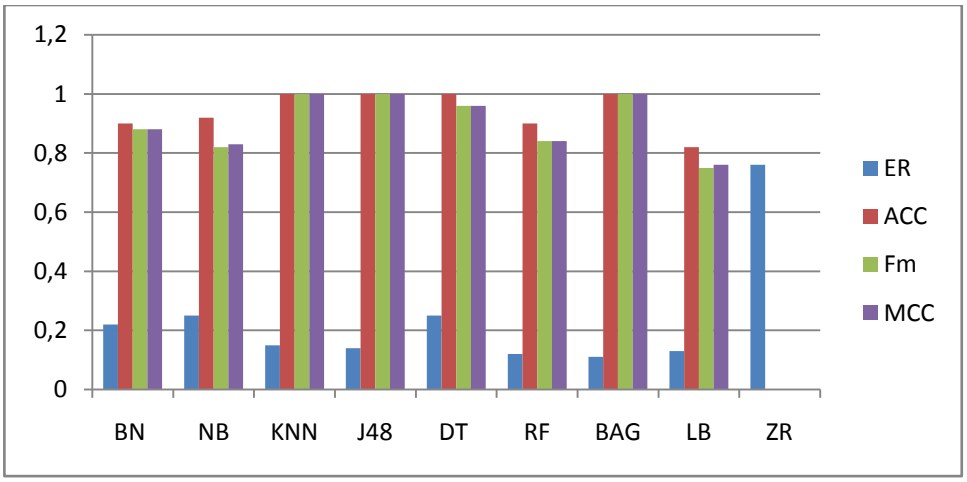


Figure 6. 11: Evaluation de FS par l'algorithme Bat pour la méthode Am-PseAAC.

Lors de l'utilisation de la stratégie Am-PseAAC (Figure 6.11), nous remarquons une diminution statistiquement significative des valeurs de mesure de performance par rapport aux MRP précédentes de telle sorte que le taux d'erreur est élevé et la précision est réduite pour tous les algorithmes de la classification standard, mais l'utilisation de la technique de FS améliore considérablement les résultats, de sorte qu'ils soient proches des résultats des deux stratégies précédentes. De plus, les classifieurs KNN, J48, DT et BAG fournissent des valeurs de précision égales à 1, mais l'algorithme LB marque une légère diminution de la valeur de précision en utilisant la technique FS.

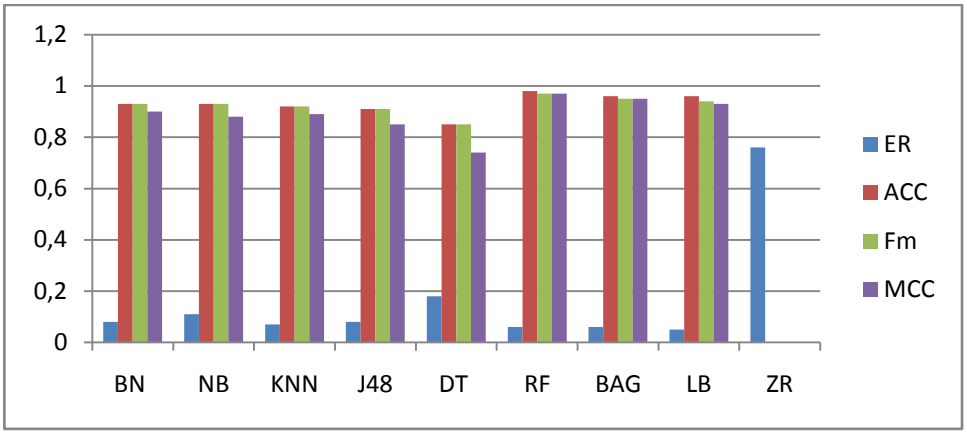


Figure 6. 12: Evaluation de FS par l'algorithme Bat pour la méthode DC.

• Généralement, les résultats de la classification standard par la méthode DC sont bons, les valeurs de précision des AFD sont proches et elles sont incluses dans l'intervalle [0,83, 0,94] de telle sorte que la valeur la plus minimale soit fournie par le classifieur DT. L'implémentation de l'algorithme BAT pour FS améliore également les performances de classification des GPCR pour tous les AFD utilisés. Notez que les classifieurs RF, BAG et LB donnent les meilleurs résultats par rapport au AFD restants, comme le montre la figure 6.12.

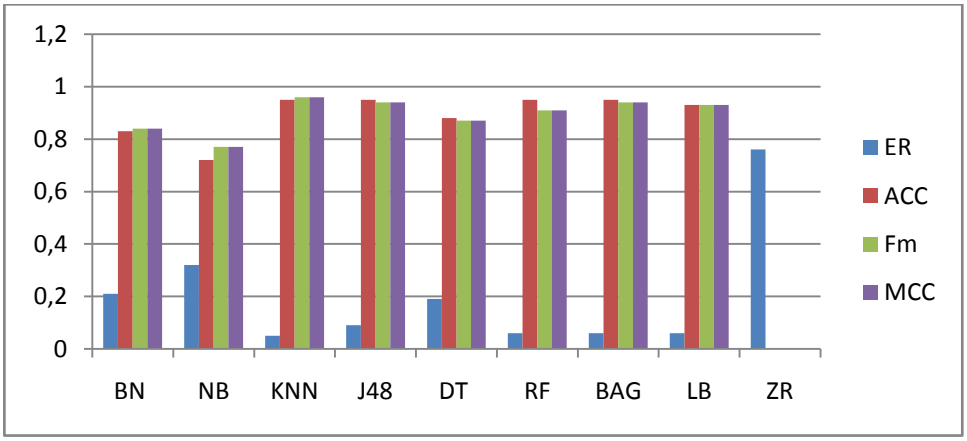


Figure 6. 13: Evaluation de FS par l'algorithme Bat pour la méthode LD.

• On peut observer l'utilité de la technique de FS dans la classification des protéines, en utilisant la stratégie LD, notamment dans l'algorithme NB (Figure 6.13), qui a fourni des valeurs de précision et un taux d'erreur égales respectivement à 0,56 et 0,48 dans la classification standard (Figure 6.7, Figure 6.8). Ces mesures devenait respectivement 0,72 et 0,32 en utilisant le meilleur sous ensemble d'attributs fournit par l'algorithme Bat. BN et NB qui ont donné de faibles résultats par rapport aux autres AFDs.

En résumé, une augmentation considérable des valeurs de mesures de performance au niveau de tous les algorithmes de classification sauf ZR qui est inapproprié pour les données biologiques. Malgré la nature difficile des séquences de protéines, tous les classifieurs produisent un bon résultat et classent les exemples d'une manière typique. Toute cette amélioration est due à l'utilisation de l'algorithme BAT pour la de sélection de caractéristique.

6.4. Comparaison avec des méthodes supplémentaires

Dans cette section, nous effectuons une étude qui compare notre approche basée sur l'algorithme de chauve-souris conçu pour la sélection de caractéristiques avec deux algorithmes bio-inspirés existants déployés sur l'environnement Weka (The PSOSearch et Evolutionary Algorithm Search). Les paramètres par défaut des algorithmes BAT, PSO et EA sont définis pour rendre la comparaison complètement équitable. Le tableau 6.4 montre les résultats de la classification des GPCR au niveau de sous-sous-famille avec FS en utilisant deux algorithmes (PSO et EA). Dans notre comparaison, trois paramètres sont pris en compte: le nombre d'attributs, la précision et le taux d'erreur. Les résultats (Tableau 6.3, Tableau 6.4) montrent que l'algorithme Bat proposé extrait un sous-ensemble d'attributs pertinent pour la base de données utilisée. En effet, la rigidité et les critères stochastiques de l'algorithme BAT sont introduits, de sorte qu'il ne génère que des sous-ensembles utiles et optimaux. Au contraire, les approches PSO et EA peuvent générer des sous-ensembles de caractéristiques inutiles, ce qui a diminué les performances de la classification.

	Classifier	**ER PSO**	**ACC PSO**	**ER EA**	**ACC EA**
PseAAC Att Number PSO/EA = 30/32	Bayes Net	0.14	0.88	0.15	0.88
	Naive Bayes	0.17	0.86	0.18	0.85
	KNN	0.12	0.87	0.12	0.87
	J48	0.09	0.9	0.09	0.9
	DecisionTable	0.22	0.79	0.22	0.79
	RF	0.06	0.93	0.06	0.92
	BAG	0.07	0.91	0.08	0.9
	LogitBoost	0.07	0.95	0.07	0.94
	ZeroR	0.76	0	0.76	0

Am-PseAAC Att Number PSO/EA = 12/14	Bayes Net	0.2	0.79	0.22	0.78
	Naive Bayes	0.23	0.75	0.27	0.72
	KNN	0.18	0.81	0.16	0.82
	J48	0.16	0.83	0.17	0.82
	DecisionTable	0.26	0.73	0.27	0.71
	RF	0.13	0.86	0.12	0.87
	BAG	0.12	0.87	0.12	0.86
	LogitBoost	0.15	0.81	0.14	0.8
	ZeroR	0.76	0	0.76	0
DC Att Number PSO/EA = 129/126	Bayes Net	0.1	0.91	0.09	0.92
	Naive Bayes	0.12	0.9	0.12	0.9
	KNN	0.06	0.93	0.06	0.93
	J48	0.1	0.9	0.09	0.9
	DecisionTable	0.2	0.8	0.19	0.82
	RF	0.07	0.94	0.06	0.95
	BAG	0.07	0.95	0.07	0.95
	LogitBoost	0.08	0.96	0.07	0.95
	ZeroR	0.76	0	0.76	0
LD Att Number PSO/EA = 7/67	Bayes Net	0.19	0.8	0.23	0.82
	Naive Bayes	0.24	0.76	0.35	0.74
	KNN	0.11	0.88	0.05	0.94
	J48	0.13	0.86	0.09	0.9
	DecisionTable	0.23	0.76	0.2	0.79
	RF	0.07	0.93	0.06	0.94
	BAG	0.08	0.94	0.07	0.93
	LogitBoost	0.08	0.9	0.07	0.91
	ZeroR	0.76	0	0.76	0

Tableau 6. 4: La performance de la classification des GPCR au niveau de ssous-sous-familles avec FS à l'aide des algorithmes PSOSearch / EA.

