REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université Mohamed Khider – BISKRA
Faculté des Sciences Exactes, des Sciences de la Nature et de la Vie
Département d'informatique

N° d'ordre : SIOD8 /M2/2018

Mémoire

Présenté pour obtenir le diplôme de master académique en

Informatique

Parcours **Système Informatique décisionnelle et optimisation**

Classification des Données Biopuces par la Méthode K-PPV

Par :

RECHACHI BOUTHAINA

Soutenu le 25/06/2018, devant le jury composé de :

Mme. Belounnar Saliha	Encadreur	Université de Biskra
Mr. Bendahmane Toufik	Président	Université de Biskra
Mr. Rahmani Salima	Membre	Université de Biskra

Remerciements

Nous remercions le bon Dieu, tout puissant, de nous avoir donné la force pour suivre, ainsi que l'audace pour dépasser toutes les difficultés.

On tient à remercier sincèrement Mme SALIHA BELLOUNAR, qui ont toujours montré à l'écoute et très disponible tout au long de la réalisation de ce mémoire.

Les jurys pour leurs efforts et leur soin apporté à mon travail.
Aux enseignants de notre université et département informatique.

Enfin, nous adresse nos plus sincères remerciements à tous nos proches et amis, qui nous ont toujours soutenue et encouragée au cours de la réalisation de ce mémoire.
Merci à tous et à toutes

Dédicaces

Je dédie ce modeste travail, aux deux êtres les plus chers à mon cœur auxquels je dois mon existence : Mon père et ma mère; vous qui étaient toujours à mes côtés pour me soutenir et m'encourager à me battre

sans jamais m'arrêter à mi-chemin; que dieu vous protège.
A ma grande sœur : Dr.Rechachi Zerarka Miled Zohra
Et autres sœurs

Et mon frère Rechachi Hichem Arbi « Que Dieu bénisse son âme »
Et autre frères

Et 'Elhani Djenaihi mon amis qui m'ont aidé durant cette application.

A tous mes amis de la promotion 2eme année master informatique 2017/2018.

Rechachi Bouthaina

RÉSUMÉ

La bio-informatique c'est un domaine de recherche qui stocker, traiter et analyser de grandes quantités de données des phénomènes biologiques, au moyen de méthodes informatiques, afin de créer de nouvelles connaissances en biologie.

La biopuce est l'une des techniques modernes qui nous aide à faire le diagnostic. L'analyse des données biopuce comprend quatre étapes : l'acquisition des données biopuce , le prétraitement des données biopuce , la sélection des gènes l'évaluation des modèles de classification.l'acquisition de données génétiques est un processus biomédical, et les autres étapes sont des processus d'exploration de données.

Ce travail vise à étudier la classification les données de bio-puces avec la méthode de classification K plus proche voisins (k-ppv) pour apprendre des modèles de décision qui permettent de prédire le comportement des exemples futurs.

Mots clés : Bio-informatique , Biopuce , Classification des données, K-ppv.

ABSTRACT

Bioinformatics is a domain. for research that store, process and analyze big data of biological phenomena, using computer methods, to create new knowledge in biology.

Microarray is a modern technique that helps us to make the diagnosis.The microarray The analysis of genetic data includes four steps: the acquisition of microarray data, the pre-processing of biochip data, the selection of genes and the evaluation of classification models.Genetic data acquisition is a biomedical process, and the other steps are data mining processes.

This work aims at investigating the classification of k (NN) method to predict future behavior.

Keywords: Bioinformatics, Microarray, Data classification, K-NN.

TABLE DES MATIÈRES

TABLE DES FIGURES

LISTE DES TABLEAUX

INTRODUCTION GÉNÉRALE

Le domaine de la bioinformatique c'est un domaine de recherche qui analyse et interprète des données biologiques, au moyen de méthodes informatiques, afin de créer de nouvelles connaissances en biologie.

La bio-informatique c'est donc un domaine de recherche qui stocker, traiter et analyser de grandes quantités de données des phénomènes biologiques, au moyen de méthodes informatiques, afin de créer de nouvelles connaissances en biologie.

La technologie des puces à ADN ou biopuces, connaît à l'heure actuelle un essor exceptionnel et suscite un formidable intérêt dans la communauté scientifique. Cette technologie a été développée au début des années 1990 et permet la mesure simultanée des niveaux d'expression de plusieurs milliers de gènes.

La technologie des puces à ADN permet de différencier des tissus tumoraux et des tissus sains à partir de la mesure simultanée d'un grand nombre de gènes au sein d'un échantillon biologique. Pour cette tâche de classification, on dispose d'un faible nombre d'échantillons alors que chaque échantillon est décrit par un très grand nombre de gènes.

technologie puce à ADN sont observées et analysées sous différentes conditions expérimentales. Ces données obtenues sont généralement analysées pour des objectifs divers et spécifiques aux maladies. Elles peuvent être utilisées pour inférer les gènes liés à un cancer, afin d'identifier les différents cancers sur la base de ces gènes.

Cela a fait l'objet de recherches algorithmiques très actives dans ce que l'on appelle la classification des données biopuce.

L'extraction de connaissances à partir de ces données peut se faire par l'utilisation des techniques d'apprentissage automatique.les méthodes de classification supervisée sont les plus utilisées pour classifier les puces à ADN.

Le traitement de ces données nécessite donc de réduire le nombre de gènes pour proposer un sous-ensemble de gènes pertinents et de construire un classifieur prédisant le type de tumeur qui caractérise un échantillon cellulaire. Il s'agit d'un problème de sélection d'attributs.

La sélection d'attributs est un problème complexe qui a déjà été largement étudié, mais les dimensions des données des biopuces nécessitent des approches spécifiques (plusieurs milliers de gènes).

L'apprentissage automatique (Machine Learning en anglais) désigne le concept selon lequel une machine arrive à déduire le raisonnement automatique des règles à partir d'un ensemble d'exemples. L'objectif est de résoudre des problèmes relativement complexes,provenant du monde réel, la machine doit apprendre à produire la sortie désirée.

Ce mémoire comporte quatre chapitres et est organisé de la manière suivante :

- Le premier chapitre nous abordons définir le bio-informatique et quelques terminologie biologie afin de dépasser à la comprendre le principe des biopuces (Etapes d'analyse,plateformes,banques et outils génomiques).
- Dans le troisième chapitre, nous entamons les différentes étapes d'analyse des puces à ADN et comment présenter cette données ? Et comment faire la sélection des gènes par méthodes statistiques après la présentation des cette données ?
- Dans le troisième chapitre sera consacré à la présentation de la définition de classification et l'apprentissage et le principe de la méthode utilisée dans ce travail .
- Le quatrième chapitre sera consacré à la présentation de l'implémentation de notre système, les différents composants nécessaires à son fonctionnement, l'environnement de développement et les structures des données utilisée.
- Enfin, nous achevons ce manuscrit par une conclusion et perspectives.

CHAPITRE

1

LA BIOINFORMATIQUE ET LES BIO-PUCES

1.1 Introduction

Le récent flot de données issues des séquences du génome et de la génomique fonctionnelle a donné naissance à un nouveau champ, la bioinformatique, qui combine des éléments de biologie et d'informatique.

Nous proposons ici une définition de ce nouveau domaine et passons en revue certaines des recherches en cours, en particulier en ce qui concerne les systèmes de régulation de la transcription.

Le terme bioinformatique est apparu pour la première fois dans une publication de Paulien Hogeweg et Ben Hesper, en référence à l'étude des processus d'information dans la communauté scientifique pour la compréhension des phénomènes biologiques.

La technologie des puces à ADN ou biopuces, connaît à l'heure actuelle un essor exceptionnel et suscite un formidable intérêt dans la communaut é scientique.

Il y a quatre étapes Les différentes phases d'une analyse les puces à ADN (Réparation des cibles, hybridation, acquisition et analyse des images, normalisation et présentation des données de biopuces)

Dans le premier chapitre nous allons présenter des la définition de terme « Bioinformatique »avec les notions et les terminologies élémentaires en biologie qui sont les bases de notre sujet de recherche dans ce mémoire pour comprendre comment marchent les biopuces à ADN.

1.2 Bioinformatique et génomique

1.2.1 Bioinformatique

La bioinformatique, une nouvelle science en pleine évolution qui attire un effort rapidement croissant de recherche et de biotechnologie [12] . bioinformatique est apparu pour la première fois dans une publication de Paulien Hogeweg et Ben Hesper, en référence à l'étude des processus d'information dans la communauté scientifique pour la compréhension des phénomènes biologiques [6] .

La plupart des définitions de la bio-informatique suggèrent l'interaction entre la biologie, les technologies de l'information et les sciences informatiques (les mathématiques). D'après Claverie, « la bioinformatique est la discipline de l'analyse de l'information biologique, en majorité sous la forme de séquences génétiques et de structures de protéines [33] .

1.2.1.1 Principe de bioinformatique

La bioinformatique sert donc à stocker, traiter et analyser de grandes quantités de données de biologie. Le but est de mieux comprendre et mieux connaître les phénomènes et processus biologiques. Grâce à ces nouvelles connaissances.

1.2.1.2 Comment ça marche ?

La bioinformatique fournit des bases de données centrales, accessibles mondialement, qui permettent aux scientifiques de présenter, rechercher et analyser de l'information. Elle propose des logiciels d'analyse de données pour les études de données et les comparaisons et fournit des outils pour la modélisation, la visualisation, l'exploration et l'interprétation des données. les chercheurs ont la possibilité de faire de nouvelles découvertes scientifiques. Des découvertes qui peuvent par exemple améliorer la qualité de vie de personnes malades grâce à la mise en place de nouveaux traitements médicaux plus efficaces [28] .

1.2.2 Terminologie génomique

1.2.2.1 La cellule

C'est la plus petite unité structurale et fonctionnelle de tous les êtres vivants. Il existe des milliers de type de cellules différents par leur forme,

leur taille, leur fonction et leur comportement. Chez les organismes dits procaryotes tels que les bactéries, le matériel génétique n'est pas contenu dans un noyau mais est libre dans tout le cytoplasme de la cellule. Par contre, les organismes complexes comme les eucaryotes qui sont pluricellulaires, l'information génétique est localisée dans un noyau. L'homme, les animaux et les plantes sont des organismes eucaryotes. La plupart de leurs cellules sont capables de grossir et se diviser. Elles sont dotées d'un métabolisme, c'est à dire la capacité d'importer des nutriments et les convertir en molécules et en énergie [19] [27] .

1.2.2.2 Acide désoxyribonucléique (ADN)

L'acide désoxyribonucléique [36] (A.D.N) est une molécule présente dans le noyau de la cellule qui joue un rôle central dans la vie cellulaire. Il renferme l'ensemble des informations nécessaires au développement et au fonctionnement d'un organisme. Cette macromolécule a une structure en double hélice constituée de deux brins antiparallèles. Un brin simple est un polymère linéaire constitué de 4 nucléotides. Un nucléotide comprend une des bases : adénosine (A), cytosine (C), guanine (G), ou thymine (T). Les couples A-T et G-C sont appelés bases complémentaires par lesquelles les deux brins vont s'associer [19] .

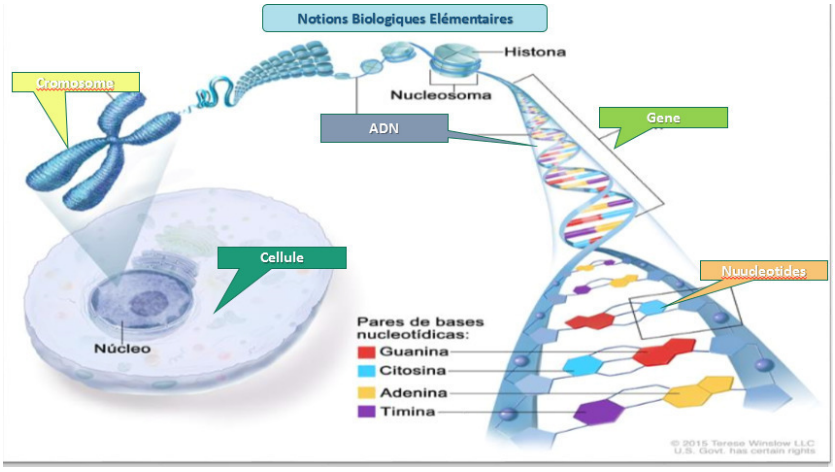


FIGURE 1.1 – Structure d'une molécule d'ADN

1.2.2.3 ARN messager (ARNm)

L'acide ribonucléique messager, ARN messager ou ARNm est une copie transitoire d'une portion de l'ADN correspondant à un ou plusieurs gènes.

L'ARNm est utilisé comme intermédiaire par les cellules pour la synthèse des protéines. Le concept d'ARN messager a été émis puis démontré par Jacques Monod, François Jacob et leurs collaborateurs en 1960.
L'ARNm est une copie simple brin linéaire de l'ADN composée d'ARN, qui comprend la région codant une protéine. La transcription des ARNm et leur traduction sont des processus qui sont l'objet de contrôles cellulaires importants et permettent à la cellule de réguler l'expression des différentes protéines dont elle a besoin pour son métabolisme [31] .

FIGURE 1.2 – Structure d'une molécule d'ARNm

1.2.2.4 Transcriptome

Le gène, unité de base de stockage de l'information génétique, est une petite séquence d'ADN. Il y a environ 6000 gènes chez les levures par exemple et 30000 chez l'homme. L'ensemble du matériel génétique d'un individu ou d'une espèce encodé dans son ADN est appelé alors son génome [19] .

En fonction de leurs besoins, les cellules utilisent à un instant donné une partie des gènes pour réaliser la synthèse des protéines nécessaires aux grandes fonctions cellulaires. Le passage du gène à la protéine se fait en deux grandes parties, la transcription et la traduction, à l'aide d'un agent essentiel l'ARNm, dit ARN messager. le gène est transcrit (synthèse de l'ARNm) puis l'ARNm est conduit hors du noyau dans le cytoplasme ou il va servir de matrice pour la synthèse des protéines pour la traduction.