En général, les résultats rapportés dans le tableau 6.4 sont approximativement similaires, en particulier pour les trois premières méthodes, par exemple les classifieurs: KNN, J48, DT et ZR ont donné les mêmes valeurs de ACC et ER pour les algorithmes PSO et EA en utilisant la méthode PseAAC et les classifieurs restants marquent une différenciation statistiquement non significative. Pour les méthodes Am-PseAAC et DC, nous marquons une légère différenciation pouvant atteindre la valeur 0,03. Quant à la méthode LD, PSO a produit des valeurs ER meilleures que celles de EA pour les classifieurs BN, NB, et elles sont plus mauvaises pour les classifieurs KNN, J48, DT, RF, BAG, LB. Cependant, presque toutes les valeurs de précision sont meilleures en utilisant l'algorithme EA par rapport à PSO.

En comparant les résultats de la classification avec la sélection des attributs en utilisant les algorithmes PSO et AE avec ceux de l'algorithme Bat, nous avons constaté que dans la plupart du temps, l'algorithme BAT donnait les meilleurs résultats. Par la suite, nous aurons plus de détails en analysant les figures ci-dessous.

L'optimisation de la classification des GPCR avec FS n'affecte pas le classifieur ZR qui reste stable en utilisant tous les algorithmes bio-inspirés pour toutes les MRP.

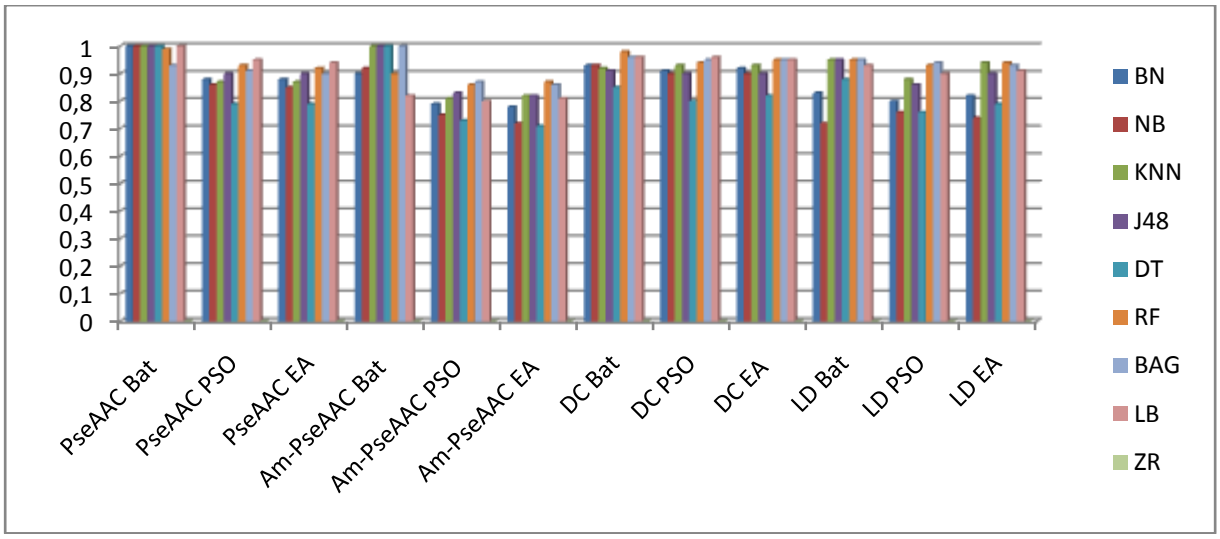


Figure 6. 14: Comparaison générale des valeurs de précision pour les MRP utilisées à l'aide des algorithmes EA / PSO et Bat.

Chaque figure présente la variation des valeurs des deux mesures de performance, la première (Figure 6.14) est dédiée aux valeurs de la précision prédictive, toutes les barres relatives à l'algorithme bat indiquent que quelle que soit la méthode utilisée, tous les AFD atteignent les meilleures précisions par rapport aux EA / PSO, à l'exception du classifieur NB qui produit les valeurs de précision / ER les plus faibles en utilisant la stratégie LD. Notez également qu'en utilisant la méthode Am-PseAAC, les classifieurs BN et NB donnent des valeurs de ER meilleures que EA et mauvaises par rapport à l'algorithme PSO.

Le FS en utilisant l'algorithme BAT a considérablement amélioré les performances de la classification des GPCR, en particulier dans les méthodes PseAAC et Am-PseAAC.

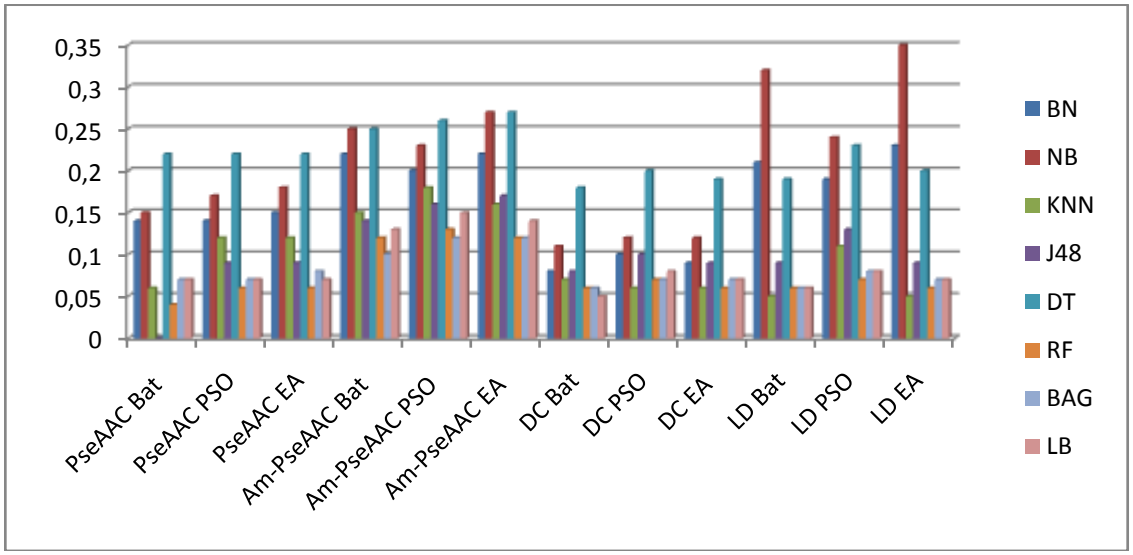


Figure 6. 15: Comparaison générale des valeurs du taux d'erreur pour le MRP utilisées à l'aide des algorithmes EA / PSO et Bat.

La deuxième figure (Figure 6.15) est consacrée aux valeurs du taux d'erreur. Il est clair que certains AFD donnent des valeurs similaires ou très proches pour tous les algorithmes de FS

tels que: BN, DT, BAG, RF et LB en utilisant la méthode PseAAC, BN, DT, RF, BAG et LB pour la stratégie Am-PseAAC. Cependant en utilisant la stratégie DC, tous les classifieurs produisent des valeurs de taux d'erreur très proches avec une différenciation statistiquement non significative pouvant atteindre 0,02, contrairement à la méthode LD qui marque une variation considérable notamment au niveau des algorithmes NB, BN, KNN, J48, DT.

Enfin, après la comparaison de l'algorithme BAT avec PSO et EA, nous avons constaté que, dans une très grande majorité, l'algorithme BAT donnait les meilleurs résultats. Par conséquent, il est adapté et efficace dans le domaine de la bioinformatique, et en particulier pour la classification des GPCR.

6.5. Conclusion

L'algorithme chauve souris a été utilisé pour optimiser les résultats dans de nombreuses applications. Dans ce travail, une nouvelle méthode basée sur l'algorithme cet algorithme est proposé et mis en œuvre pour la sélection des caractéristiques et la classification des GPCR en utilisant plusieurs méthodes de représentation des protéines. Cet algorithme utilise les avantages de la capacité d'écholocation et les résultats montrent un comportement prometteur.

Le système proposé a abouti à une taille réduite du sous-ensemble de caractéristiques, à une précision de classification accrue, à une faible complexité de calcul et à une convergence rapide. Notre approche aussi a abouti à des performances comparables aux meilleurs résultats rapportés dans la littérature.

À l'avenir, nous aimerions utiliser la méthode proposée basée sur l'algorithme de chauve-souris sur des applications bioinformatiques supplémentaires qui nécessitent une optimisation tel que: l'extraction des motifs protéiques et l'expression de gènes.

Conclusions et perspectives

Conclusions

Avec les progrès récents des techniques de séquençage génomique, le nombre de séquences protéiques disponibles pour l'analyse a considérablement augmenté. Les processus de la prédiction de la fonction d'une protéine sont trop coûteux pour répondre à cette demande. Cependant, connaître la fonction d'une protéine est extrêmement important dans plusieurs domaines tels que la médecine, la pharmacologie et l'agriculture. Par conséquent, il est nécessaire de trouver des modèles informatiques capables de prédire la fonction des protéines. C'est un domaine de recherche ouvert et très intéressant en bioinformatique, car les modèles existants ne fonctionnent pas encore assez bien. Les bases de données des GPCR rassemblent des informations pertinentes liées aux caractéristiques physico-chimiques des protéines, qui ont été prises en compte dans certaines recherches, la plupart limitées à quelques ensembles de caractéristiques.

Cependant, certains auteurs ont constaté que nous ne connaissons que les fonctions de 5% des protéines découvertes. Face à ce scénario, nous avons besoin de méthodes de calcul pour automatiser et faciliter le processus d'identification de la fonction d'une protéine. Actuellement, plusieurs procédés expérimentaux sont utilisés à cet effet. Cependant, aucune des approches existantes n'est en mesure de prédire avec une bonne précision la fonction d'un grand ensemble de protéines. Par conséquent, le problème de la prédiction de la fonction des protéines est un défi pour la biologie moléculaire et la bioinformatique en général.

Une pléthore d'informations concernant la structure des protéines est disponible sur des ensembles de données publics. Cependant, nous ne savons pas quelles variables ou caractéristiques sont les plus pertinentes pour distinguer la fonction d’une protéine.

Nous considérons dans ce travail l'hypothèse qu'une prédiction efficace de la fonction des protéines nécessite des informations provenant de différentes caractéristiques biologiques, autrement dit, une bonne représentation des chaines protéiques, sous forme des vecteurs d'attributs numériques. Ainsi, il est nécessaire de trouver un ensemble de caractéristiques qui représente le mieux la fonction protéique, pour résoudre le problème de classification. Cela signifie que nous devons sélectionner les informations les plus pertinentes parmi différents échantillons d'une base de données et construisez un modèle de classification à l'aide de ceux-ci. Dans les problèmes d'apprentissage automatique supervisé tels que la classification, la sélection des caractéristiques et la réduction de la dimensionnalité peuvent optimiser le

temps de calcul nécessaire à l'algorithme de fouille de données, tout en améliorant la précision de la prédiction et en rendant le modèle plus facile à interpréter.

La sélection de caractéristiques est un problème d'optimisation combinatoire, qui vise à trouver le sous-ensemble optimal de caractéristiques d'un critère donné parmi plusieurs possibilités dans un espace de solution. Dans le contexte de la prédiction de la fonction protéique, il n'est pas possible de garantir que les caractéristiques choisies et la taille du sous ensemble sont les meilleures pour un ensemble de données.

De plus, il y a deux problèmes majeurs dans la tâche de la prédiction computationnelle des fonctions protéiques avec des algorithmes de classification, qui sont le choix de la méthode de représentation de protéines et le choix de l'algorithme de fouille de données. Ce sont des problèmes ouverts dans le scénario conventionnel de la classification à plat (où il n'y a pas de relations hiérarchiques entre les classes), car il y a beaucoup de choix et il n'est pas clair quel stratégie de représentation et quel algorithme de classification sont les meilleurs.

Ceci nous a incités à envisager l'utilisation des approches bio-inspirées dans le cadre de notre travail pour la fouille de données en bioinformatique pour résoudre les problème cités au dessus afin d'améliorer la précision prédictive. Nous nous sommes particulièrement intéressés à deux familles différentes d'approches bio-inspirées. La première est celle des Algorithmes Génétiques qui s'inspirent de l'évolution naturelle et que nous avons appliquée, avec succès, au problème de sélection du meilleur couple (MRP/AFD). La deuxième est celle de l'algorithme BAT inspiré du comportement d'écholocation des chauves-souris pour résoudre le problème de sélection de caractéristiques afin d'optimiser la classification et diminuer les erreurs tel qu'il est présenté dans notre travail antérieur [BEK 20]. Ces deux approches ont fait leurs preuves dans plusieurs domaines applicatifs comme il est illustré dans le chapitre 2.

Une étude préalable des méthodes de représentation de séquences des RCPG a été faite dans ce travail (Chapitre 4). Cette étude empirique où différentes approches d'extraction de caractéristiques pour représenter les protéines sont comparées à l'aide d'un pool précis et diversifié de classifieurs. Nous obtenons un certain nombre d'observations statistiquement robustes selon le comportement des différentes mesures de performance testées ici. Les principales conclusions selon [BEK 18, BEK 19], qui peuvent être tirées des résultats sont:

- ✓ La stratégie d'extraction de caractéristiques donne de bons résultats avec certains classificateurs mais pas avec les autres.
- ✓ Le même algorithme d'exploration de données marque différents résultats en utilisant différentes approches de représentation des protéines.

- ✓ Pour différents ensembles de données, la meilleure méthode de représentation des protéines et les algorithmes d'apprentissage automatique peuvent également être différents.
- ✓ Optimiser l'identification des fonctions des GPCR, nécessite un meilleur choix d'une méthode de représentation protéique, pour une bonne classification. Ce choix est varié en fonction du contexte de travail, du jeu de données d'entrée et surtout de l'algorithme de classification utilisé car certains algorithmes sont sensibles au bruit, à l'ambiguïté et au flou des données.

Perspectives

Les études expérimentales réalisées, nous a permis de déterminer l'approche bio-inspirée qui nous semble la plus prometteuse pour la prédiction de fonctions des protéines et c'est dans cette voie que nous souhaitons continuer.

De nombreuses perspectives peuvent être envisagées suite à cette thèse, notamment :

- ✓ L'utilisation d'autres méthodes d'extraction de caractéristiques pour générer différents vecteurs d'attributs numériques en terme de taille et contenu, ainsi que d'autres classifieurs pour confirmer et enrichir les démonstrations obtenues dans le chapitre 4.
- ✓ Il serait également important d'utiliser plusieurs bases de données ou ensembles de données de différentes protéines telle que : les enzymes.
- ✓ La comparaison avec d'autres travaux existants dans la littérature constitue une autre perspective de nos travaux.
- ✓ Bien que l'algorithme BAT que nous avons utilisé nous donnait des résultats prometteurs et parfaits, nous projetons d'évaluer d'autres méthodes bio-inspirées telles que les essaims de particules et les colonies d'abeilles artificielles qui figurent parmi les modèles les plus récents et représentent des algorithmes de faible complexité, qui ont eu beaucoup de succès dans de nombreux domaines d'application durant la dernière décennie.

Bibliographie

[AGH 08] Aghdam MH, Ghasem-Aghaee N, Basiri ME, Application of ant colony optimization for feature selection in text categorization. In Proceeding of the fifth IEEE congress on evolutionary computation, IEEE Press. 2008

[AHM 15] Ahmad S, Kabir. M, Hayat. M, Identification of Heat Shock Protein Families and J-Protein Types by incorporating Dipeptide Composition into Chou‘s general PseAAC‘. Computer Methods and Programs in Biomedicine. 2015

[AKH 12] Akhtar, S., A. R. Ahmad, and E. M. A. Rahman. A metaheuristic bat inspired algorithm for full body human pose estimation. In IEEE conference on computer and robot vision9:369–375. 2012

[ALB 02] Alberts B, Johnson A, Lewis J, et al., Molecular Biology of the Cell. 4th edition. New York: Garland Science; 2002.

[AL-B 20] Al-Betar M.A., Alomari O.A., Abu-Romman S. M., A TRIZ-inspired bat algorithm for gene selection in cancer classification. Genomics. 2020

[AL-S 04] Al-Shahib A., Chao H., Aik C., Tan Mark G., and Gilbert D., "An Assessment of Feature Relevance in Predicting Protein Function from Sequence. In the Proceedings of the Fifth International Conference on Intelligent Data Engineering and Automated Learning (IDEAL04)," 2004.

[AL-S 05] Al-Shahib A., Breitling R., and Gilbert D., "Franksum: New Feature Selection Method For Protein Function Prediction," International Journal Of Neural Systems, vol. 15, pp. 259–275, 2005.