De manière générale, pouvoir comparer le transcriptome de différents types cellulaires, dans différentes conditions, ou pouvoir analyser l'ensemble du transcriptome d'une cellule à plusieurs phases de son cycle cellulaire ou dans diverses conditions pathologiques, doit permettre de mieux comprendre le fonctionnement cellulaire sur le plan fondamental.

Les méthodes d'analyse du transcriptome les plus utilisées reposent sur la technologie des puces à ADN car elles permettent de visualiser simultanément le niveau d'expression de plusieurs milliers de gènes dans un contexte physiologique ou pathologique particulier [25] .

— **La transcription :** est, en biologie moléculaire, un mécanisme qui permet de « recopier » les données des gènes, ce qui permet leur utilisation pour créer de la matière biologique en assemblant des des acides animés en protéines selon le code génétique. Formation d'un brin complémentaire(ARNm) à la portion du brin original d'ADN codant pour une protéine.

— **La traduction :** est l'étape de synthèse des protéines en utilisant l'ARNm qui est conduit hors du noyau dans le cytoplasme ou il va servir de matrice pour la synthèse des protéines.

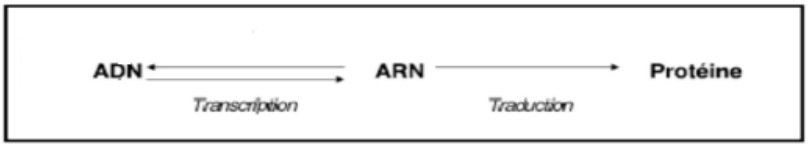


FIGURE 1.3 – processus de transcriptome

1.2.3 La méthode P C R

1.2.3.1 Définition

La PCR (Polymerase Chain Reaction) [1] est une technique d'amplification génétique in vitro qui a été conçue au début des années 80 par un chercheur américain, Kary Mullis, travaillant au sein d'une firme biotechnologique californienne. Cette technique, qui a révolutionné les approches expérimentales en biologie moléculaire, a été publiée pour la première fois en 1985 dans la revue scientifique "Science". En 1993, Kary Mullis recevra le Prix Nobel de Chimie pour sa découverte de la PCR.
La PCR (Polymerase Chain Reaction) est une technique d'amplification d'ADN in vitro. Elle permet d'obtenir un très grand nombre de copies d'une séquence d'ADN choisie [27] .

1.2.3.2 Principe de la PCR

La PCR [32] est basée sur le mécanisme de réplication de l'ADN in vivo : l'ADN bicaténaire est déroulé en ADN monocaténaire, puis dupliqué et réenroulé, selon des cycles répétitifs comprenant les trois étapes suivantes :

Dénaturation de l'ADN par fusion à haute température pour convertir l'ADN bicaténaire en ADN monocaténaire. Cette étape est réalisée à une température comprise entre 93 et 96C.

Hybridation à l'ADN cible de deux oligonucléotides utilisés comme amorces. Cette hybridation a lieu à une température comprise entre 55 et 65C.

Extension de la chaine d'ADN par addition de nucléotides à partir des amorces en utilisant l'ADN polymérase (2) comme catalyseur en présence d'ions Mg2+. La température optimale de travail de l'ADN polymérase est de 72C [27] [19] .

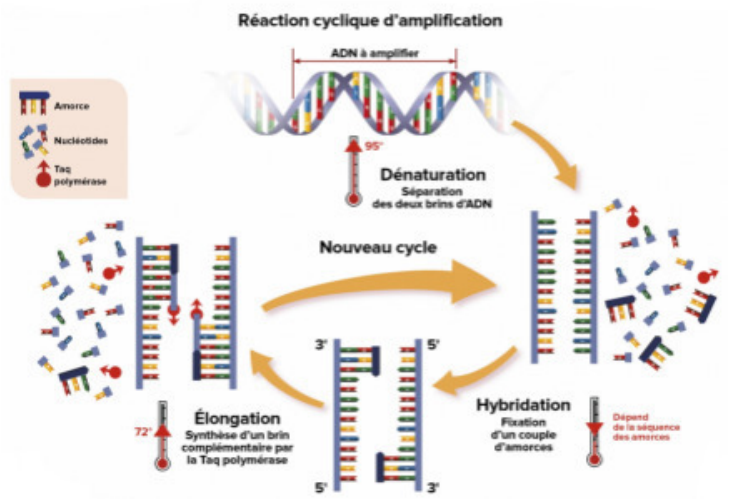


FIGURE 1.4 – L'amplification de l'ADN dans la PCR.

1.3 La technologie puces à ADN

Les puces à ADN , également appelées biopuces, Une puce ADN (appelée DNA microarray en anglais) [21] ou encore puces à gènes, permettent de mesurer les niveaux d'expression simultanés de plusieurs dizaines de milliers de gènes dans un prélèvement. Cette technologie a été publiée pour la première fois en 1995 (Schena et al., 1995) [13] . les puces à ADN jouent aujourd'hui un

rôle prépondérant, tant par leur relative simplicité de mise en œuvre que par leurs nombreux champs d'application. Elles permettent l'analyse simultanée de grand nombre des gènes dans un échantillon biologique sain ou malade [5] .

Cependant, cette technologie génère une grande diversité de données qui implique (les bases de données et les banques de données, s.d.) un important travail de bio-informatique. Aussi, un grand nombre de techniques liées à l'informatique sont nécessaires à l'analyse des données issues de cette technologie analyse d'images, stockage et gestion des informations, techniques de normalisation, analyses statistiques, représentation graphiques, techniques d'extraction de connaissances [25]

Dans cette partie, nous descrivons d'abord les détails de technologie puces a ADN, qui constituent la plupart des données que nous avons utilisées.

1.3.1 Historique

Les premières puces à ADN sont apparues en 1993, mais leur concept date de 1987 (cité dans Bellis et Casellas, 1997). Suite logique aux anciennes méthodes de northern blotting (Alwine et al., 1977) et d'expression différentielle (Liang et Pardee, 1992), la technologie des puces à ADN est basée sur le principe d'hybridation développé par Southern (1974). Ce principe stipule que deux fragments d'acides nucléiques complémentaires peuvent s'associer et se dissocier de façon réversible sous l'action de la chaleur et de la concentration saline du milieu. Il s'agit simplement d'une miniaturisation du système classique de reverse dot blot (Lennon et Lehrach, 1991) qui a vu le jour grâce à une technologie pluridisciplinaire intégrant l'électronique (techniques de dépôt), la chimie (préparation des lames et greffes des sondes oligonucléotidiques ou syn-thèse in situ), l'analyse d'images (acquisition des données) et l'informatique (interprétation des données). Depuis leur apparition, les puces à ADN suscitent un intérêt inversement proportionnel à leur taille, avec pour preuve l'explosion du nombre de publications qui leur sont dédiées depuis 2001 [33] .

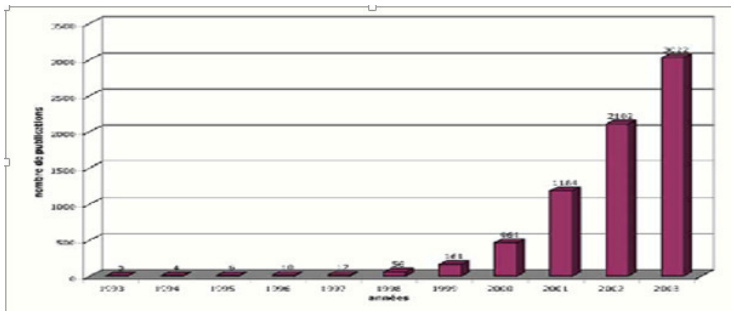


FIGURE 1.5 – Nombre de publications concernant les puces à ADN de 1994 à 2004

Historiquement les macroarrays, les microarrays et les « véritables » puces à ADN correspondent à trois méthodes différentes d'analyse (Lagoda et Regad, 2000). Les macroarrays utilisaient des clones d'ADN complémentaire (ADNc) disposés sur des membranes de nylon (avec un espacement de l' ordre du millimètre) en association avec des cibles radioactives. Les microarrays, plus miniaturisés, comportaient quelques milliers de gènes représentés par des produits PCR déposés tous les 200 à 400 microns sur une lame de verre et des cibles marquées par fluorescence. Enfin, les « véritables » puces à ADN associaient à chacun des gènes d'un organisme un ensemble d'oligonucléotides synthétisés in situ. La première de ces puces à ADN s'appelait la « Gene ChipTM HIV PRT ». Commercialisée en 1998 par Affymetrix (Santa Clara, CA, USA), elle avait été conçue pour l'analyse des mutations de la transcriptase inverse et de la protéase du virus HIV (cité dans Hinfray, 1997). La même année a vu le développement de la première puce à oligonucléotides dédiée à une bactérie, contenant un sous-ensemble de cent gènes de Streptococcus pneumoniae [33]

Aujourd'hui ces trois distinctions n'ont plus vraiment lieu d'être, d'autant plus que ces techniques sont utilisées de façon croisée, comme le montre l'exemple de puces à ADN utilisant des produits PCR et des cibles radioactives. Les terminologies « puce à ADN » et « microarray » sont donc employées de façon indifférente. Les termes de « biopuce » ou « microréseau » sont également employés dans la littérature française [33] [34] .

1.3.2 Principe Puce à ADN

L'ADN est la forme de stockage de l'information génétique de tous les êtres vivants. La technologie des puces à ADN ou biopuces, connaît à l'heure actuelle un essor exceptionnel et suscite un formidable intérêt dans la communauté scientifique. Cette technologie a été développée au début des années 1990 et permet la mesure simultanée des niveaux d'expression de plusieurs milliers de gènes, voire d'un génome entier [30] .

L'utilité de ces informations est scientifiquement incontestable car la connaissance du niveau d'expression d'un gène constitue une avancée vers sa fonction, mais également vers le criblage de nouvelles molécules et l'identification de nouveaux médicaments et de nouveaux outils de diagnostic [12] .

De cette manière, on peut comparer les gènes exprimés dans différentes cellules (par exemple : cellules différenciées, cellules tumorales). [5] .

1.3.3 Etapes d'une analyse par puces à ADN

Les différentes phases d'une analyse par puces ADN Une puce ADN (DNA microarray) sont indiquées dans la figure suivante :

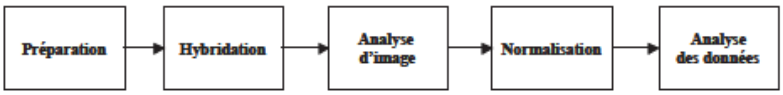


FIGURE 1.6 – Etapes d'une analyse par puces à ADN.

Préparation des cibles Il s'agit d'extraire les ARN messager d'un échantillon biologique à analyser. Une mauvaise purification peut conduire à une augmentation des bruits de fond sur la lame [7]

L'hybridation Consiste à marquer les deux échantillons pour ensuite les hybrider et les nettoyer. Les échantillons sont marqués par des substances fluorescentes, Cy3 (en vert) et Cy5 (en rouge). Ce processus d'hybridation est réalisé dans une station fluidique (four) pour favoriser les liaisons entre séquences complémentaires [7] .

Acquisition et analyse des images L'obtention des images est réalisée par lecture des puces sur des scanners de haute précision, adaptés aux marqueurs utilisés. Le procédé de détection combine deux lasers, pour exciter les fluorochromes Cy3 et Cy5. On obtient alors deux images dont le niveau de gris représente l'intensité de la fluorescence lue. Si on remplace les niveaux de gris par des niveaux de vert pour la première image et par des niveaux de rouge pour la seconde, on obtient en les superposant une image en fausses couleurs composée de spots allant du vert au rouge quand un des fluorophores domine, en passant par le jaune (même intensité pour les deux fluorophores). Le noir symbolise l'absence de signal. L'intensité du signal de fluorescence pour chaque couple (gène,spot) est proportionnel à l'intensité d'hybridation donc à l'expression du gène ciblé (voir Figure suivante) [21]

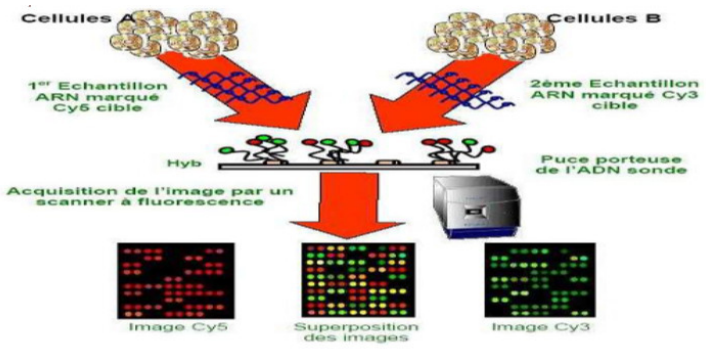


FIGURE 1.7 – Processus d'acquisition de l'image

Normalisation des données La normalisation consiste à ajuster l'intensité globale des images acquises sur chacun des deux canaux rouge et vert, de manière à corriger les différences systématiques entre les échantillons sur la même lame, qui ne représentent pas de variations biologiques entre les échantillons et qui tendent à déséquilibrer le signal de l'un des canaux par rapport à l'autre [7] .

Présentation des données de puces à ADN Après les transformations, les données recueillies pour l'étude d'un problème de donné sont regroupées sous forme matrice avec une ligne par couple (gène, sonde) et une colonne par échantillon. Les matrices de données qui sont actuellement disponibles ont donc les caractéristiques suivantes : la grande dimensionnalité due au nombre élevé de descripteurs (gènes) et le nombre limité d'échantillons [7] .