[ALS 14] Alsariera Y. A, Alamri H. S., Nasser A. M., Majid M. A., Zamli K. Z., Comparative Performance Analysis of Bat Algorithm and Bacterial Foraging Optimization Algorithm using Standard Benchmark Functions. 2014 8th Malaysian Software Engineering Conference (MySEC). IEEE, 2014

[ANG 94] Angeline, P. J. 1994. Genetic Programming and Emergent Intelligence. Advances in Genetic Programming. Ed. K. E. Kinnear. Cambridge, MA. MIT Press.
[AZA 11] Azad A. K. M., Shahid S., Noman N., and Lee H. Prediction of plant promoters based on hexamers and random triplet pair analysis. Algorithms for Molecular Biology 6:19. 2011.

[BAG 16] Bagherzadeh-Khiabani F, Ramezankhani A, Azizi F, Hadaegh F, Steyerberg EW, Khalili D.A, ‘Tutorial on variable selection for clinical prediction models: feature selection methods in data mining could improve the results’. J Clin Epidemiol, Vol 71, pp.76–85. 2016

[BAR 71] Bartholomew, D.J., Operational Research Quarterly (1970-1977), 1971. 22(2): p. 199-201.

[BAS 18] Basith, S.; Manavalan, B.; Shin, T.H.; Lee, G. iGHBP: Computational identification of growth hormone binding proteins from sequences using extremely randomised tree. Comput. Struct. Biotechnol. J. 16, 412–420. 2018

[BEK 18] Bekhouche S., Ben Ali Y., Comparative analysis on features extraction strategies for GPCR classification, In Proceedings of Conference: 2018 4th International Conference on Computer and Technology Applications (ICCTA). 2018.

[BEK 19] Bekhouche S., Ben Ali Y., Optimizing the identification of GPCR function, In Proceedings of the New Challenges in Data Sciences: Acts of the Second Conference of the Moroccan Classification Society - SMC '19, 1-5. 2019

[BEK 20] Bekhouche S., Ben Ali Y., Feature Selection in GPCR Classification Using BAT Algorithm, International Journal of Computational Intelligence and Applications, Vol. 19, No. 01, 2050006. 2020

[BEN 07] Bennani and S. Guerif. Sélection de variable en apprentissage numérique non supervisé. In cap 07 : conférence francophone sur l'apprentissage automatique, 2007.

[BEN 16] Ben Othman M.T., Survey of the use of genetic algorithm for multiple sequence alignment, Journal of Advanced Computer Science & Technology, 5 (2) pp. 28-33, (2016).

[BEN 00] Bentley D.R.,"The human genome project—an overview," Medicinal Research Reviews, vol.20, pp.189–196, 2000.

[BER 08] Bertolazzi P., Felici G., Festa P., and Lancia G., "Logic classification and feature selection for biomedical data," Computers & Mathematics with Applications, vol. 55, pp. 889-899, 2008.

[BHA 04a] Bhasin, M. and G.P. Raghava, Classification of nuclear receptors based on amino acid composition and dipeptide composition. J Biol Chem, 279(22): p. 23262-6. 2004

[BHA 04a] Bhasin, M., Raghava, G.P.S. GPCRpred: An SVMbased Method for Prediction of Families and Subfamilies of G-Protein Coupled Receptors', Nucleic Acids, vol. 32, pp. 383–389. 2004.

[BHA 04b] Bhasin, M. and G.P. Raghava, ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. Nucleic Acids Res, 2004. 32(Web Server issue): p. W414-9.

[BHA 04c] Bhasin, M., Raghava, G.P.S.: GPCRpred: An SVM-based Method for Prediction of Families and Subfamilies of G-Protein Coupled Receptors, Nucleic Acids Research, vol. 32 (suppl 2), pp. W383–W389, 2004

[BRO 84] Broto, P., G. Moreau, and C. Vandycke, Molecular Structures – Perception, Auto-correlation Descriptor Eur. J. Med. Chem., 1984. 19: p. 71-78.

[BYW 16] Bywater, Robert P. Comparison of algorithms for prediction of protein structural

features from evolutionary data. PLoS ONE 11(3):e0150769. doi:10.1371/journal.pone.0150769. . 2016

[CAI 03] Cai C. Z., Han L. Y., Ji Z. L., Chen X., and Chen Y. Z., "SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence " Nucleic Acids Research, vol. 31, pp. 3692–3697 2003.

[CAI 19] Cai X., Zhang J.,Liang H., Wang L., Wu Q., An ensemble bat algorithm for large‑scale optimization, International Journal of Machine Learning and Cybernetics. Vol. 10, pp.3099–3113. 2019

[CAO 13] Cao, D.S., Q.S. Xu, and Y.Z. Liang, propy: a tool to generate various modes of Chou's PseAAC. Bioinformatics, 2013. 29(7): p. 960-2.

[CAP 07] Capra J. A. and Singh M., "Predicting functionally important residues from sequence conservation," Bioinformatics, vol. 23, pp. 1875-1882, 2007.

[CHA 15] Chawla M., Duhan M., Bat Algorithm: A Survey of the State-of the-Art, Applied Artificial Intelligence: An International Journal 29:6, 617-634, 2015.

[CHE 14] Chen S., Montgomery J., Bolufe-Rohler A., Measuring the curse of dimensionality and its effects on particle swarm optimization and differential evolution. Appl Intell. doi:10.1007/s10489-014-0613-2. 2014

[CHE 03] Chen X. An improved branch and bound algorithm for feature selection. Pattern Recognition Letters, 24(12):1925-1933. 2003.

[CHE 06] Chen Y., Abraham A., and Yang B., "Feature selection and classification using flexible neural tree," Neurocomputing, vol. 70, pp. 305-313, 2006.

[CHO 00] Chou KC . Prediction of Protein Subcellar Locations by Incorporating Quasi-Sequence Order Effect." Biochemical and Biophysical Research Communications, 278, 477-483. 2000.

[CHO 01] Chou, K.C.: Prediction of Protein Cellular Attributes using Pseudo-Amino Acid Composition, Proteins, vol. 43, pp. 246–255, 2001.

[CHO 05] Chou K-C., Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes‘. Bioinformatics, vol. 21, pp.10–19. 2005

[CHO 06] Chou KC, Shen H-B., Predicting protein subcellular location by fusing multiple classifiers‘. J Cell Biochem, vol.99, pp.517–527. 2006.

[CHO 11 a] Chouaib H., Sélection de caractéristiques: méthodes et applications. Thèse présentée pour l'obtention du grade de Docteur de l'université Paris Descartes. 2011.

[CHO 11] Chou, K.C., Some remarks on protein attribute prediction and pseudo amino acid composition. J Theor Biol, 2011. 273(1): p. 236-47

[CHO 16] Chowdhury B., Garai A., and Garai G. An optimized approach for annotation of large eukaryotic genomic sequences using genetic algorithm. bioRxiv, 2016.

[COH 04] Cohen, J.: Bioinformatics—An Introduction for Computer Scientists, ACM Computing Surveys, vol. 36, issue 2, pp. 122–158, 2004

[COR 07] Correa ES., Freitas AA., and Johnson CG., (2007) 'Particle Swarm and Bayesian Networks Applied to Attribute Selection for Protein Functional Classification', GECCO'07, Jul y 7–11, 2007, pp. 2651- 2658. London, England, United Kingdom.

[COR 79] Cornish-Bowden, A., How reliably do amino acid composition comparisons predict sequence similarities between proteins? J Theor Biol, 1979. 76(4): p. 369-86

[COS 08] Costa E.P., Lorena A.C., Carvalho A.C.P.L.F., Freitas A.A., 'Top-Down Hierarchical Ensembles of Classifiers for Predicting G-Protein-Coupled-Receptor Functions'. LNBI, pp.35-46. 2008

[COV 91] Cover, T. M. et Thomas, J. A. Elements of information theory. Wiley-Interscience, New York, NY, USA. 1991.

[CUI 07] Cui J., Han L.Y., Li H., Ung C.Y., Tang Z.Q., Zheng C.J., Cao Z.W. and Chen Y.Z. Computer prediction of allergen proteins from sequence derived protein structural and physicochemical properties', Mollecular Immunology, Vol. 44, pp.514–520. 2007

[CUI 19] Cui, Z.; Zhang, J.; Wang, Y.; Cao, Y.; Cai, X.; Zhang, W.; Chen, J. A pigeon-inspired optimization algorithm for many-objective optimization problems. Sci. China Inf. Sci 2019.

[DAS 97] Dash, M. et Liu, H.. Feature selection for classification. Intelligent Data Analysis, 1:131_156, 1997.

[DAV 07] Davies, M., Secker, A., Freitas, A., Mendao, M., Timmis, J., and Flower, D. On the hierarchical classification of G protein-coupled receptors. Bioinformatics 23, 23 (2007), 3113-3118.

[DAV 08] Davies, M.N., Secker, A., Freitas, A.A., Clark, E., Timmis, J., Flower, D.R.: Optimizing Amino Acid Groupings for GPCR Classification, Bioinformatics, vol. 24, issue 18, pp. 1980–1986, 2008.

[DEA 06] Dean R. A. Variable selection for model-based clustrering. Journal of the American Statistical Association, 473 :169–178, 2006.

[DIA 05] Diamandis EP, van der Merwe DE. Plasma protein profiling by mass spectrometry for cancer diagnosis: opportunities and limitations. Clin Cancer Res; 11:963–5. 2005

[DRE 00] Drews J., Drug discovery: a historical perspective. Science 287, pp.1960–1964. 2000

[DUD 00] Duda, R. O., Hart, P. E., Stork, D. G. Pattern Classification (2nd Edition). Wiley-Interscience. 2000.

[FOR 93] Forrest, Stephanie. "Genetic algorithms: principles of natural selection applied to computation." Science, vol.261, p.872-878, 1993

[FRE 07] Freitas, A. A., and de Carvalho, A. C. P. L. F.'A tutorial on hierarchical classification with applications in bioinformatics. Research and Trends in Data Mining Technologies and Application', Vol. 99(7), pp.175–208. 2007

[FUR 00] Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M. et Haussler, D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics, 16(10):906-914. 2000

[GAN 13] Gandomi A. H., Yang X-S., Alavi A. H., Talatahari S., Bat algorithm for constrained optimization tasks, Neural Computing and Applications, vol. 22 Issue 6, pp.1239–1255. 2013

[GAO 06] Gao, Q.B., Wang, Z.Z., (2006) Classification of G Protein Coupled Receptors at Four Levels', Protein Engineering, Design & Selection, vol. 19, issue 11, pp. 511–516.

[GAO 13] Gao Q.B., Ye X.F., He J., Classifying G-Protein Coupled Receptors to the Finest Subtype Level', Biochemical and Biophysical Research Communications, vol.439, pp.303–308. 2013.

[GEA 54] Geary, R.C., The Contiguity Ratio and Statistical Mapping. The Incorporated Statistician, 1954. 5(3): p. 115-146.

[GHA 15] Ghaheri A., Shoar S., Naderan M., and Hoseini S. S. 2015. The Applications of Genetic Algorithms in Medicine. Oman Medical Journal 30(6), pp. 406–416. http://doi.org/10.5001/omj.2015.82

[GOL 99] Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.Science;286:531–7. 1999

[GOL 99] Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A. et Bloomfield, C. D. Molecular classification of cancer : class discovery and class prediction by gene expression monitoring. Science, 286:531-537. 1999

[GOV 11] Govindan, G. and A.S. Nair. Composition, Transition and Distribution (CTD). A dynamic feature for predictions based on hierarchical structure of cellular sorting. in India Conference (INDICON), Annual IEEE. 2011.

[GRA 74] Grantham, R., Amino acid difference formula to help explain protein evolution. Science, 1974. 185(4154): p. 862-4.

[GU 09] Gu, Q., Ding, Y.: Binary Particle Swarm Optimization based Prediction of G Protein-coupled Receptor Families with Feature Selection, ACM/SIGEVO Summit on Genetic and Evolutionary Computation (GEC), pp. 171–176, 2009.

[GU 15] Gu, Q., Ding, Y.S., Zhang, T.L.: An Ensemble Classifier based Prediction of G

Protein-Coupled Receptor Classes in Low Homology, Neurocomputing, vol. 154, pp.110–118, 2015.

[GUO 19] Guo S.S., Wang J-S., Ma X-X. Improved Bat Algorithm Based on Multi-population Strategy of Island Model for Solving Global Function Optimization Problem' Computational Intelligence and Neuroscience, Volume 2019, 12 pages. 2019

[GUY 03] Guyon, I. et Elisseeff, A. An introduction to variable and feature selection. J. Mach. Learn. Res., 3:1157-1182. 2003

[HAI 89] Haining, R.P., Geography, 1989. 74(1): p. 81.

[HAN 04] Han L. Y., Cai C. Z., Ji Z. L., Cao Z. W., Cui J., and Chen Y. Z., "Predicting functional family of novel enzymes irrespective of sequence similarity: a statistical learning approach," Nucl. Acids Res., vol. 32, pp. 6437-6444, December 7, 2004

[HAR 73] Harris, C.E. and D.C. Teller, Estimation of primary sequence homology from amino acid composition of evolutionary related proteins. J Theor Biol, 1973. 38(2): p. 347-62.

[HAS 01] Hastie, T., Tibshirani, R. et Friedman, J. The Elements of Statistical Learning. Springer series in statistics. Springer, New York. 2001.

[HOL 06] Holden, N., Freitas, A.A.: Hierarchical Classification of G-Protein-Coupled Receptors with a PSO/ACO Algorithm, IEEE Swarm Intelligence Symposium (SIS), pp. 77–84, 2006.

[HOL 08] Holden N., Freitas AA. Improving the Performance of Hierarchical Classification with Swarm Intelligence. E. Marchiori and J.H. Moore (Eds.): EvoBIO, LNCS 4973, pp. 48–60, 2008

[HOL 09] Holden N., Freitas AA. Hierarchical classification of protein function with ensembles of rules and particle swarm optimisation. Soft Comput 13, pp.259–272. 2009

[HUA 04] Huang, Y., Cai, J., Ji, L., Li, Y.: Classifying G-Protein Coupled Receptors with Bagging Classification Tree, Computational biology and chemistry, vol. 28, issue 4, pp. 275–280, 2004.

[HUA 06] Huang C.-L. and Wang C.-J., "A GA-based feature selection and parameters optimization for support vector machines," Expert System with Application, vol. 31, pp. 231–240, 2006.

[HUA 07] Jinjie Huang a,b,*, Yunze Cai b , Xiaoming Xu, A hybrid genetic algorithm for feature selection wrapper based on mutual information. Pattern Recognition Letters 28 1825–1844. 2007

[HUA 08] Huang C. L. and Dun J.-F., "A distributed pso–svm hybrid system with feature selection and parameter optimization,"Applied Soft Computing, vol. 8, p. 1381–1391, 2008.

[HUA 08] Huang, C.-J., Yang, D.-X. et Chuang, Y.-T. Application of wrapper approach and composite classifier to the stock trend prediction. Expert Syst. Appl., 34:2870-2878. 2008.

[HWA 13] Hwang K., Ha B., Ju S., and Kim S. 2013. Partial AUC maximization for essential gene prediction using genetic algorithms. BMB Reports 46:41-46. doi: 10.5483/BMBRep.2013.46.1.159.

[IQB 14] Iqbal M.J., Faye I., Belhaouari Samir B., and Said A.M., Efficient Feature Selection and Classification of Protein Sequence Data in Bioinformatics, Hindawi Publishing Corporation, The Scientific World Journal, Volume 2014, 12 pages. 2014.

[ISH 00] Ishibuchi, H. et Nakashima, T. Multi-objective pattern and feature selection by a genetic algorithm. In Proc. of Genetic and Evolutionary Computation Conference (GECCO'2000, pages 1069-1076. Morgan Kaufmann. 2000.

[ISM 16] Ismail H.D., Smith M. and KC D.B, FEPS: Feature Extraction from Protein Sequences webserver. 2016.

[JAI 97] Jain, A. et Zongker, D. Feature selection : Evaluation, application, and small sample performance. IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, 19:153-158. 1997.

[JEF 04] Jeffries N.O, Performance of a genetic algorithm for mass spectrometry proteomics, BMC Bioinformatics, 5(1):180. 2004.

[JEN 02] Jensen L. J., Gupta R., Blom N., Devos D., Tamames J., Kesmir C., Nielsen H., Staerfeldt H. H., Rapacki K., Workman C., Andersen C. A. F., Knudsen S., Krogh A., Valencia A., and Brunak S., "Prediction of Human Protein Function from Post-translational Modifications and Localization Features," Journal of Molecular Biology, vol. 319, pp. 1257-1265, 2002.

[JEN 05] Jensen R. and Shen Q., "Finding rough set reducts with ant colony optimization," Journal of Fussy Sets and Systems, vol. 149, pp. 5-20, 2005.

[JOH 94] John, G. H., Kohavi, R. et Pfleger, K. Irrelevant features and the subset selection problem. In MACHINE LEARNING : PROCEEDINGS OF THE ELEVENTH INTERNATIONAL, pages 121-129. Morgan Kaufmann. 1994.

[JOH 97] John, G. H. Enhancements to the data mining process. Thèse de doctorat, Stanford, CA, USA. UMI Order No. GAX97-23376. 1997. In Chouaib H., Sélection de caractéristiques: méthodes et applications. Thèse présentée pour l'obtention du grade de Docteur de l'université Paris Descartes.

[JUD 97] Judson R.S., Genetic algorithms and their use in chemistry. In: Lipkowitz KB, Boyd DB (eds) Rev. Computational Chemistry, vol 10. Wiley-VCH, Weinheim, pp 1–73. 1997.