1.4 Plateformes

Il existe actuellement deux types de puces à ADN qui dominent le marché :

— Les puces à ADNc qui fonctionnent avec des micros points contenant des fragments d'ADN sur un support de verre. La société Agilent est l'une des plus grandes industries qui les commercialisent [27] [19] .

— Les puces à oligonucléotides qui reposent sur le principe de synthèse in situ de milliers de séquences distinctes d'oligonucléotides. La société Affymetrix est l'unique détenteur de cette technologie [27] [19] .

1.4.1 Technologie Agilent

Les puces à ADNc de la technologie Agilent ont été les premières puces à être développées. Le pionnier en la matière fut Patrick Brown et ses associés de l'université de Stanford. Elles sont construites grâce à des machines robots qui déposent des points appelés spots contenant des fragments d'ADN (50-150 m) dans une lamelle de verre (Figure 2.5) [18] .

FIGURE 1.8 – Puce à ADNc d'Agilent Technologies

1.4.1.1 Avantages

L'avantage des puces à ADNc de Agilent Technologies est le faible coût qu'elles comportent grâce à un prix très abordable. L'utilisation de ces puces ne nécessite pas de matériel spécifique pour effectuer les expériences et son accessibilité facilite la récupération des résultats à des fins d'analyse. Notamment l'importation de données en utilisant des équipements universels existants dans la plupart des laboratoires de recherche [18] .

1.4.2 Technologie Affymetrix

Affymetrix était une société américaine qui fabriquait des puces à ADN ; il était basé à Santa Clara, Californie, États-Unis. La société a été acquise

par Thermo Fisher Scientific en mars 2016 .

Elles dérivent à l'origine d'un projet de séquençage par hybridation. Les sondes sont des oligonucléotides synthétisés par une technique de photolithographie. Cette technique consiste à diriger une lumière sur des sites spécifiques de la puce ce qui active la réaction d'oligosynthèse. On ajoute également des oligonucléotides dont la séquence varie pour une seule base pour confirmer que le signal obtenu pour chacun des gènes est bien spécifique [19] .

On hybride une seule expérience par puce et l'intensité de fluorescence mesurée par un scanner permet une mesure de l'abondance relative de chacun des ARNm présent dans l'échantillon biologique étudié.

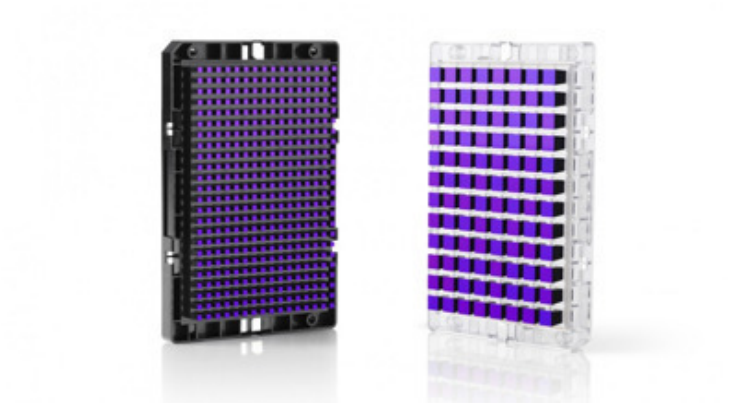


FIGURE 1.9 – Puce à oligonucléotides d'Affymetrix

1.4.2.1 Avantages

La synthèse d'oligonucléotides comprend plusieurs avantages notamment la vitesse, la spécificité et la reproductibilité. La vitesse de génération des données sur la puce est un avantage crucial, puisque il suffit juste de repérer les séquences de gènes d'intérêt de l'ADN, donc on ne perd pas de temps à la manipulation des ressources d'ADNc telles que la préparation et la détermination précise de la manipulation clones bactériens, les produits de la Réaction en Chaine par Polymérase (PCR) ou des ADNc, réduisant ainsi le risque de contamination. Cependant, avant la fabrication de la matrice, la connaissance préalable de la séquence du génome est nécessaire pour concevoir les ensembles d'oligonucléotides, et lorsque cela n'est pas disponible, d'autres méthodes d'impression du matériel génétique isolé peuvent être utilisées [19] .

1.5 Bases des données de biologie

Une base de données est un ensemble structure et organise permettant le stockage de grandes quantites d'informations afin d'en faciliter leur utilisation.*..

Pour permettre le stockage et l'organisation des données biologiques à différents niveaux, de nombreuses bases de données ont été mise en place telles que des bases de données de :

- séquences : GenBank, EMBL Nucleotide Sequence Database et DNA Data Bank of Japan (DDBJ), Eucaryotic Promoter Database (EPD) ;
- protéines : UniProt, Protein Data Bank (PDB), InterPro, Institut européen de bioinformatique (EBI) ;
- génomiques spécialisées : Saccharomyces Genome Database (SGD), FlyBase, Worm-Base, The Arabidopsis Information Resource, Zebrafish Information Network
- polymorphismes génétique : dbSNP, Hapmap ;
- voies de signalisation : KEGG, REACTOME, Panther, NCBICP.

La bioinformatique est le traitement automatique de l'information biologique, sous forme de données. La base de données (données primaires) est donc la matière première à partir de laquelle la bioinformatique va produire d'autres données (données secondaires) et construire de nouvelles connaissances.

Toutes ces bases sont interconnectées grâce à l'utilisation d'un identifiant unique pour caractériser une séquence, un gène, un transcrit ou une protéine, comme c'est le cas sur le site du National Center for Biotechnology Information (NCBI) [9] .

Les principales utilisations d'une base de données de microréseaux sont de stocker les données de mesure, de gérer un index interrogeable et de mettre les données à la disposition d'autres applications pour analyse et interprétation (directement ou via des téléchargements d'utilisateurs) [11] .

1.5.1 Gene Expression Omnibus

Gene Expression Omnibus [16] est un entrepôt public à haute capacité de traitement des données génomique et protéomique, essentiellement MIAME. Il a été établi en 2000 au National Center for Biotechnology Information (NCBI) [9] . Les données expérimentales peuvent être soumises en remplissant un formulaire sur le web ou comme un paquet de fichiers, feuille de

calcul, fichier texte SOFT (Simple Omnibus Format in Text) ou fichier XML MINiML (MIAME Notation in Markup Language). Les fichiers sont stockés sous la forme de 3 types d'enregistrement basiques [4] :

— Plateform : Description du tableau
— Sample : Description d'un échantillon biologique et les résultats de son hybridation
— Series : Description de l'expérience réalisée sur un groupe d'échantillon Basées sur les études expérimentales soumises, les données dans GEO sont organisées dans des objets de haut niveau représentés par le type Dataset (Jeu de données), qui est une collection d'échantillons biologiques comparables ayant été traités sur la même plateforme et dont les mesures sont les résultats de ce traitement et de calculs cohérents sur ce jeu de données, et Profils, qui correspondent au niveau d'expression d'un gène dans tous les échantillons d'un jeu de données .

1.5.2 ArrayExpress

ArrayExpress est une base de données publique pour stocker et fournir un accès à des données de génomique fonctionnelle à haut débit. ArrayExpress se compose de deux composants spécialisés à des fins distinctes : le référentiel ArrayExpress des données expérimentales archivées accessibles au public et l'ArrayExpress Data Warehouse des profils d'expression génique [29] .

Elle est constituée de 3 composantes [29] :

— ArrayExpress repository : Qui est conforme au standard MIAME. Les expériences peuvent être soumises à cet entrepôt grâce à l'outil en ligne MIAMExpress ou en chargeant des tableurs (MAGE-TAB de préférence) .
— ArrayExpress Warehouse : qui est une base de données de gènes, sélectionnés à partir de l'ArrayExpress repository, dont les profils d'expression sont indexés .
— ArrayExpress Atlas : Qui est une nouvelle base de données résumée pour interroger les gènes d'expression organisés et classés à travers de multiples expériences et conditions.

1.5.3 La MGED (Microarray Gene Expression Data Society)

La MGED a initié le développement et la promotion de standard pour le stockage et le partage des données de puces à ADN basées sur l'expression des gènes et du résultat des études effectuées sur ces données. Parmi ces standards l'on peut citer le MIAME (Minimum Information About a Microarray Experiment). MIAME est un standard conceptuel décrivant l'information minimum requise pour une interprétation et une vérification propre des expériences des puces à ADN tandis que MAGEML et MAGE-TAB sont des standards définissant le format MIAME (conformité de la description des données et des expériences) [27] .

1.5.3.1 MIAME

Le standard MIAME requiert que les informations suivantes soient fournies pour les publications basées sur les expériences de puce à ADN [27] [19] :

1. Les données brutes résultant de l'analyse de l'image de chaque puce (fichiers CEL).
2. Les données finales après le prétraitement qui est la matrice d'expression des gènes
3. Les informations essentielles à propos de l'annotation de l'échantillon et des facteurs
4. expérimentaux.Le plan expérimental incluant les relations entre échantillons, puces et fichiers de données.
5. expérimentaux.Le plan expérimental incluant les relations entre échantillons, puces et fichiers de données.
6. Les protocoles de traitement expérimentaux des données. Le standard MIAME ne requiert pas que les données soient dans un format spécifique, il recommande toutefois l'utilisation du format MAGE-TAB ou MAGE-ML.

1.6 Les outils de traitement

1.6.1 Les outils d'analyse d'image

L'analyse automatique de l'image générée par la puce à ADN est dans un premier temps nécessaire. Cela permet notamment de repérer et séparer les spots, d'éliminer les spots défectueux, et d'annoter chaque spot avec son intensité lumineuse globale, afin d'obtenir des résultats numériques exploitables pour un traitement automatisé [27] .

Le principe général de l'analyse d'image est de convertir l'image en valeurs numériques quantifiant l'expression des gènes. Il existe des logiciels d'analyse d'image comme : ScanAlyze [17] , Genepix Pro [38] .

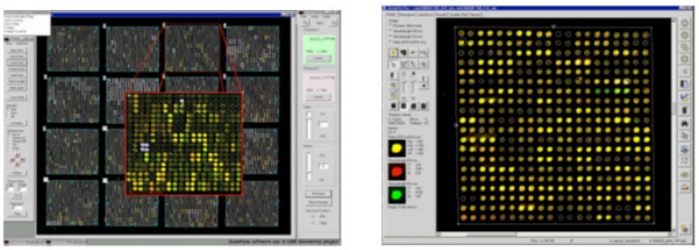


FIGURE 1.10 – isualisation d'un scan à l'aide de ScanAlyze et GenePix Pr

1.6.2 Langage R

R est actuellement l'outil le plus utilisé pour le traitement numérique des données biologiques. R est un outil d'analyses statistiques et graphiques qui possède son propre langage de programmation. R est très vite devenu disponible pour les systèmes d'exploitation Windows et Mac-OS.
il existe plusieurs langage proposent différentes solutions pour le traitement et l'analyse des données biologiques le projet BioConductor,Les projets BioPerl, BioJava, BioPython etc.

FIGURE 1.11 – La page d'accueil langage R

1.6.3 Le projet BioConductor

BioConductor est une initiative de collaboration entre statisticiens, mathématiciens,biologistes et développeurs afin de créer des outils informatiques (algorithmes, logiciels) pour résoudre des problèmes de biologie et de bioinformatique [15] .

Né en 2000, BioConductor, associé à R, reçoit en 2002 le titre de Insightful Innovation AwardOpen Source Open Development Software Project. Dédiées à l'analyse des données de génomique, les librairies disponibles sur le site de-BioConductor permettent non seulement l'analyse des données de puces à ADN (e.g. librairies Affy, marray, limma) mais aussi des expériences SAGE (SAGElyzer), de la spectrométrie de masse (PROcess) ou encore l'annotation des gènes (GOsta) [3] .

1.7 Synthèse du chapitre

Nous avons présenté dans ce chapitre la définition de un domaine de recherche de bioinformatique et des notions élémentaires en biologie qui sont la base de notre sujet de recherche dans ce mémoire.

Nous avons présenté les différentes étapes d'une analyse par puce à ADN, telles que la préparation des cibles et l'hybridation, acquisition et analyse des images et transformation des données.

Et aussi nous avons présenté banques de données génomiques publiques et les différents outils de traitement tels que les outils d'analyse des images et les outils de traitement numérique des données biologiques.

CHAPITRE

2

ETAPES D'ANALYSE DES DONNÉES DE BIOPUCES "MICROARRAY"

2.1 Introduction

La technologie d'analyse des données volumineuses peut extraire les informations génétiques liées aux maladies et aux médicaments à partir de données génétiques massives et fournir de nouvelles idées pour le développement de médicaments ainsi que le diagnostic et le traitement des maladies. Par conséquent, les grandes données ont des effets positifs sur la recherche sur le cancer [35].

L'analyse des données génétiques comprend quatre étapes : l'acquisition des données génétiques, le prétraitement des données génétiques, la sélection des gènes et l'établissement et l'évaluation des modèles de classification. Parmi ces étapes, l'acquisition de données génétiques est un processus biomédical, et les autres étapes sont des processus d'exploration de données [35].

En général la base de données est très dense, composée de plusieurs milliers d'attributs. Le temps de traitement est très grand. La sélection d'attributs consiste à déterminer le sous ensemble optimal suffisant d'attributs [21].

L'extraction de connaissances à partir de ces données peut se faire par l'utilisation de techniques d'apprentissage automatique . Cependant, ces données contiennent un très grand nombre des attributs [20].

la sélection des attributs consiste à évaluer chaque attribut pour lui assigner un score de pertinence qui permet un classement des attributs. Les attributs les mieux classés c'est-à-dire les plus pertinents seront sélectionnés pour la phase du traitement [27].

2.2 Etapes du prétraitement des données

Le résultat de l'acquisition d'une puce ADN par le scanner est une image dans laquelle chaque spot coloré correspond à l'expression d'un gène détectée par une sonde. Ces données brutes doivent subir des étapes de prétraitement avant de pouvoir être analysées [13] .