[JUD 08] Judson R.S., Genetic Algorithms for Protein Structure Prediction. In: Floudas C., Pardalos P. (eds) Encyclopedia of Optimization. Springer, Boston, MA. 2008

[KAC 10] Kachouri, R., Djemal, K. et Maaref, H. Adaptive feature selection for heterogeneous image databases. In Djemal, K. et Deriche, M., éditeurs : Second IEEE

International Conference on Image Processing Theory, Tools 38 ; Applications, 10, Paris, France. 2010

[KAV 13] Kaveh, A., and P. Zakian. Enhanced bat algorithm for optimal design of skeletal structures. Asian Journal of Civil Engineering 15(2):179–212. 2013

[KAV 17] Kavakiotis I., Tsave O., Salifoglou A., Maglaveras N., Vlahavas I., Chouvarda I., 'Machine Learning and Data Mining Methods in Diabetes Research'. Computational and Structural Biotechnology Journal, Vol. 15, pp. 104–116. 2017

[KEE 05] Keedwell E., Narayanan A., Intelligent Bioinformatics: The application of artificial intelligence techniques to bioinformatics problems. John Wiley & Sons Ltd. DOI:10.1002/0470015721. 2005

[KHA 11] Khan, K., A. Nikov, and A. Sahai. A fuzzy bat clustering method for ergonomic screening of office workplaces. InAdvances in intelligent and soft computing, 101:59–66. Berlin, Heidelberg: SpringerVerlag. 2011

[KHA 15] Khan Z.U., M. Hayat, Khan M.A., Discrimination of acidic and alkaline enzyme using Chou's pseudo amino acid composition in conjunction with probabilistic neural network model', Journal of theoretical biology, vol. 365, pp. 197 -203. 2015

[KHA 17] Khan M., Hayat M., Khan S.A., Iqbal N., UnbDPC: Identify mycobacterial membrane protein types by incorporating un-biased dipeptide composition into Chou's general PseAAC. Journal of Theoretical Biology, vol. 415, pp 13–19. 2017.

[KHA 10] Khan A, Majid A, Choi T-S., Predicting protein subcellular location: exploiting amino acid based sequence of feature spaces and fusion of diverse classifiers'. Amino Acids Vol. 38, N° 3, pp. 47–35. 2010.

[KIM 10] Kim, H. S. Bat intelligent hunting optimization with application to multiprocessor scheduling (PhD thesis, Department of Electrical Engineering and Computer Science, Case Western Reserve University, Cleveland, OH). 2010

[KIN 00] King R. D., Karwath A., Clare A., and Dehaspe L., "Accurate Prediction of Protein Functional Class from Sequence in the Mycobacterium tuberculosis and Escherichia coli Genomes using Data Mining," Yeast, vol. 17, pp. 283-293, 2000.

[KIN 01] King R. D., Karwath A., Clare A., and Dehaspe L., "The utility of different representations of protein sequence for predicting functional class," Bioinformatics, vol. 17, pp. 445–454, 2001.

[KIR 92] Kira, K. et Rendell, L. A. The feature selection problem : Traditional methods and a new algorithm. In AAAI, pages 129{134, Cambridge, MA, USA. AAAI Press and MIT Press. 1992.

[KOH 97] Kohavi, R. et John, G. H. Wrappers for feature subset selection. Artif. Intell., 97:273-324. 1997.

[KOL 96] Koller, D. et Sahami, M. Toward optimal feature selection. pages 284-292. Morgan Kaufmann. 1996.

[KOM 12] Komarasamy, G., and A. Wahi. An optimized K-means clustering technique using Bat algorithm. European Journal of Scientific Research84:263–273. 2012

[KON 13] König, C., Cruz-Barbosa, R., Alquézar, R., Vellido, A.: SVM-based Classification of Class C GPCRs from Alignment-free Physicochemical Transformations of their Sequences, in A. Petrosino, L. Maddalena, P. Pala (Eds), New Trends in Image Analysis and Processing, Lecture Notes in Computer Science, 8158, Springer Berlin Heidelberg, pp. 336–343, 2013.

[KOT 07] Kotsiantis S. B., "Supervised Machine Learning: A Review of Classification Techniques," Informatica, vol. 31, pp. 249-268, 2007.

[KUB 97] Kubat M. and Matwin S. Addressing the Curse of Imbalanced Training Sets: One Sided Selection. Proceedings of the Fourteenth International Conference on Machine Learning. pp. 179-186, San Francisco, CA, July 8-12 1997.

[KUM 10] Kumari T., Pant B., Pardasani K.R., 'A SVM Model for AAC Based Classification of Class B GPCRs', World Congress of Biomechanics (WCB), Springer Berlin Heidelberg, pp.1607–1610. 2010

[KUM 15] Kumar, R., Srivastava, A., Kumari, B., Kumar, M., Prediction of β-lactamase and its class by Chou's pseudo-amino acid composition and support vector machine'. Journal of Theoretical Biology, vol 365, pp.96–103. 2015

[KUM 17] Kumar, R., Kumari, B., Kumar, M., Prediction of endoplasmic reticulum resident proteins using fragmented amino acid composition and support vector machine'. PeerJ. vol. 4 doi: 10.7717/peerj.3561. 2017

[KUN 99] Kuncheva, L. I. et Jain, L. C. Nearest neighbor classifier : Simultaneous editing and feature selection. 1999

[LAP 02] Lapinsh, M., Prusis, P., Lundstedt, T., and Wikberg, J. E. S. Proteochemometrics modeling of the interaction of amine g-protein coupled receptors with a diverse set of ligands. Molecular Pharmacology 61, 6 (2002), 1465-1475.

[LEE 08] Lee B. J. and Ryu K. H., "Feature Extraction from Protein Sequences and Classification of Enzyme Function," in International Conference on Biomedical Engineering and Informatics, pp. 138-142. 2008.

[LEI 14] Leijôto L.F., Rodrigues T.A.O., Zarate L.E. and Nobre C.N., A Genetic algorithm for the selection of features used in the prediction of protein function. In Proceedings of IEEE 14th International Conference on Bioinformatics and Bioengineering, 2014.

[LES 04] Leslie CS, Eskin E, Cohen A,et al. Mismatch string kernels for discriminative protein classification. Bioinformatics; 20:467–76. 2004

[Li 17] Li J. and Liu H., Challenges of Feature Selection for Big Data Analytics. IEEE Computer Society, vol 17, pp 1541-1672. 2017

[LI 06] Li, Z.R., et al., PROFEAT: a web server for computing structural and

physicochemical features of proteins and peptides from amino acid sequence. Nucleic Acids Res, 2006.

[LI 08] Li, Y. et Guo, L. Tcm-knn scheme for network anomaly detection using feature based optimizations. In Proceedings of the 2008 ACM symposium on Applied computing, SAC '08, pages 2103-2109, New York, NY, USA. ACM. 2008

[LI 10] Li, Z., Zhou, X., Dai, Z., Zou, X. Classification of G-Protein Coupled Receptors based on Support Vector Machine with Maximum Relevance Minimum redundancy and Genetic Algorithm', BMC bioinformatics, vol. 11, issue 1, pp. 325–340. 2010

[LIN 07] Lin H., Li Q-Z., Using Pseudo Amino Acid Composition to Predict Protein Structural Class: Approached by Incorporating 400 Dipeptide Components'. J Comput Chem, vol.28, pp.1463–1466. 2007

[LIN 12] Lin, J. H., C. W. Chou, C. H. Yang, H. L. Tsai, and I. H. Lee. 2012. A bio-inspired optimization algorithm for modeling the dynamics of biological systems. In theThird international conference on innovationsin bio-inspired computing and applications, 206–211. IEEE. 2012.

[LIU 02] Liu, H., Motoda, H. et Yu, L. Feature selection with selective sampling. In Proceedings of the Nineteenth International Conference on Machine Learning, ICML '02, pages 395-402, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc. 2002.

[LUS 01] Luscombe N. M., Greenbaum D., and Gerstein M., "What is bioinformatics? A proposed definition and overview of the field. Methods of Information in Medicine,vol.40,no.4,pp.346–358, 2001.

[MA 08] Ma S. and Huang J., Penalized feature selection and classification in bioinformatics. BRIEFINGS IN BIOINFORMATICS.VOL 9. NO 5. 392-403. 2008

[MAN 13] Manning T., Sleator R. D., and Walsh P. Naturally selecting solutions: The use of genetic algorithms in bioinformatics. Bioengineered 4:266-278. 2013

[MAQ 13] Maqsood H., Khan A., Prediction of Membrane Protein Types Using Pseudo-Amino Acid Composition and Ensemble Classification', International Journal of Computer and Electrical Engineering, Vol. 5, No. 5. 2013

[MAT 02] Mathé C., Sagot M., Schiex T., and Rouzé P. 2002. Current methods of gene prediction, their strengths and weaknesses. Nucleic Acids Research 19:4103-4117.

[MAT 14] V. MATHIVET, L' Intelligence Artificielle pour les développeurs Concepts et implémentations en C#. Editions ENI - Décembre. ISBN : 978-2-7460-9215-0. 2014

[MOO 04] Moore J. H., Hahn L. W., Ritchie M. D., Thornton, T. A. and White, B. C. Routine discovery of complex genetic models using genetic algorithms. Applied Soft Computing 4:79-86. doi:10.1016/j.asoc.2003.08.003. 2004.

[MOR 50] Moran, P.A., Notes on continuous stochastic phenomena. Biometrika, 1950. 37(1-2): p. 17-23.

[MUN 17] Munoz S., Guerrero F.D, Kellogg A., Heekin A.M, Leung M-Y, 'Bioinformatic prediction of G protein Coupled receptor encoding sequences from the transcriptome of the foreleg, including the Haller's organ, of the cattle tick, Rhipicephalus australis'. PLoSONE, vol 12. 2017

[MUS 12] Musikapun, P., and P. Pongcharoen. Solving multi-stage multi-machine multi-product scheduling problem using Bat algorithm. In Second international conference on management and artificial intelligence (IPEDR), 35:98–102. Singapore: IACSIT Press, 2012.