Les 3 étapes du prétraitement des données de la sont :

1. La Correction du bruit de fond : par puce, correction de l'intensité de chaque sonde en utilisant l'information disponible sur cette puce.
2. La normalisation entre les puces : détection et correction des différences systématiques entre les puces.
3. Le résumé : calcul d'une mesure d'expression par gène et puce à partir des intensités des sondes des sondes.

2.2.1 Correction du bruit de fond (Background Correction)

La correction du bruit de fond : par puce, correction de l'intensité de chaque sonde en utilisant l'information disponible sur cette puce [29] .

Le bruit de fond apporte une information supplémentaire dont il serait dommage de ne pas tenir compte, mais comment le prendre en considération ?

Après l'hybridation, une puce à ADN est scannée pour pouvoir générer des fichiers où les résultats de l'hybridation sont traduits numériquement. On obtient dans ces fichiers une quantité énorme d'information .

2.2.1.1 Correction par Robust Multi-Array Analysis (RMA) :

Lors du développement de RMA[27], les auteurs ont jugé que les sondes MM posaient plus de problème qu'elles n'en solutionnaient et ont proposé de ne plus utiliser que les valeurs des sondes PM. Cette correction se fait en utilisant un modèle basé sur la distribution empirique des intensités de sondes. Le modèle observé (Y) est une somme de la composante « bruit » (B) et de la composante «signal » (S)

$$Y = B + S \tag{2.1}$$

Avec B une distribution proche d'une distribution normale et S une composante exponentielle. Pour éviter la possibilité de valeurs d'expression négative, il est nécessaire de tronquer la distribution normale à zéro.

2.2.2 Normalisation

Une normalisation précise est essentielle pour obtenir des résultats fiables et reproductibles dans les études d'expression génique. Cette étude démontre la variabilité associée au tissu dans la sélection de ces gènes de normalisation et de souligne l'importance de la sélection des gènes de référence corrects à la fois pour le modèle animal et le tissu étudié [30] .

une normalisation est appliquée aux données dans le but de faire ressortir les différences réellement dues aux variations d'expression des transcrits entre les échantillons [13] .

Avant l'application d'une transformation logarithmique, la plupart des intensités mesurées sont faibles, la transformation logarithmique permet de recentrer la distribution et de la rendre symétrique, ce qui facilite l'utilisation des statistiques. A noter que la transformation logarithmique à base 2 est la plus utilisée .

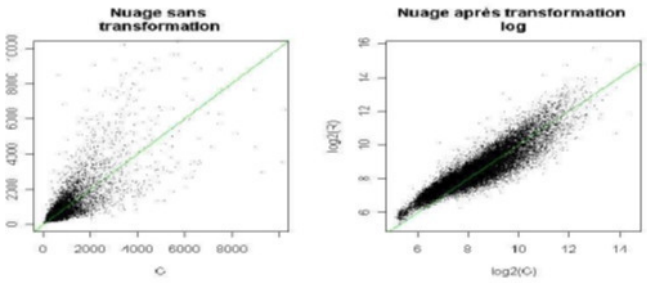


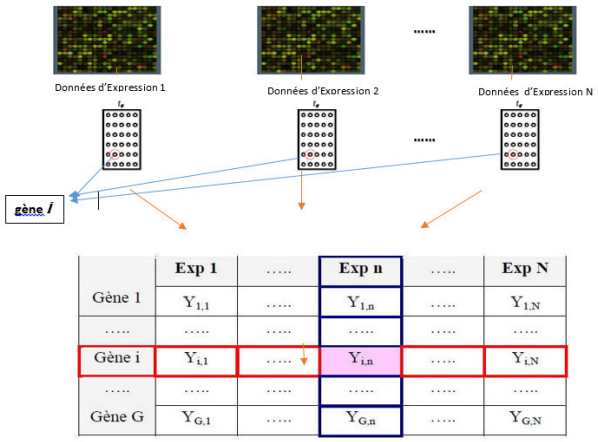
FIGURE 2.1 – Nuage de points pour une biopuce , avant et aprés la transformation log. Sur cette figure, on peut voir 2 nuages de points correspondant au même jeu de données, à gauche sans aucune transformation, à droite avec un passage au logarithme de base 2.

2.2.3 Sommarisation

C'est une étape propre à toute plateforme pour laquelle un même transcrit est sondé par plusieurs sondes que l'on doit résumer en une seule valeur d'expression.

2.3 La représentation des données

Le processus d'extraction de connaissances à partir des données des biopuces est décrit dans la figure 2.2. Il consiste d'une part à analyser les images et à normaliser les données et d'autre part à appliquer des méthodes statistiques et informatiques pour obtenir de la connaissance utile pour les experts [13] .



	Exp 1	.....	Exp n	.....	Exp N
Gène 1	$Y_{1,1}$	.....	$Y_{1,n}$	.....	$Y_{1,N}$
.....	.....	.....	.....	.....	.....
Gène i	$Y_{i,1}$	.....	$Y_{i,n}$	.....	$Y_{i,N}$
.....	.....	.....	.....	.....	.....
Gène G	$Y_{G,1}$	.....	$Y_{G,n}$	.....	$Y_{G,N}$

FIGURE 2.2 – représentation des données biopuce 'microarray'

Cette étape implique la représentation de la matrice de données pour la classification. Les données de puces à ADN se présentent sous la forme d'une « matrice d'expression » on notera Y , la mesure d'expression du gène i dans l'expérience , pour (G étant le nombre total de gènes) et (N étant le nombre total d'expériences) [25].

Il y a deux manières d'analyser les données d'expression. L'analyse peut porter sur la recherche de similitudes de comportements entre les gènes ou alors entre les expériences. Dans le premier cas, chaque gène est caractérisé par un vecteur d'expression et dans le second cas se sont les expériences qui sont caractérisées par des vecteurs [21] .

2.4 Sélection des attributs pour traitement des données

La sélection d'attributs : En général la base de données est très dense, composée de plusieurs milliers d'attributs. Le temps de traitement est très grand. La sélection d'attributs consiste à déterminer le sous ensemble optimal suffisant d'attributs [23] .

La sélection d'attributs est un problème complexe qui a déjà été largement étudié, mais les dimensions des données des biopuces nécessitent des approches spécifiques (plusieurs milliers de gènes)

Le principe de la sélection des attributs consiste à évaluer chaque attribut pour lui assigner un score de pertinence qui permet un classement des attributs. Les attributs les mieux classés c'est-à-dire les plus pertinents seront sélectionnés pour la phase du traitement [35] .

L'avantage de la sélection est qu'elle peut être utilisée lorsqu'on travaille avec un très grand nombre d'attributs car elles sont de complexité raisonnable.

Pourquoi la sélection d'attributs :

— Dimension des entrées telle que coût de l'apprentissage trop grand
— Apprentissage moins coûteux(Faciliter l'apprentissage, Meilleure performance en classification)

La mesure de pertinence utilisée dans une méthode filtre peut être une mesure statistique classique telle que la **t-statistique** et le test de **Fisher**. Certaines mesures de filtrage ont été proposées spécifiquement pour la sélection de gènes telles que **B/W**.

a) t-statistique Le filtre t-score, ou t-test, est basé sur le test t de Welch (Welch, 1947), utilisé en statistiques pour tester la significativité de la différence entre les moyennes d'une variable dans deux groupe. Ce test fait l'hypothèse d'une distribution normale de la variable testée, ou d'un échantillon "suffisamment" grand (classiquement, n>30), et il tient compte, contrairement au t de Student, du cas où la variance de la variable testée est inégale entre les deux groupes. Sa formule est [13] [21] :

$$t = \frac{x1 - x2}{\sqrt{\frac{s1^2}{n1} + \frac{s2^2}{n2}}} (2.2)$$

Dans une utilisation pour la sélection de variables, ce t-score est réalisé sur toutes les variables entre les deux groupes à classifier. Puis les variables sont classées selon la valeur absolue de leur score, et le filtre garde les d meilleures variables (ce d étant un paramètre soit à définit a priori, soit à optimiser par une validation croisée imbriquée), ou encore toutes les variables ayant un score supérieur à un seuil à définir lui aussi.

Bien que d'une part il soit très simple, et que d'autre part l'hypothèse de normalité de la distribution des variables n'est pas nécessairement réalisée sur les données puces, ce filtre a permis, sur données puces, d'obtenir des

sélections parmi les plus stables (ou les moins instables !), tout en permettant de bonnes performances de classification [29] .

b) Fisher Le test de Fisher est défini comme suit :

$$f = \frac{x1 - x2}{s1 + s2} \tag{2.3}$$

où xk et sk2 sont la moyenne et l'écart-type de l'attribut pour la classe k = 1, 2.un score important indique donc que les moyennes des 2 classes sont significativement différentes.

c) BW : Le score discriminant BW est basé sur le rapport entre dispersion entre classes et dispersion intra-classes pour un attribut j, ce rapport est obtenu comme suit :

$$BW = \frac{\sum i \sum j (y_{i=K})(x_{kj} - x_j)}{\sum i \sum j (y_{i=K})(x_{ij} - x_{kj})} \tag{2.4}$$

où x j et x kj dénotent respectivement la moyenne d'un attribut j à travers tous les échantillons et à travers les échantillons appartenant à la classe k seulement [27] [21] .

d) IM Information Mutuelle : L'information mutuelle est une mesure de la dépendance statistique entre deux variables. Dans le contexte de la sélection de variables, il s'agit donc de mesurer la dépendance entre chaque variable explicative (gène...) et la variable d'intérêt (classe) : plus cette dépendance est forte, plus le filtre considère la variable pertinente. L'information mutuelle I(X,Y) entre une variable X et une variable Y est définit par [29] :

$$MI(X,Y) = \sum_{y \varepsilon Y} \sum_{x \varepsilon X} p(X = x, Y = y) log \frac{p(X = x, Y = y)}{p(X = x)p(Y = y)} \tag{2.5}$$

e) RFE(Backward Feature Elimination) La recursive feature elimination (RFE) (ou backward feature elimination) consiste à éliminer progressivement les variables les moins discriminantes. Pour cela, on part de l'ensemble de variables initial, sur lequel on construit un classifieur, puis on élimine une certaine proportion des variables les moins discriminantes (on peut les éliminer une à une, mais sur des données hautes dimensions on prend généralement des paliers plus larges, par exemple 10/100 [29].

le but de l'élimination récursive des fonctions (RFE) consiste à sélectionner des entités en considérant de manière récursive des ensembles d'entités de plus en plus petits [24].

— Premièrement, l'estimateur est formé sur l'ensemble initial de caractéristiques et l'importance de chaque caractéristique.
— Ensuite, les fonctionnalités les moins importantes sont supprimées de l'ensemble actuel de fonctionnalités.
— Cette procédure est répétée de manière récursive sur l'ensemble élagué jusqu'à ce que le nombre désiré de caractéristiques à sélectionner soit finalement atteint [24].

2.5 Synthèse du chapitre

L'analyse des données génétiques comprend quatre étapes : l'acquisition des données génétiques, le prétraitement des données génétiques, la sélection des gènes et l'établissement et l'évaluation des modèles de classification. Puis nous avons présenté les différents critères statistiques utilisés pour la sélection des gènes comme les tests statistique et Manual information et RFE.

Et nous allons voir dans le chapitre suivant les différentes méthodes de classification des données,nous nous concentrons sur la méthode K plus proche voisins

CHAPITRE

3

CLASSIFICATION DES DONNÉES DE BIOPUCES MICROARRAY

3.1 Introduction

Les méthodes d'apprentissage automatique sont d'un intérêt majeur en recherche diagnostique, plus précisément dans l'identification des combinaisons de bioinformatique qui constitueront les futurs tests de diagnostic.

L'apprentissage automatique est l'un des domaines phares de l'intelligence artificielle. Il concerne l'étude et le développement de modèles quantitatifs permettant à un ordinateur d'accomplir des tâches sans qu'il soit explicitement programmé à les faire. Apprendre dans ce contexte revient à reconnaître des formes complexes et à prendre des décisions intelligentes. Compte tenu de toutes les entrées existantes, la complexité pour y arriver réside dans le fait que l'ensemble des décisions possibles est généralement très difficile à énumérer. Les algorithmes issus de ce domaine sont utilisés par plusieurs autres domaines, tels que la vision par ordinateur, la reconnaissance de forme, la recherche d'information, la bioinformatique, la fouille de données et beaucoup d'autres

Les méthodes de classification ont pour but d'identifier les classes auxquelles appartiennent des objets à partir de certains paramètres descriptifs. Elles s'appliquent à un grand nombre d'activités humaines et conviennent en particulier au problème de la prise de décision automatisée. La procédure de classification sera extraite automatiquement à partir d'un ensemble d'exemples. Un exemple consiste en la description d'un cas avec la classification correspondante. Un système d'apprentissage doit alors, à partir de cet ensemble d'exemples, extraire une procédure de classification, il s'agit en effet d'extraire une règle générale à partir des données observées. La procédure générée devra classifier correctement les exemples de l'échantillon et avoir un bon pouvoir prédictif pour classifier correctement de nouvelles descriptions. Les méthodes utilisées pour la classification sont nombreuses, citons : la méthode K-Plus Proches Voisins (K-PPV), des Séparateurs à Vastes Marges (SVM), les Réseaux de Neurones, etc...

3.2 Classification

La classification est la tâche consistant à attribuer une classe à une donnée qu'on veut classer, ou autrement dit, à assigner une donnée à une classe de données.
Plus formellement, soit X £ R d un ensemble représentant un espace à d dimensions, appelé l'espace des instances. La donnée x £ X est appelée une instance et représente un point dans l'espace X. L'instance x est présentée sous forme d'un vecteur de taille d, x = (x(1) ;..... ; x(d)), où chaque composante x(i) £ R est une valeur discrète ou continue. Soit Y un ensemble fini de classes où chaque classe y £ Y est présentée sous forme d'une valeur discrète appelée étiquette de classe (un nom ou identifiant unique pour la classe) [8].