[NAK 17] Nakano F. K., Mastelini S. M., Barbon S., Cerri R., 'Stacking Methods for Hierarchical Classification', In Proceedings of 16th IEEE International Conference on Machine Learning and Applications. 2017

[NAR 77] Narendra, P. M. et Fukunaga, K. A branch and bound algorithm for feature subset selection. IEEE Trans. Comput., 26: 917-922. 1977.

[NAV 12] Naveed, M., Khan, A.U.: GPCR-Mpredictor: Multi-level Prediction of G ProteinCoupled Receptors using Genetic Ensemble, Amino Acids, vol. 42, issue 5, pp. 1809–1823, 2012.

[NAY 18] Nayyar, A., Garg, S., Gupta, D., & Khanna, A. Evolutionary computation: theory and algorithms. In Advances in Swarm Intelligence for Optimizing Problems in Computer Science (pp. 1-26). Chapman and Hall/CRC. 2018

[NEM 09] Nemati S., Basiri M.E, Ghasem-Aghaee N., Aghdam M.H, 'A novel ACO–GA hybrid algorithm for feature selection in protein function prediction'. Expert Systems with Applications, Vol. 36. pp.12086–12094. 2009.

[NGO 16] Ngo T., Kufareva I., Coleman J.LJ., Graham R. M, Abagyan R. and Smith N. J. Identifying ligands at orphan GPCRs: current status using structure-based approaches. British Journal of Pharmacology. Vol. 173, pp.2934–2951. 2016

[NIZ 11] Nizam A., Ravi J., and Subburaya K., Cyclic Genetic Algorithm for Multiple Sequence Alignment. International Journal of Research and Reviews in Electrical and Computer Engineering (IJRRECE) Vol. 1, No. 2, June 2011.

[NOT 97] Notredame C., O'Brien E. A., and Higgins D. G. 1997. RAGA: RNA sequence alignment by genetic algorithm. Nucleic Acids Research 25:4570-4580

[ONG 07] Ong, S.A., et al., Efficacy of different protein descriptors in predicting protein functional families. BMC Bioinformatics, 2007. 8: p. 300.

[OOI 03] Ooi C.H., and Tan, P. Genetic algorithms applied to multi-class prediction for the analysis of gene expression data. Bioinformatics 19:37-44. 2003

[PAN 06] Pandey G., Kumar V., and Steinbach H., "Computational Approaches for Protein Function Prediction: A Survey," TR 06-028, 2006.

[PAP 04] Papasaikas, P.K., Bagos, P.G., Litou, Z.I., Promponas, V. J., Hamodrakas, S.J.: PRED-GPCR: GPCR Recognition and Family Classification Server, Nucleic Acids Research, vol. 32 (suppl 2), pp. W380–W382, 2004.

[PAR 03] Park KJ, Kanehisa M. Prediction of protein subcellular locations by support vector machines using compositions of amino acid and amino acid pairs. Bioinformatics; 19:1656–63. 2003

[PAT 95] Pattern Recognition Letters, (1995).Vol. 16, pp. 801-808.

[PEA 95] Pearson, K., Note on Regression and Inheritance in the Case of Two Parents. Proceedings of the Royal Society of London, 1895. 58: p. 240-242.

[PEA 04] Pearson W., "Finding protein and nucleotide similarities with FASTA, Current Protocols in Bioinformatics, chapter 3, unit3.9, 2004

[PED 96] Pedersen J.T., Moult J, Genetic algorithms for protein structure prediction. Current Opinion in Structural Biology, Vol. 6, pp. 227–231. 1996

[PED 97] Pedersen J. T. and Moult J. Protein folding simulations with genetic algorithms and a detailed molecular description. Journal of Molecular Biology 269:240-259. 1997

[PEI 98] Pei, M., Punch, W.F., and Goodman, E.D. "Feature Extraction Using Genetic Algorithms", Proceeding ofInternational Symposium on Intelligent Data Engineering and Learning'98 (IDEAL'98), Hong Kong, Oct. 1998

[PEN 10] Peng, Z.L., Yang, J.Y., Chen, X.: An Improved Classification of G-Protein Coupled Receptors using Sequence-Derived Features, BioMed Central Bioinformatics, vol. 11, 2010

[PET 93] Petrilli, P., Classification of protein sequences by their dipeptide composition. Comput Appl Biosci, 1993. 9(2): p. 205-9

[PIY 02] Piyathilake C, Johannig GL. Cellular vitamins, DNA methylation and cancer risk. Am Soc Nutri Sci; 132:2340S–2344S. 2002

[PRA 14] Pramanik, S., & Setua, S. K. A steady state Genetic Algorithm for Multiple Sequence Alignment. 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI). 2014.

[RAH 11] Rahman S.A., Bakar A.A., Hussein Z.A.M., Feature Selection and Classification of Protein Subfamilies Using Rough Sets, 2009 International Conference on Electrical Engineering and Informatics. 5-7 August 2009, Selangor, Malaysia. 2009.

[RAO 08] 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, pp. 495–503, 2008.

[REH 13] Rehman ZU., Mirza MT., Khan A, Xhaard H., Predicting G-Protein-Coupled

Receptors Families Using Different Physiochemical Properties and Pseudo Amino Acid Composition‘. Methods in Enzymology, Vol. 522, pp. 61-79. 2013

[REH 11] Rehman, Z., Khan, A. G-Protein-Coupled Receptor Prediction using Pseudo Amino-Acid Composition and Multiscale Energy Representation of Different Physiochemical Properties, Analytical Biochemistry, vol. 412, pp. 173–182, 2011.

[REN 10] Ren, X.-M. and J.-F. Xia, Prediction of Protein-Protein Interaction Sites by Using Autocorrelation Descriptor and Support Vector Machine, in Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence, D.-S. Huang, et al., Editors. 2010, Springer Berlin Heidelberg. p. 76-82.

[REY 05] Rey S, Gardy JL, Brinkman FSL. Assessing the precision of high-throughput computational and laboratory approaches for the genome-wide identification of protein subcellular localization in bacteria. BMC Genomics ;6:162. 2005

[SAE 07] Saeys Y., Inza I., Larranaga P., ‘A review of feature selection techniques in bioinformatics’, BIOINFORMATICS, vol. 23 no. 19, pp.2507–2517. 2007.

[SAI 12] Saidi R., Aridhi S., Maddour M.: Feature extraction in protein sequences classification : a new stability measure. ACM-BCB '12, Orlando, FL, USA, October 2012.

[SAN 98] Sandberg, M., Eriksson, L., Jonsson, J., Sjostrom, M., and Wold, S. New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. Journal of Medical Chemistry 41 (1998), 2481-2491.

[SAN 18] Santos B.C., Nobre C.N., Zarate L. E., Multi-objective genetic algorithm for feature selection in a protein function prediction context, 2018 IEEE Congress on Evolutionary Computation (CEC), Rio de Janeiro, Brazil. 2018.

[SAN 19] Santos B.C., Rodrigues M.W., Pinto L.N.C., Nobre C.N., Zárate L. E., Feature selection with genetic algorithm for protein function prediction. In proceedings IEEE International Conference on Systems, Man and Cybernetics (SMC). Italy. 2019
[SAS 02] Sastry, K., and Goldberg, D.E. (2002). Genetic algorithms, efficiency enhancement, and deciding well with differing fitness variances. Proceedings of the Genetic and Evolutionary Computation Conference, 528–535.

[SCH 94] Schneider, G. and P. Wrede, The rational design of amino acid sequences by artificial neural networks and simulated molecular evolution: de novo design of an idealized leader peptidase cleavage site. Biophys J, 1994. 66(2 Pt 1): p. 335-44.

[SEC 07] Secker, A., Davies, M., Freitas, A., Timmis, J., Mendao, M., and Flower, D. An experimental comparison of classification algorithms for the hierarchical prediction of protein function. Expert Update (the BCSSGAI Magazine) 9, 3 (2007), 17-22.

[SEC 09] Secker, A., Davies, M.N., Freitas, A.A., Timmis, J., Clark, E., Flower, D.R. An Artificial Immune System for Clustering Amino Acids in the Context of Protein Function Classification‘, Journal of Mathematical Modelling and Algorithms, vol.8, issue 2, pp.103–123. 2009

[SEC 10] Secker A., Davies, M.N., Freitas, A.A., Clark,E. & Timmis, J., Flower, D.R., Hierarchical classification of G-Protein-Coupled Receptors with data-driven selection of attributes and classifiers', International Journal of Data Mining and Bioinformatics, vol.4, pp.191-210. 2010.

[SEL 05] Selzer, P., & Ertl, P.: Identification and Classification of GPCR Ligands using Self-Organizing Neural Networks, QSAR & Combinatorial Science, vol. 24, issue 2, pp. 270–276, 2005.

[SET 13] Settouti N., Hafa A., Approche Filtre pour la sélection des gènes pertinents des données biopuces du Cancer du Côlon. 2013.

[SHE 08] Shen, H.B. and K.C. Chou, PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition. Anal Biochem, 2008. 373(2): p. 386-8.