Le classifieur se présente alors sous forme d'une fonction de classification h (appelée aussi modèle de classification) permettant d'associer une donnée x £ X à une étiquette de classe y £ Y. (équation 3.1).

$$h : |X \longmapsto Y, x \longmapsto y = h(x) \tag{3.1}$$

3.3 Apprentissage automatique

3.3.1 Définition

L'apprentissage automatique (en anglais machine learning) est une branche de l'intelligence artificielle qui s'intéresse à l'extraction d'informations par des méthodes automatisées qui réalisent un « apprentissage » à partir de données

On parle d'apprentissage artificiel (ou apprentissage automatique) lorsqu'un programme a la capacité d'améliorer ses performances à partir de données acquises en cours de fonctionnement. Ce type de système est utilisé pour résoudre des taches trop complexes pour les algorithmes classiques.

L'apprentissage automatique est le processus de la construction des modèles à partir d'un ensemble de données (exemples). Le modèle sera créé complètement ou bien à partir l'amélioration d'un modèle partiel ou général. L'apprentissage étant un sujet central depuis les débuts de l'intelligents artificielle [8].

Dans l'apprentissage automatique, les tâches sont généralement classées en grandes catégories. Ces catégories sont basées sur la façon dont l'apprentissage est reçu ou comment le feedback sur l'apprentissage est donné au système développé.

Deux des méthodes d'apprentissage automatique les plus largement adoptées sont l' apprentissage supervisé qui forme des algorithmes basés sur des données d'entrée et de sortie étiquetées par l'homme et l'apprentissage non supervisé qui ne fournit pas à l'algorithme des données étiquetées pour lui permettre de trouver une structure et de découvrir une logique dans données entrées. Explorons donc ces méthodes plus en détail [8].

3.3.2 Types d'apprentissage

Les algorithmes d'apprentissage peuvent se catégoriser selon le mode d'apprentissage qu'ils emploient :

3.3.2.1 L'apprentissage non supervisé

Quand le système ne dispose que d'exemples, mais sans étiquettes, et que le nombre de classes et leur nature n'ont pas été prédéterminés, on parle d'apprentissage non supervisé (ou clustering). Dans ce cas le but d'apprentissage est de grouper les exemples selon leurs attributs en basant sur la notion de la similarité qui est généralement calculée selon la fonction de distance entre paires d'exemples [2]

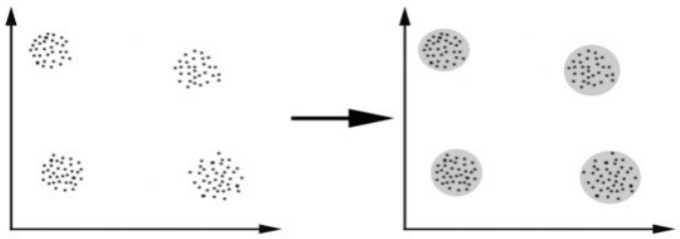


FIGURE 3.1 – Illustration de regroupement en clusters

3.3.2.2 L'apprentissage supervisé

Si les classes sont prédéterminées et les exemples connus, le système apprend à classer selon un modèle de classement ; on parle alors d'apprentissage supervisé (ou d'analyse discriminante) [2] .

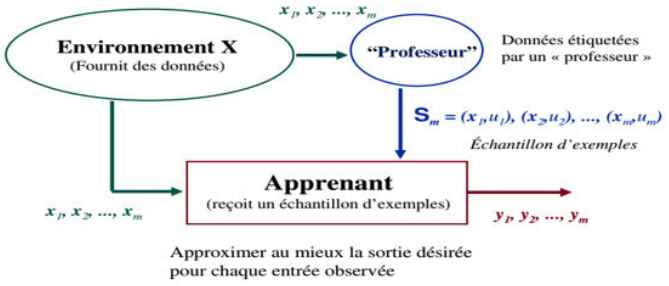


FIGURE 3.2 – L'apprentissage supervisé

3.4 K-Plus Proches Voisins (K-PPV)

3.4.0.1 Présentation

Le principe général des KPPV est relativement simple. Pour chaque profil distance de notre base de test, on calcule la distance de celui-ci à tous les profils distance de la base d'apprentissage. La distance utilisée reste à définir en tenant compte du type de données à classifies. Nous reviendrons sur ce point un peu plus tard dans le rapport. Une fois que les distances entre le profil testé et chaque profil distance de la base d'apprentissage ont été calculées, il reste à prendre une décision. Le profil distance de la base de test sera en fait classé dans la classe revenant majoritairement parmi les K Plus Proches Voisins. Dans le cas de notre application, lorsqu'une seule classe ne ressort pas majoritairement, on choisira parmi les classes encore en course, cette possède le voisin les plus proche du profil distance testé. On effectue cette opération pour chaque profil distance de la base de test [13] .

3.4.1 Principe de la technique k-ppv

Définition les K-nn (K-nearest neighbor) : Méthode d'apprentissage supervisée qui raisonnent avec le principe sous-jacent : " Dis moi qui sont tes amis, je te dirais qui tu es". Elle diffère des méthodes d'apprentissages traditionnelles car aucun modèle n'est induit à partir d'exemple. A chaque fois que l'on veut classer un nouvel individu, on refait tourner l'algorithme et on cherche de nouveaux amis [23] .

Exemple : Si l'on veut prédire la probabilité de survenue d'un cancer chez un nouveau patient on procède en deux étapes :

1. On recherche selon les caractéristiques de ce patient les patients qui lui ressemble .

1. Si parmi ces “voisins”, il y a eu plus de cancer, alors le patient a une forte probabilité d’avoir le cancer.

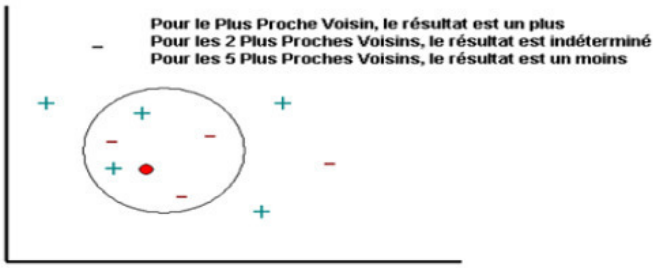


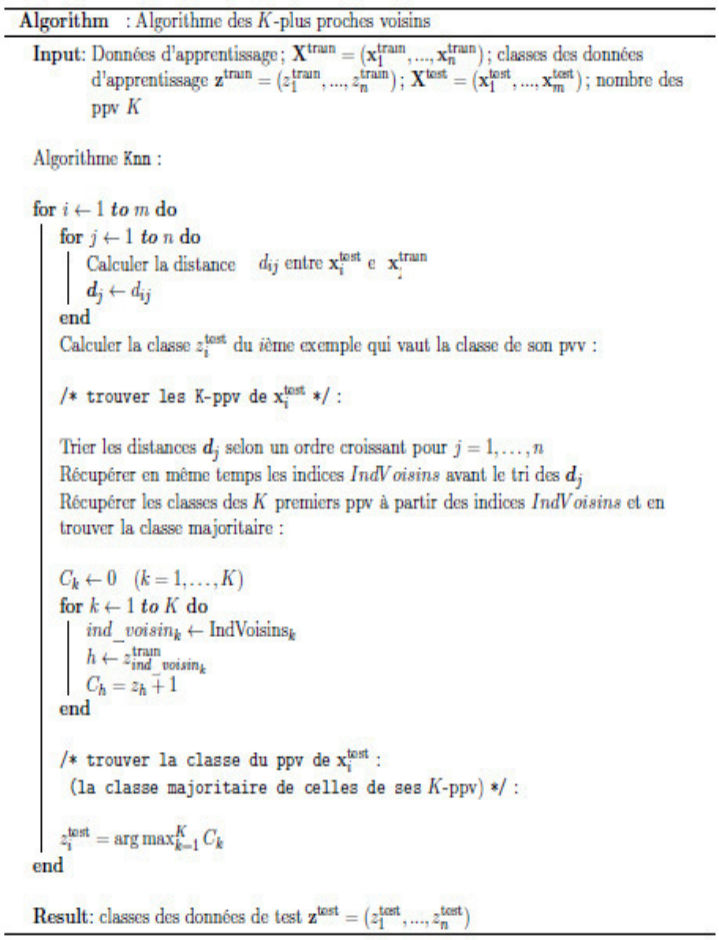
FIGURE 3.3 – L’illustrer l’analyse des K Plus Proches Voisin

Afin d’illustrer l’analyse des K Plus Proches Voisins, considérons la problématique de la classification d’un nouvel objet (point de requête) parmi un ensemble d’exemples connus. C’est le schéma reporté ci-dessous, qui matérialise les exemples connus (instances) par des signes plus et moins et le point de requête par un cercle rouge. Notre problème consiste à déterminer le résultat (affecter à une classe) du point de requête en fonction d’un certain nombre de voisins les plus proches. En d’autres termes, nous voulons savoir si nous devons classer le point de requête comme un signe plus ou comme un signe moins [10] .

3.4.2 Algorithme

1. initialisation, choix de :
   — Nombre de classes, Valeur de k, exemples initiaux, mesure de similarité
2. pour chaque vecteur d’objet à classer :
   — mesurer la distance du vecteur avec tous les autres déjà classés
   — déterminer la liste des k vecteurs les plus proches de lui (k-ppv)
   — déterminer la classe la plus représentée dans la liste des k-ppv et affecter notre vecteur à cette classe.

Algorithme de la méthode k-ppv tel que le paramètre k >= 1 [10] :



Algorithm : Algorithme des K-plus proches voisins

Input: Données d'apprentissage ; $\mathbf{X}^{train} = (\mathbf{x}_1^{train}, ..., \mathbf{x}_n^{train})$; classes des données d'apprentissage $\mathbf{z}^{train} = (z_1^{train}, ..., z_n^{train})$; $\mathbf{X}^{test} = (\mathbf{x}_1^{test}, ..., \mathbf{x}_m^{test})$; nombre des ppv K

Algorithme `Knn` :

for $i \leftarrow 1$ ***to*** m **do**
  for $j \leftarrow 1$ ***to*** n **do**
    Calculer la distance d_{ij} entre $\mathbf{x}_i^{test}$ e $\mathbf{x}_j^{train}$
    $\boldsymbol{d}_j \leftarrow d_{ij}$
  end
  Calculer la classe z_i^{test} du $i$ème exemple qui vaut la classe de son pvv :

  `/* trouver les K-ppv de` $\mathbf{x}_i^{test}$ `*/` :

  Trier les distances $\boldsymbol{d}_j$ selon un ordre croissant pour $j = 1, \ldots, n$
  Récupérer en même temps les indices $IndVoisins$ avant le tri des $\boldsymbol{d}_j$
  Récupérer les classes des K premiers ppv à partir des indices $IndVoisins$ et en trouver la classe majoritaire :

  $C_k \leftarrow 0 \quad (k = 1, \ldots, K)$
  for $k \leftarrow 1$ ***to*** K **do**
    $ind\_voisin_k \leftarrow \text{IndVoisins}_k$
    $h \leftarrow z_{ind\_voisin_k}^{train}$
    $C_h = z_h + 1$
  end

  `/* trouver la classe du ppv de` $\mathbf{x}_i^{test}$:
   `(la classe majoritaire de celles de ses` K`-ppv) */` :

  $z_i^{test} = \arg\max_{k=1}^{K} C_k$
end

Result: classes des données de test $\mathbf{z}^{test} = (z_1^{test}, ..., z_n^{test})$

FIGURE 3.4 – Algorithme de la méthode k-ppv k >= 1

3.4.3 Quelques règles sur le choix de k

Le paramètre k doit être déterminé par l'utilisateur : k £ N. En classification binaire, il est utile de choisir k impair pour éviter les votes égalitaires [8] .

Le meilleur choix de k dépend du jeu de donnée. En général, les grandes valeurs de k réduisent l'effet du bruit sur la classification et donc le risque de sur-apprentissage, nécessaire pour des petites bases d'apprentissage, mais rendent les frontières entre classes moins distinctes.
Il convient donc de faire un choix de compromis entre la variabilité associée à une faible valeur de k contre un 'oversmoothing' ou sur lissage (i.e gommage

des détails) pour une forte valeur de k. Un bon k peut être sélectionné par diverses techniques heuristiques, par exemple, de validation-croisée. Nous choisirons la valeur de k qui minime l'erreur de classification.

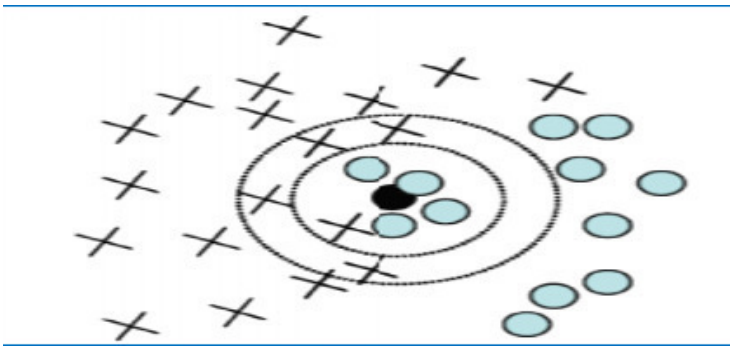


FIGURE 3.5 – Le choix de K influence de décision : pour K=5La décision est de classer l'objet « noir » dans la classe « rond ».pour K=9, La décision est de classer l'objet dans la classe « croix ».

3.4.4 Les avantages et les inconvénients K-PPV

Avantages

- Apprentissage rapide,
- Méthode facile à comprendre,
- Taux de prédictif est souvent bon ,
- Qu'elle ne pose aucune hypothèse sur la forme des classes à apprendre,
- Adapté aux domaines où chaque classe est représentée par plusieurs prototypes et où les frontières sont irrégulières (ex. Reconnaissance de chiffre manuscrits ou d'images satellites).

Inconvénient

- Mais La performance de cette méthode diminue lorsque la dimension augmente, puisque pour chaque nouvelle classification, il est nécessaire de calculer toutes les distances de x à chacun des exemples d'apprentissage,
- La performance dépend fortement de k, le nombre de voisins choisi et il est nécessaire d'avoir un grand nombre d'observations pour obtenir une bonne précision des résultats.

3.4.5 Evaluation du modèle

L'apprentissage supervisé utilise une partie des données pour calculer un modèle de décision qui sera généralisé sur l'ensemble du reste de l'espace. Il est très important d'avoir des mesures permettant de qualifier le comportement du modèle appris sur les données non utilisées lors de l'apprentissage. Ces métriques sont calculées soit sur les exemples d'entrainement eux mêmes ou sur des exemples réservés d'avance pour les tests.

La métrique intuitive utilisée est la précision du modèle appelée aussi le taux de reconnaissance. Elle représente le rapport entre le nombre de donnée correctement classées et le nombre total des données testées. L'équation suivante donne la formule utilisée.

Généralement, la précision est donnée sous forme de pourcentage ce qui nécessite de multiplier la précision de l'équation précédente par 100 [14] .

3.5 Synthèse du chapitre

K-PPV est une méthode de classification rapide qui montre de bonnes performances dans la résolution de problèmes varies. Cette méthode a montré son efficacité dans de nombreux domaines d'applications tels que le traitement d'image, la catégorisation de textes ou le diagnostiques médicales et ce même sur des ensembles de données de très grandes dimensions.

CHAPITRE

4

CONCEPTION ET RÉALISATION

4.1 Introduction

les données biopuces sont représentés sous forme d'une matrice tel que les échantillons jouer de role les individus et les genes jouer de role caractéristique (attributs).

L'objectif du travail est appliquer des K plus proches voisins (KNN) comme méthode de classification et utilisé les critères de précision pour évaluer la performance des classificateurs sur les sous-ensembles de gènes sélectionnés.

Nous avons proposé différents méthodes de classification tel que KNN et SVM et RN pour obtenir des résultats compétitifs.

Dans ce chapitre nous présentons la conception de notre système en commençant par sa conception générale puis sa conception détaillée en spécifiant les différents éléments composant notre système et précisant son fonctionnement .

4.2 Conception globale du système

La technologie d'analyse des données volumineuses peut extraire les informations génétiques liées aux précision de la catégorie diagnostique d'un exemple (échantillon). L'analyse des données génétiques comprend quatre étapes :

1. Les jeux des données biopuces : Comme une données primaires nous avons accéder à des fichier 'file-GEO.soft' .Ce fichier contient une description de l'expérience réalisée sur un groupe d'échantillon et une description sur les noms des gènes et leur niveau d'expression dans les échantillons. Alors pour préparer notre base en divisé sous deux formats,tel que :
   - Fichier de texte comme fichier d'information détaillée de base de notre base traitée.
   - Fichier sous forme CSV qui contient matrice numérique d'une biopuces sur une maladie spécifié.
2. Le prétraitement des données biopuces : Pour dirigé chaque base GEO à la traitement, il faut faire 3 étape d'analyse des donnes biopuces (voir le chapitre3).
3. La sélection des gènes : ETape finale d'analyse des donnes biopuces pour identifier le sous-ensemble genes de taille minimale nécessaire et suffisant pour l'amélioration taux de classification.
4. l'évaluation des modèles de classification : Cette étape implique la représentation de la matrice de données pour la classification. Pour l'extraction de connaissances à partir des données des biopuces en applique le classifieur pour évaluer l'efficacité et la performance de ce modèle de classification.
   La méthode utilisée pour la classification des données bio-puces dans notre travail est une méthode de classification supervisé qui est le K-ppv(KNN).

Dans ce travail, nous avons résumé la conception globale notre travial sur le schéma général d'analyse des données de biopuces :

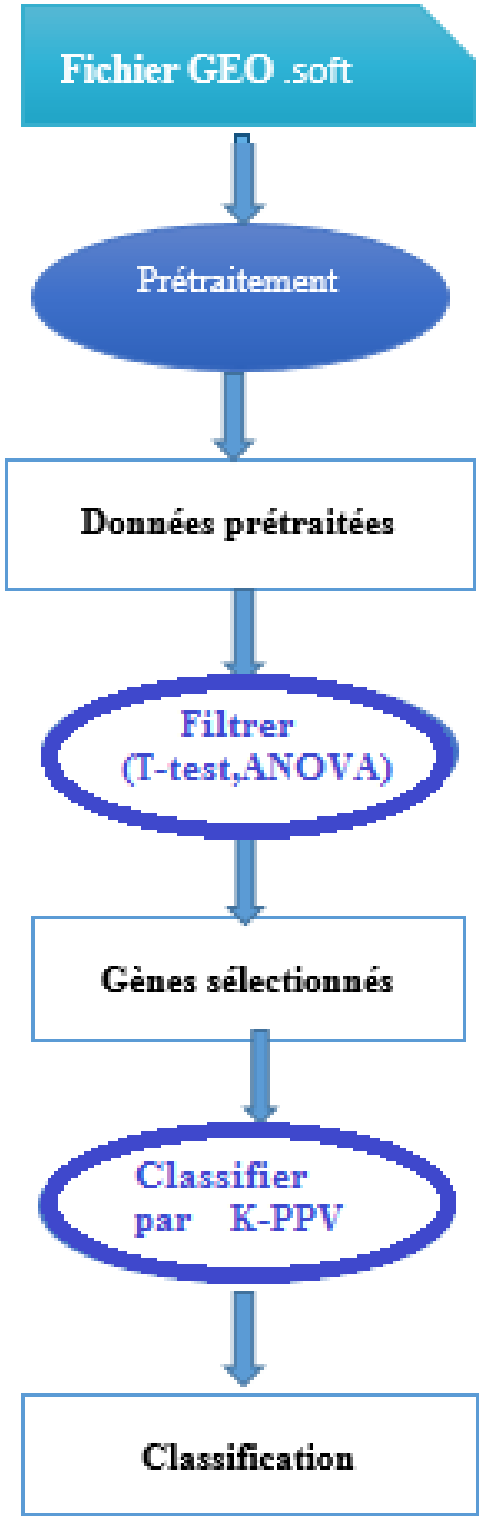


FIGURE 4.1 – le schéma général d'analyse des données de biopuces

4.3 Conception détaillée du système

4.3.1 Accéder à des données biopuces GEO

Nous avons utilisé la base de données stocke des DataSets d'expression génique GEO (Gene Expression Omnibus) , facilement accessible et qui est utilisé dans de nombreux travaux concernant la classification des données de puces à ADN .

Dans cet travail , 3 jeux des données de gènes de cancer utilisés :
— le cancer du prostate.
— le cancer du côlon.
— la tumeur du Hypertensions pulmonaires : PBMC..

On peut télécharger ces fichiers GEO du site :https ://www.ncbi.nlm.nih.gov/geo/

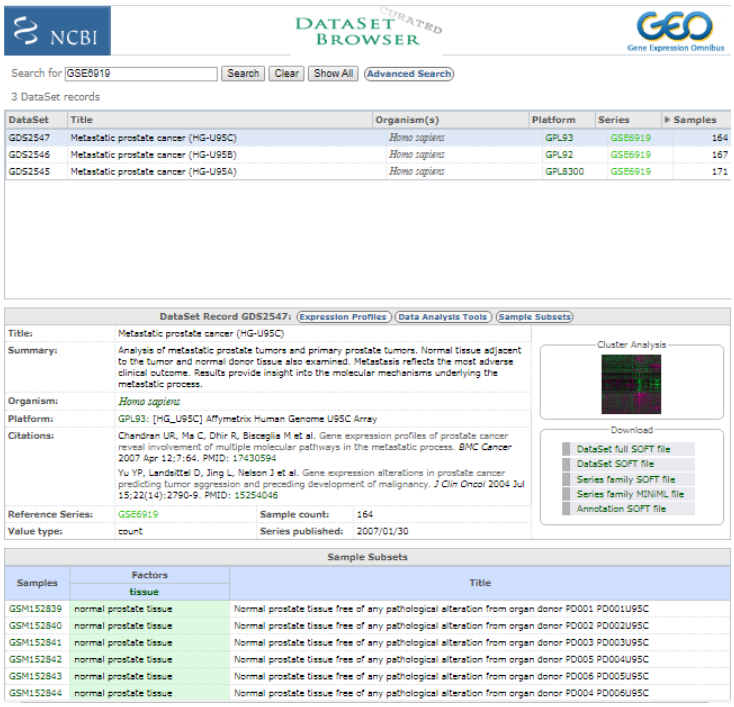


FIGURE 4.2 – Capture de site Datasets GEO

Les informations détaillées sur les jeux de données sont répertoriées dans le tableau 4.1

	Jeux des données			
Datasets	**Echantillons**	**Genes**	**Classes**	**Référence**
Cancer de prostate	164	12 580	4	GSE6919
Cancer de colon	50	25 800	2	GSE5232
PBMC	140	7 730	3	GSE33463

TABLE 4.1 – Les informations détaillées sur les jeux de données utilisées

4.3.2 Le prétraitement des données biopuces

dans ce processus contiens deux étapes principales : La correction du bruit du fond La première étape dans notre prétraitement est la correction du bruit du fond. Les bases des données GEO sont stockées avec correction du bruit du fond.

4.3.3 Normalisation

La normalisation est un autre concept important nécessaire pour changer toutes les caractéristiques à la même échelle.

Cela permet une convergence plus rapide sur l'apprentissage et une influence plus uniforme pour tous les poids.

Nos bases des données GEO téléchargés peut être déjà passé par la normalisation, sinon nous appliquerons les normalisé.

4.3.4 sélection des gènes

Pour résoudre le difficile problème de la sélection de gènes distingués dans les jeux de données sur l'expression des gènes du cancer.

Nous avons utilisé un certain nombre de méthodes et techniques pour réduire le nombre de profils d'expression à un sous-ensemble contenant les gènes les plus significatifs.

algorithme de sélection des sous-ensembles de gènes basé sur :

— SelectKBEST.

— SelectPercentile.

4.3.5 Classifier les données

En divisée La base (Matrice) en deux parties : une partie d'entrainement(apprentissage) et une partie de test. Le modèle est construit à partir de la partie d'entrainement et testé sur la partie de test. Si le modèle est acceptable, il est utilisé sinon il est révisé en essayant un autre jeu de paramètres (tunning). La figure suivante illustre le fonctionnement de cette étape.

Nous avons adopté des K les plus proche voisins (K-ppv) comme outil de classification.

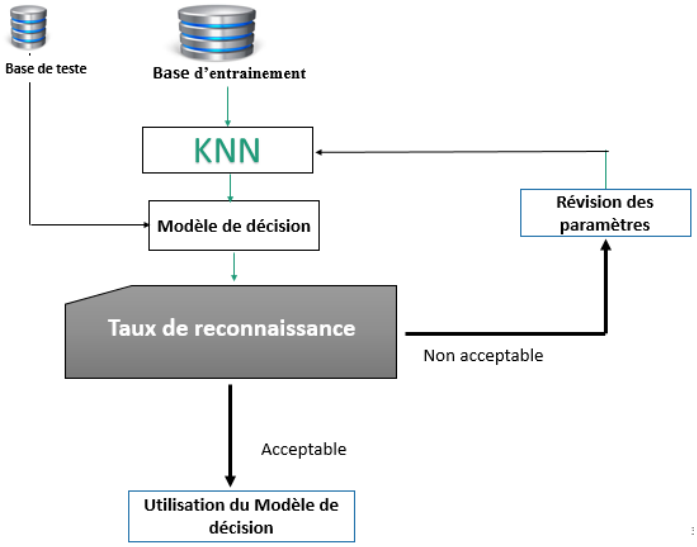


FIGURE 4.3 – Construction du modèle de décision

4.4 Realisation

4.4.1 L'environnement de travail :

Notre système est développé en utilisant le matériel suivant :

— Ordinateur portable : Intel(R) Cor(TM) i5 CPU @ 1.70 GHz, RAM (4Go)

— système d'exploitation : Windows7

4.4.1.1 Python

C'est un langage multiparadigme et multiplateforme, sa syntaxe simple et facile à apprendre Python est un langage de programmation puissant et facile à apprendre. Il nous fait gagner du temps, L'interpréteur Python et sa vaste bibliothèque standard sont disponibles librement. C'est pourquoi on optera pour python.

4.4.1.2 PyCharm 2018.1.3

est un environnement de développement intégré (IDE) utilisé dans la programmation informatique, spécifiquement pour le langage Python. Il est développé par la société tchèque JetBrains. [2] Il fournit une analyse de code, un débogueur graphique, un testeur d'unité intégré, l'intégration avec des systèmes de contrôle de version (VCS) et prend en charge le développement Web avec Django [22] .

4.4.1.3 Présentation l'outil Weka 3.8

Weka(Waikato Environment for Knowledge Analysis) est un ensemble d'outils permettant de manipuler et d'analyser des fichiers de données, publié sous la licence GNU. Il se compose d'une collection d'algorithmes d'apprentissage pour les tâches data mining, dont les arbres de décision, les réseaux de neurones , les SVM et les KNN.