[SHR 10] Shrivastava, S, Pardasani, K. R., Malik, M. M. SVM Model for Identification of human GPCRs'. Journal of computing, Vol.2, issue 2. 2010

[SIE 89] Siedlecki, W., Sklansky J., A note on genetic algorithms for large-scale feature selection, Pattern Recognition Letters, Vol. 10, Page 335-347, (1989).

[SIK 07] Riyaz Sikora, Selwyn Piramuthu, Framework for efficient feature selection in genetic algorithm based data mining, European Journal of Operational Research 180 (2007) 723–737.

[SIL 11] Silla Jr C.N., Freitas A.A., Selecting Different Protein Representations and Classification Algorithms in Hierarchical Protein Function Prediction', Intelligent Data Analysis, vol. 15, pp.979–999. 2011.

[SÏR 10] Sîrbu A., Ruskin, H. J., and Crane M. 2010. Comparison of evolutionary algorithms in gene regulatory network model inference. BMC Bioinformatics 11:59. http://www.biomedcentral.com/1471-2105/11/59

[SOM 04] Somol, P., Pudil, P. et Kittler, J. Fast branch & bound algorithms for optimal feature selection. IEEE Pattern Analysis and Machine Intelligence, 26: 900-912. 2004.

[SUR 06] Surgand J-S., Développement de nouvelles méthodes bioinformatiques pour l'étude des récepteurs couplés aux protéines G. Thèse présentée pour obtenir le grade de Docteur de l'Université Louis Pasteur Strasbourg I. 2006

[SPE 90] Spears, W. M. and De Jong K. A. 1990. Using genetic algorithms for supervised concept learning. Proceedings of the 2nd International IEEE Conference on Tools for Artificial Intelligence. pp. 335-341. Herndon, VA, USA.

[TAH 13] Taherian, H., I. N. Kakhki, and M. R. Aghaebrahimi. Application of an improved SVR based bat algorithm for short-term price forecasting in the Iranian pay-as-bid electricity market. In 3rd International conference on computer and knowledge engineering, 161–166. IEEE. 2013

[TAH 15] Tahmasebipour K. and Houghten S. Disease-gene association using a genetic algorithm. Conference paper presented at 2015 IEEE Conference on Computational

Intelligence in Bioinformatics and Computational Biology, held at Niagara Falls, Canada in August 2015. doi:10.1109/CIBCB.2015.7300331. 2015
[THA 16] Tharwat, A.; Hassanien, A.E.; Elnaghi, B.E. A BA-based algorithm for parameter optimization of Support Vector Machine. Pattern Recognit. Lett, 93, 13–22. 2016

[TIB 04] Tibshirani R, Hastie T, Narasimhan B, Soltys S, et al. Sample classification from protein mass spectrometry, by 'peak probability contrasts'. Bioinformatics; 20:3034–44. 2004

[TON 08] Tong, J.C. and Tammi, M.T., Prediction of protein allergenicity using local descriptions of amino acid sequence, Frontiers in Bioscience, Vol. 13, pp.6072–6078. 2008.

[TUP 17] Tupe KA, Wakchaure Prof.MA., Big Data Feature Selection Data Stream Mining. International Journal Of Engineering And Computer Science, Vol. 6 Issue 7, pp. 22041-22044. 2017

[TUS 01] Tusher, V. G., Tibshirani, R. et Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. 98. 2001.
[UMA 07] Umar S. and Golan Y., "Enzyme function prediction with interpretable models," Computational Systems Biology, pp. 1-33, 2007.

[UMA 17] Umarov R. K. and Solovyev V. V. Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks. PLoS ONE, 12(2): e0171410. http://doi.org/10.1371/journal.pone.0171410. 2017

[WAN 06] Wang H, Fu Y, Sun R, et al. A SVM score for more sensitive and reliable peptide identification via tandem mass spectrometry. Pac Symp Biocomput;11:303–14. 2006.

[WAN 07] Wang X., Yang J., Teng X., Xia W., and Jensen R., "Feature selection based on rough sets and particle swarm optimization," Pattern Recognition Letters, vol. 28, 2007.

[WAN 19] Wang Y., Wang P., Zhang J., Cui Z., Cai X., Zhang W. and Chen J., A Novel Bat Algorithm with Multiple Strategies Coupling for Numerical Optimization. Mathematics, 7, 135; doi:10.3390/math7020135. 2019

[WEI 16] Wei L., Xing P., Shi G., Ji Z., and Zou Q., Fast prediction of protein methylation sites using a sequence-based feature selection technique. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2016

[WES 01] West M, Blanchette C, Dressmna H, et al. Predicting the clinical status of human breast cancer by using gene expression profiles. Proc Natl Acad Sci USA; 98:11462–67. 2001

[WES 04] Weston J, Leslie C, Zhou D,et al. Semi-supervised protein classification using cluster kernels.Adv Neural Inf Process Syst;16:595–602. 2004

[WU 13] Wu G., Pedrycz W., Li H., Qiu D., Ma M., and Liu J. 2013. Complexity Reduction in the Use of Evolutionary Algorithms to Function Optimization: A Variable Reduction Strategy. The Scientific World vol. 2013, Article ID 172193.

[XIA 10] Xiao, X., Qiu, W.R.: Using Adaptive K-Nearest Neighbor Algorithm and Cellular Automata Images to Predicting G-Protein-Coupled Receptor Classes, Interdisciplinary Sciences: Computational Life Sciences, vol. 2, issue 2, pp. 180–184, 2010.
[XIA 11] Xiao, X., Wanga, P., Chou, K.C.: GPCR-2L: Predicting G Protein-Coupled Receptors and their Types by Hybridizing Two Different Modes of Pseudo Amino Acid Compositions, Molecular BioSystems, vol. 7, pp. 911–919, 2011.

[YAN 04] Yang H, Mukomel Y, Fink E. Diagnosis of ovarian cancer based on mass spectra of blood samples. In: Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics. The Hague, The Netherlands; 3444–50. 2004

[YAN 10] Yang, X.S. A New Metaheuristic Bat-Inspired Algorithm. Comput. Knowl. Technol. 2010, 28, 65–74.

[YAN 12a] Yang, X. Bat algorithm for multi-objective optimization. Int. J. Bio-Inspired Comput. 2012, 3, 267–274.

[YAN 12b] Yang, X. S., and A. H. Gandomi. 2012. Bat algorithm: A novel approach for global engineering optimization.Engineering Computations29:464–483.

[YAN 13] Yang, X.S., Karamanoglu, M.: Swarm Intelligence and Bio-Inspired Computation: An Overview, in.Yang, X.S., Cui, Z., Xiao, R., Gandomi, A.H., Karamanoglu, M. (Eds.), Swarm intelligence and bio-inspired computation: theory and applications, Elsevier, pp. 3–23, 2013.

[YAN 16] Yang C., Moi S., Lin Y., and Chuang L. Genetic algorithm combined with a local search method for identifying susceptibility genes. JAISCR 6:203-212. doi:10.1515/jaiscr-2016-0015. 2016.

[YU 06] Yu CS, Chen YC, Lu CH, et al. Prediction of protein subcellular localization. Proteins; 64:643–51. 2006

[ZEK 11] Zekri, M., Alem, K., Souici-Meslati, L.: Identification Methods of G Protein Coupled Receptors, International Journal of Knowledge Discovery in Bioinformatics, Vol. 2, No. 4, pp. 35–52, 2011.

[ZEK 15] Zekri M, Approches Bio-inspirées pour la Fouille de Données en Bioinformatique. THESE Présentée en vue de l'obtention du diplôme de Doctorat 3ème Cycle. Annaba, Algérie. 2015.

[ZHA 12] Zhang Z., Wu J., Yu J., Xiao J. 'A brief review on the evolution of GPCR: conservation and diversification'. Open Journal of Genetics, Vol. 2, pp.11-17. 2012

[ZUK 04] Zukiel R, Nowak S, Barciszewska A, et al. A simple epigenetic method for the diagnosis and classification of brain tumors.Mol Cancer Res; 2:196–202. 2004 Protein Engineering, Design & Selection, vol.19, issue 11, pp. 511–516. 2006.

Webographie

[1] https://babel.cegep-ste-foy.qc.ca/profs/gbourbonnais/pascal/nya/genetique/notesadn/adn4.htm

[2] https://www.sareptatherapeutics.ch/fr/applications-therapeutiques

[3] http://biochimej.univ-angers.fr/Page2/COURS/7RelStructFonction/2Biochimie/5Signalisation/4RCPGetProteinesG/1RCPGetProtG.htm

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

[5] http://Endocytose-des-r%C3%A9ception-coupl%C3%A9s-aux-prot%C3%A9ines-G-_-m%C3%A9decine_sciences%20.html

[6] https://palli-science.com

[7] https://pharmacomedicale.org/pharmacologie/pharmacologie-medicale-vue-d-ensemble/32-differents-types-de-structure-sur-lesquelles-agissent-les-medicaments/58-recepteurs-couples-aux-proteines-g

[8] http://www.123bio.net/cours/liaison/partie45.html

[9] https://www.gpcrdb.org/

[10] https://www.cs.waikato.ac.nz/ml/weka.

[11] https://cran.r-project.org/