Weka contient des outils pour le prétraitement des données, classification, régression, clustering, règles d'association, et la visualisation. Il est aussi bien adapté pour le développement de nouveaux programmes d'apprentissage

4.4.2 Bibliothèques utilisées

— Pandas.py : cette bibliothèque permet la manipulation et l'analyse des données. Elle propose en particulier des structures de données et des opérations de manipulation de tableaux numériques et de séries temporelles.
— Numpy.py : est un module de calcul scientifique permettant d'intégrer des matrices ou des tableaux à N-dimensions à Python tout en offrant une série d'opérateurs classiques de haut niveau sur ces derniers.
— Matplotlib.py : Matplotlib est une bibliothèque de traçage Python 2D qui produit des figures de qualité de publication dans une variété de formats papier et d'environnements interactifs entre plates-formes.
— Scikit-learn :
Machine Learning dans Python qui contient : une large gamme d'algorithmes d'apprentissage machine, Des outils simples et efficaces pour l'exploration de données et l'analyse de données, Accessible à tous et réutilisable dans différents contextes, Construit sur NumPy, SciPy, et matplotlib, Open source, commercialement utilisable - licence BSD.

4.4.3 Construction du Modèle De Classification

1 lecture des données nous avons diviséla notre ensemble de données en ses attributs et Classe(étiquettes)

— On commence par lire la base d'apprentissage à l'aide de la bibliothèque 'Pandas' pour charger et analyser le fichier de données.
data = pd.read _ csv("namefile", sep=';')

— On commence par lire fichier description la base par :
fichier = open("namefile.txt", "r")

2 Divisé la base(traing-test) Nous avant diviser chaque bases de données en deux parties inégales grace à la bibliothèque 'sklearn.model_selection' de divisé en apprentissage et tests par l'instruction suivante :

X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.33)
— 67‰des données seront utilisé pour d'apprentissage(training)

— 33‰des données seront utilisé pour le test (test)

3 Normalisation Pour la normalisation, nous utilisons le **StandardScaler()**, qui change toutes les fonctionnalités entre 0 et 1 :

$$X'_i = \frac{(X_i - X_i^{min})}{(X_i^{max} - X_i^{min})} \quad (4.1)$$

4 Sélection et Réduire des Gènes

1. Sélection des gènes avec **F-test** pour réduire les caractéristiques Nous utilisons cette fonction de sélection par défaut : les 10‰les plus significatifs.
   selector_Genes = SelectPercentile(f_classif, percentile=10).
2. Fonction **SelectKBest** prenant deux tableaux X et y, et renvoyant une paire de tableaux (scores, pvalues) ou un tableau avec des scores. **K** : int ou "all", optionnel, par défaut = 10 Nombre de fonctionnalités principales à sélectionner
   selector_genes = SelectKBest(score_func=f_classif, k=5)

5 Evaluation du Modèle de Classification) Nous avons adopté des K les plus proche voisins (K-ppv) comme outil de classification. et utilisé les critères de précision pour évaluer la performance des classificateurs sur les sous-ensembles de gènes sélectionnés.

nous avons exécutées une boucle de 1 à 40. Dans chaque itération,calculées l'erreur moyenne pour les valeurs prédites. Et en suite nous tracerons dans la diagramme suivant les valeurs d'erreur par rapport aux valeurs K grâce à la bibliothèque 'Matplotlib'.

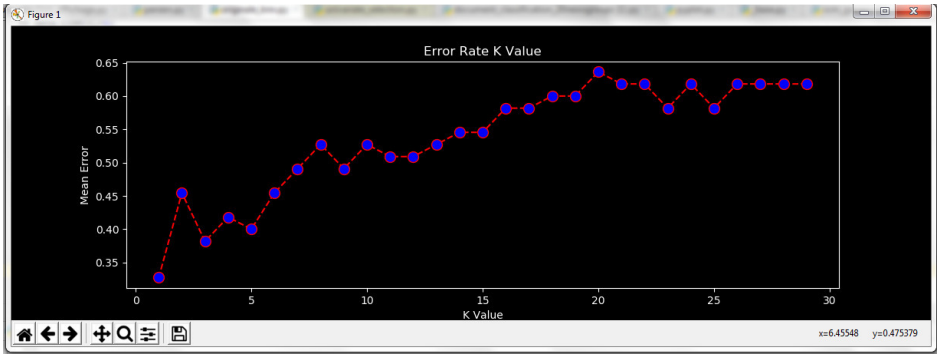


FIGURE 4.4 – Comparaison du taux d'erreur avec la valeur K

Discussion le diagramme De la sortie, nous pouvons voir le moins l'erreur moyenne correspondant la valeur K =3. Alors La valeur K=3 c'est la meilleur valeur qui donne la meilleur de la précision de prédiction sur la base**'cancer_Prostate'** .

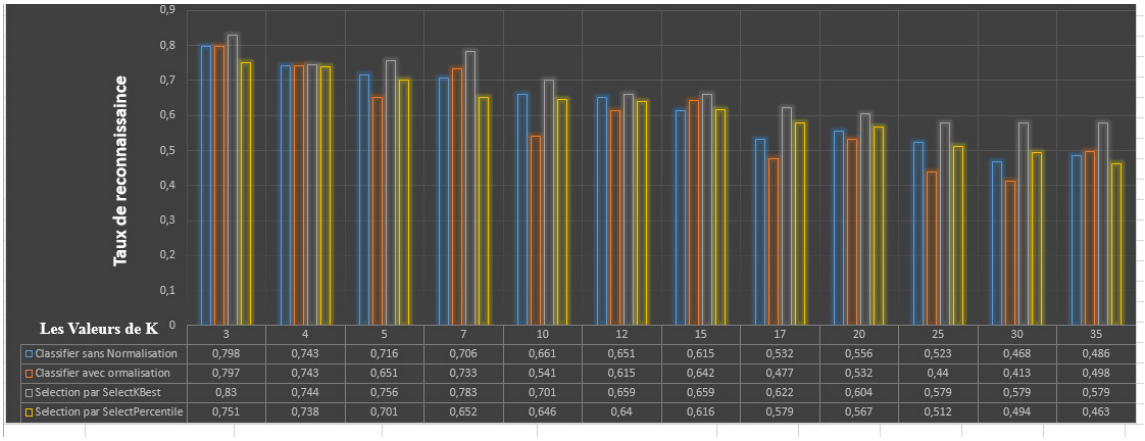
4.4.4 Résultats et discussions

On applique la base de maladie cancer_Prostate,et nous avant représentées les résultant sur diagramme qui exprimé :

Les taux de reconnaissance 'TR' tel que c'est le rapport entre le nombre d'exemples correctement classés et le nombre total d'exemples testés.

— 'Taux de reconnaissance' sans normalisation et avec la normalisation.
— 'Taux de reconnaissance ' avec la normalisation
— Et aussi 'taux de reconnaissance' avec la sélection par les deux méthode (SelectKBest , SelectPercentile) .

La résultat résume dans le diagramme suivant :



Les Valeurs de K	3	4	5	7	10	12	15	17	20	25	30	35
Classifier sans Normalisation	0,798	0,743	0,716	0,706	0,661	0,651	0,615	0,532	0,556	0,523	0,468	0,486
Classifier avec ormalisation	0,797	0,743	0,651	0,733	0,541	0,615	0,642	0,477	0,532	0,44	0,413	0,498
Selection par SelectKBest	0,83	0,744	0,756	0,783	0,701	0,659	0,659	0,622	0,604	0,579	0,579	0,579
Selection par SelectPercentile	0,751	0,738	0,701	0,652	0,646	0,64	0,616	0,579	0,567	0,512	0,494	0,463

FIGURE 4.5 – Evolution du taux de reconnaissance .

4.4.4.1 Discussion le diagramme :

— chaque itération de calcule l'évaluation nous remarquons :

1. La normalisation étape principe sous la précision de prédiction.
2. La sélection des attributs(gènes) par la méthode **selectKBest** augmenté la précision de classification.
3. Meilleure performance en classification avec le sélection des attributs. Alors Apprentissage moins coûteux et plus Facilite l'apprentissage,

— Le résultat de taux de reconnaissance correspondance avec Le taux d'erreur avec K.
— pour construire un modèle de décision. Les résultats en terme de taux de reconnaissance sont satisfaisants.

4.4.5 Weka

Weka contient des outils pour le prétraitement des données, classification, régression, clustering, règles d'association, et la visualisation. Il est aussi bien adapté pour le développement de nouveaux programmes d'apprentissage. Nous avons utilisées différentes méthode pour comparer les plus efficace :

4.4.5.1 Fichier Database de Weka

Fichier CSV : Sauvegarder le fichier de données d'apprentissage et de test au niveau de l'outil Excel sous la forme .csv (comma separated value).

Fichier ARRF : Un fichier ARFF (Attribute-Relation File Format) est un fichier texte ASCII qui décrit une liste d'instances partageant un ensemble d'attributs [37] . sur structure suivante :

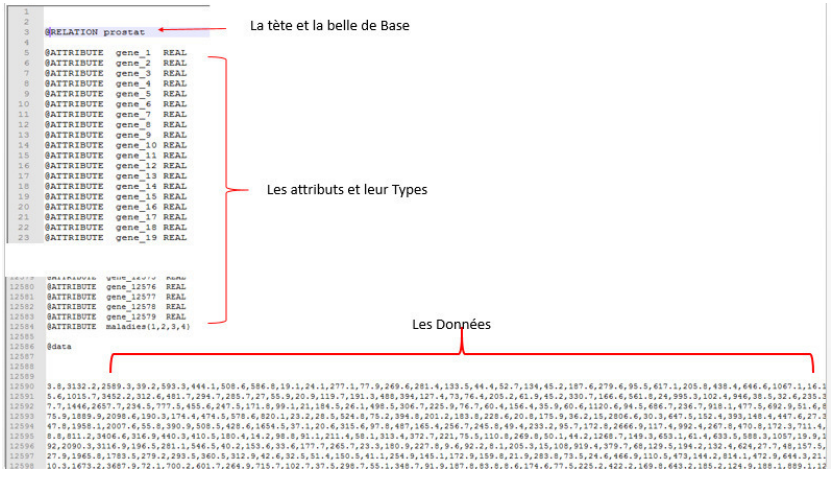


FIGURE 4.6 – La forme de fichier ".arff"

l'expérimentation le fichier par WEKA :

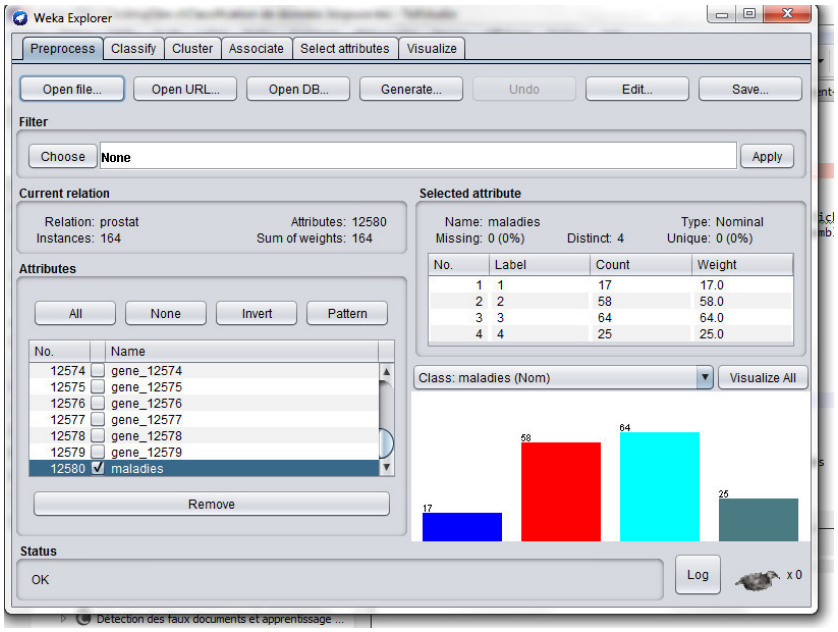


FIGURE 4.7 – exploration fichier .arrf

Cette interface permet de :
— Charger les deux fichiers d'entrainement et de test.
— Modifier les paramètres d'algorithme(si nécessaire).
— Faire l'apprentissage.
— Prédire la classe en utilisant les données de test.
— Afficher le résultat de prédiction.
— Enregistrer le modèle.

Présentation des Tests : Trois méthodes de tests ont été utilisées :
— Using Training Set : Le classificateur est évalué sur la manière dont sont prédits les classes des instances testées.
— Cross Validation : Le classificateur est évalué par validation croisée.
— Percentage Split : Le classificateur est évalué sur la manière dont est prédit un certain pourcentage des données.

Pour nous aider à visualiser les résultats de classification Weka fourni un **rapport d'analyse** après chaque phase d'apprentissage et de validation.

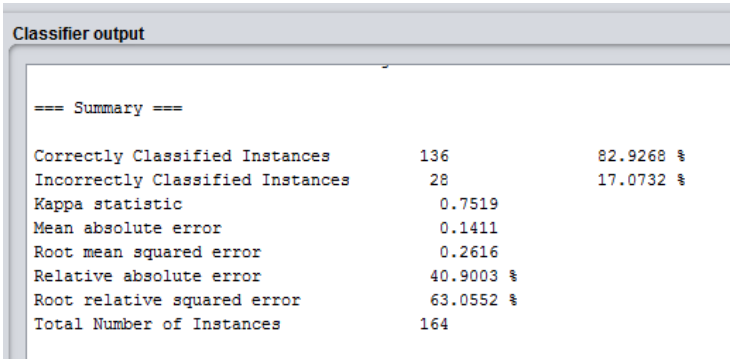


FIGURE 4.8 – Le rapport d'analyse Weka

Dans ce rapport la donnée qui nous intéresse est le pourcentage d'instances correctement classifiées. C'est la quantité que nous tenterons d'optimiser au cours de ce projet. Dans cet exemple, 82.9268 ‰des instances ont été correctement classifiées, c'est-à-dire parmi 164 échantillons il y a 136 échantillons sont reçus correctement.

On Applique Les tests des algorithme d'apprentissage Supervisée sur notre base **cancer_Prostate** avec les réglage d'option par défaut avec**Using Trainig** :

— IBK : l'option 'nombre de voisins' modifiée en valeurs entre [3,50].
— SMO : l'option 'valeurs de gamma' modifiée en valeurs entre [0.1,5].
— j48 : l'option 'MinNumObj' modifiée en valeurs entre [2,40].

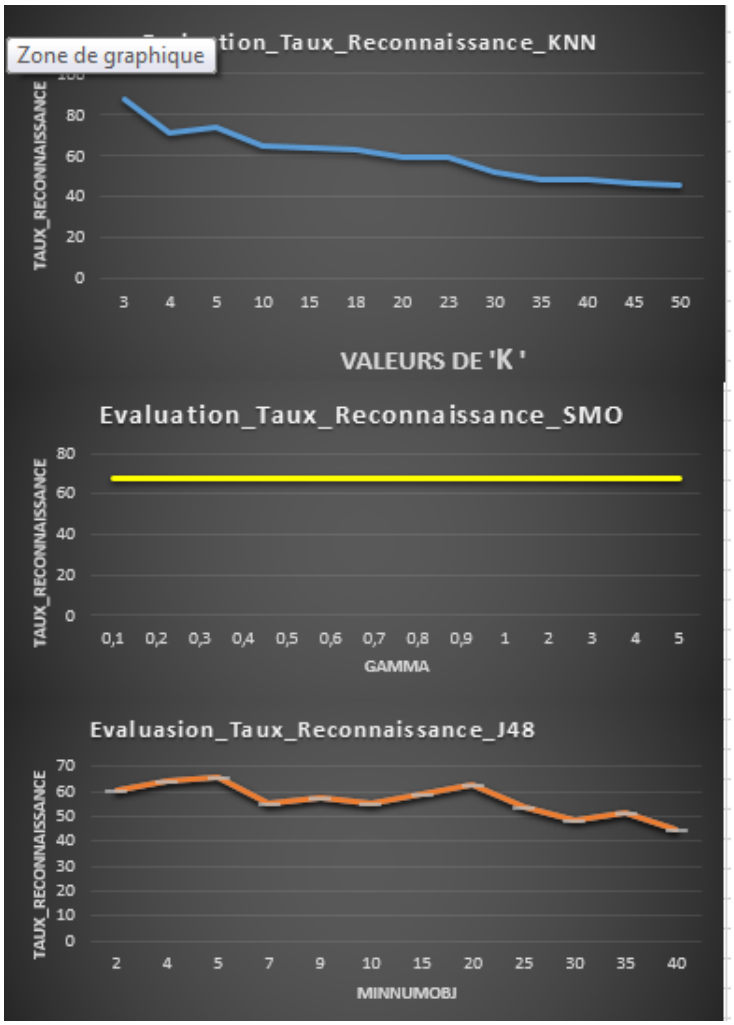


FIGURE 4.9 – Evaluation les 3 méthodes de classification KNN,SVM, Tree J48

4.4.5.2 Discussion des résultat :

Ceci a pour but une étude comparative des algorithme de classification Arbre de décision J48, et Séparateurs a Vaste Marge SVM et Les K plus proche voisins K-ppv
Les jeux de données Classifiées sont cancer_Prostate :
Pour nous aider à visualiser les résultats de classification Weka fourni un rapport d'analyse après chaque phase d'apprentissage et de validation.
Nous avant remarquées :
— En remarque le diagramme KNN et SVM sont perturbés qui in-

diqués Les méthodes effectues aveec notre base par à apport la méthode de J48 .

— Evaluation TR sous diagramme de SVM sont stables avec la pourcentage 67,8571‰c'est suffisant .
— Evaluation de TR sous diagramme de J48 le performances généralement acceptable tel que TR=60,7143‰avec gamma=2
— Les excellents résultats du jeux de donnés cancer_Prostate sont L'algorithme KNN avec la valeur de k=3 avec TR=28.9268 ‰.

La classification est une tâche très importante dans le data mining, et qui consomme beaucoup de recherches pour son optimisation (Normalisation et La sélection les attributs sont des méthodes de amélioration la précision de prédiction)
Les excellents résultats du jeux de donnés cancer_Prostate sont L'algorithme KNN avec la valeur de k=3 .

```
=== Summary ===

Correctly Classified Instances         136               82.9268 %
Incorrectly Classified Instances        28               17.0732 %
Kappa statistic                          0.7519
Mean absolute error                      0.1411
Root mean squared error                  0.2616
Relative absolute error                 40.9003 %
Root relative squared error             63.0552 %
Total Number of Instances              164
```

FIGURE 4.10 – rapport d'analyse de résultat méthode 3NN

4.5 Synthèse du chapitre

Nous avons représenté l'implémentation de notre système : L'environnement et les outils de développement. Ensuite nous avons présenté quelques expérimentations effectuées sur la base de données de maladie et présenté les résultats obtenus enter les différant Algorithmes de classification et son diagrammes de évaluation taux de reconnaissance.

CONCLUSION GÉNÉRALE

A travers ce projet d'étude, nous avons découvert le domaine de la Bioinformatique, son objet, ses méthodes et ses outils .

Pour remédier ce problème, Nous avons proposé une méthode basée sur l'apprentissage automatique pour construire notre système de diagnostic par des méthodes simples et rapides. Avec l'utilisation croissante des Technologie puce à ADN nous avons classifié cette données par la méthode de d'apprentissage KNN .

KNN est un algorithme de classification simple mais puissant. Il ne nécessite aucune formation pour faire des prédictions, ce qui est généralement l'une des parties les plus difficiles d'un algorithme d'apprentissage automatique.

WEKA c'est une outils permettant de manipuler et d'analyser des fichiers de données. Il se compose d'une collection d'algorithmes d'apprentissage pour les tâches data mining, nous avons appliquées les arbres de décision, les SVM et les KNN.

Le système que nous proposons utilise l'apprentissage automatique supervisé à partir des données de **biopuce**, pour construire un modèle de décision. Les résultats en terme de taux de reconnaissance sont satisfaisants.
Pour les perspectives et les travaux futur, nous proposons des idées qui peuvent améliorer notre système de **classification des données biopuces**, telles que :

— Etudier et réduire nombre de gènes au niveau du laboratoire pour améliorer les performances du système de précision de prédiction.
— Extraction caractéristiques les plus significatif pour limiter les étapes difficile comme le filtrage et sélection attributs.
— Stocker les données biopuce sous forme plus fonctionnelles pour diriger à la classification des méthodes simples et rapides comme (K-PPV).

BIBLIOGRAPHIE

[1] Atsushi Akane, Kazuo Matsubara, Hiroaki Nakamura, Setsunori Takahashi, and K Kimura. Identification of the heme compound copurified with deoxyribonucleic acid (dna) from bloodstains, a major inhibitor of polymerase chain reaction (pcr) amplification. *Journal of Forensic Science*, 39(2) :362–372, 1994.

[2] AI Alpaydin. Neural models of incremental supervised and unsupervised learning. 1990.

[3] DESeq Anders. Huber, 2010 https ://www. bioconductor. org/packages/release/bioc/html. *DESeq. html.*

[4] Tanya Barrett, Dennis B Troup, Stephen E Wilhite, Pierre Ledoux, Dmitry Rudnev, Carlos Evangelista, Irene F Kim, Alexandra Soboleva, Maxim Tomashevsky, and Ron Edgar. Ncbi geo : mining tens of millions of expression profiles—database and tools update. *Nucleic acids research*, 35(suppl_1) :D760–D765, 2006.

[5] François Bertucci, Béatrice Loriod, Rebecca Tagett, Samuel Granjeaud, Daniel Birnbaum, Catherine Nguyen, and Rémi Houlgatte. Puces à adn : technologie et applications. *Bulletin du cancer*, 88(3) :243–52, 2001.

[6] Philippe Bordron. *Analyse des systèmes bactériens : une approche in silico pour intégrer les connaissances du vivant.* PhD thesis, Université de Nantes, 2012.

[7] Amina BOUBLENZA. *Coopération entre classifieurs hétérogènes pour la reconnaissance des données médicales.* PhD thesis, 07-05-2017, 2017.

[8] Mohamed-Rafik Bouguelia. *Classification et apprentissage actif à partir d'un flux de données évolutif en présence d'étiquetage incertain.* PhD thesis, Université de Lorraine, 2015.

[9] NCBI Resource Coordinators. Database resources of the national center for biotechnology information. *Nucleic acids research*, 44(Database issue) :D7, 2016.

[10] Antoine Cornuéjols and Laurent Miclet. *Apprentissage artificiel : concepts et algorithmes.* Editions Eyrolles, 2011.

[11] AJ COZZONE, P DESSEN, C GAUTIER, D KAHN, B LABEDAN, and JL RISLER. Bases de données et outils d'analyse pour la génomique bactérienne. 2000.

[12] Frédéric Dardel and François Képès. *Bioinformatique : Génomique et post-génomique.* Editions Ecole Polytechnique, 2002.

[13] David Dernoncourt. *Stabilité de la sélection de variables sur des données haute dimension : une application à l'expression génique.* PhD thesis, Université Pierre et Marie Curie-Paris VI, 2014.

[14] Abdelhamid Djeffal. Cours fouille de données avancée. *Université Mohamed Khider Biskra*, 2015.

[15] Steffen Durinck, Yves Moreau, Arek Kasprzyk, Sean Davis, Bart De Moor, Alvis Brazma, and Wolfgang Huber. Biomart and bioconductor : a powerful link between biological databases and microarray data analysis. *Bioinformatics*, 21(16) :3439–3440, 2005.

[16] Ron Edgar, Michael Domrachev, and Alex E Lash. Gene expression omnibus : Ncbi gene expression and hybridization array data repository. *Nucleic acids research*, 30(1) :207–210, 2002.

[17] MB Eisen. Scanalyze, 1999.

[18] L Greillier, P Roll, F Barlesi, A Robaglia-Schlupp, A Fraticelli, P Cau, and P Astoul. Apport des puces à adn dans le diagnostic étiologique des pleurésies : étude de faisabilité. *Revue des Maladies Respiratoires*, 24(7) :859–867, 2007.

[19] Abdelillah HASSAM, Ismael Abraham OUATTARA, Mme Lynda ZAOUI, Mlle Khadija HENNI, and Mr Rahal SD. Construction d'un workflow d'analyse de données issues de puces à adn. *Gene Expression*, 1 :6, 2014.

[20] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. Unsupervised learning. In *The elements of statistical learning*, pages 485–585. Springer, 2009.

[21] José Crispin Hernandez Hernandez. *Algorithmes Métaheuristiques hybrides pour la sélection de gênes et la classification de données de biopuces.* PhD thesis, Université d'Angers, 2008.

[22] Quazi Nafiul Islam. *Mastering PyCharm.* Packt Publishing Ltd, 2015.

[23] Oliver Kramer. K-nearest neighbors. In *Dimensionality Reduction with Unsupervised Nearest Neighbors*, pages 13–23. Springer, 2013.

[24] Oliver Kramer. Scikit-learn. In *Machine Learning for Evolution Strategies*, pages 45–53. Springer, 2016.

[25] Nolwenn Le Meur. *De l'acquisition des données de puces à ADN vers leur interprétation : importance du traitement des données primaires.* PhD thesis, Université de Nantes. Unité de Formation et de Recherche de Médecine et des Techniques Médicales, 2005.

[26] Frank Meyer. Recommender systems in industrial contexts. *arXiv preprint arXiv :1203.4487*, 2012.

[27] Omar MOUSSATI. *Classification des données de biopuces.* PhD thesis, Université des Sciences et de la Technologie d'Oran Mohamed Boudiaf, 2016.

[28] Hof Nathalie. decouverte sciences et environnement -qu-est-ce-que-la-bioinformatiquee. *Thèse d'université en Sciences, Strasbourg No. 266*, page 165, 2013.

[29] H Parkinson, Misha Kapushesky, Mohammadreza Shojatalab, Niran Abeygunawardena, Richard Coulson, Anna Farne, Ele Holloway, N Kolesnykov, P Lilja, Margus Lukk, et al. Arrayexpress—a public database of microarray experiments and gene expression profiles. *Nucleic acids research*, 35(suppl_1) :D747–D750, 2006.

[30] Julie Peyre. *Analyse statistique des données issues des biopuces à ADN.* PhD thesis, Université Joseph-Fourier-Grenoble I, 2005.

[31] Yann Ponty. *Etudes combinatoire et génération aléatoire des structures secondaires d'ARN.* PhD thesis, Université Paris-Sud, 2003.

[32] Paul Rabinow. *Making PCR : A story of biotechnology.* University of Chicago Press, 2011.

[33] Nancie Reymond. *Bioinformatique des puces à ADN et application à l'analyse du transcriptome de Buchnera aphidicola*. PhD thesis, INSA de Lyon, 2004.

[34] Christophe Ronsin. *L'histoire de la biologie moléculaire : pionniers & héros*. De Boeck Supérieur, 2005.

[35] Qiang Su, Yina Wang, Xiaobing Jiang, Fuxue Chen, and Wencong Lu. A cancer gene selection algorithm based on the ks test and cfs. *BioMed research international*, 2017, 2017.

[36] Dr Colette Vendrely. *Colette Vendrely,... L'Acide désoxyribonucléique du noyau des cellules animales, son rôle possible dans la biochimie de l'hérédité...* Éditions du" Bulletin biologique de la France et de la Belgique, 1952.

[37] Ian H Witten and Eibe Frank. Weka. *Machine Learning Algorithms in Java*, pages 265–320, 2000.

[38] Sang Hwa Yang, Jong Sik Kim, Tae Jeong Oh, Myung Soon Kim, Sun Woo Lee, Suk Kyung Woo, Hyun Sill Cho, Yung Hyun Choi, Young Ho Kim, Sun Young Rha, et al. Genome-scale analysis of resveratrol-induced gene expression profile in human ovarian cancer cells using a cdna microarray. *International journal of oncology*, 22(4) :741–750, 2003.