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

N° d'ordre : SIOD 2/M2/2018

Mémoire

Présenté pour obtenir le diplôme de master académique en

Informatique

Parcours : **système d'information d'optimisation et de décision (SIOD)**

Pré-traitement de données de biopuces et la sélection de gènes par une approche bio-inspirée

---

Par :

SELMI AYMEN TAKIE EDDINE

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

Cherif Foudil	Professeur	Président
Djerou Leila	M.C.A	Rapporteur
Ben Seghir Nadia	M.C.B	Examinateur

Dédicace

Tous les mots ne sauraient exprimer la gratitude, l'amour, le respect, la reconnaissance, c'est tous simplement que je dédie ce modeste travail à :

À ma chère mère, Tu représentes pour moi le symbole de la bonté par excellence, la source de tendresse et l'exemple du dévouement qui n'a pas cessé de m'encourager et de prier pour moi. Ta prière et ta bénédiction m'ont été d'un grand secours pour mener à bien mes études.

À mon chèr père,Aucune dédicace ne saurait exprimer l'amour, l'estime, le dévouement et le respect que j'ai toujours eu pour vous. Ce travail est le fruit de tes sacrifices que tu as consentis pour mon éducation et ma formation.

À mon très cher frère RAMI qui présent dans tous mes moments d'examens par son soutien moral et ses belles surprises sucrées. Je te souhaite un avenir plein de joie, de bonheur, de réussite et de sérénité. Je t'exprime à travers ce travail mes sentiments de fraternité et d'amour.

À mon ami Younes.

À tous mes collègues de l'Université de biskra .

Remerciement

Tout d'abord, je remercie le Dieu, notre créateur de m'avoir donné la force, la volonté et le courage afin d'accomplir ce travail modeste.

Je veux adresser mes remerciements les plus sincères à mon encadreur madame DJEROU Leila qui a proposée le thème de ce mémoire, pour ses conseils et ses dirigées du début à la fin de ce travail pour l'attention et la disponibilité dont elle a su faire preuve au cours de la réalisation de ce mémoire.

Je tiens également à remercier messieurs les membres de jury pour l'honneur qu'ils m'ont fait en acceptant de siéger à notre soutenance.

J'ai une attention particulière pour Mademoiselle SFAKSI Sara, Doctorante à l'université de Biskra, pour sa précieuse aide, ses remarques et ses conseils pertinents.

Finalement, je tiens à exprimer ma profonde gratitude à ma famille qui m'a soutenue et encouragée tout au long de mes études.Ainsi que l'ensemble des enseignants qui ont contribué à ma formation.

RÉSUMÉ

Le diagnostic médical est un élément très important dans le domaine de reconnaissance et traitement des maladies. La biopuce (puce à Adn) est l'une des techniques modernes qui nous aide à faire le diagnostic. Elle permet d'étudier simultanément le niveau d'expression de plusieurs milliers de gènes ou groupe de gènes dans des conditions différentes (physiologique ou pathologique).

Les données issues des biopuces sont obtenues à partir d'un protocole complexe où plusieurs étapes peuvent introduire du bruit dans les données. De plus ces données sont décrites par des milliers d'attributs (gènes). Afin d'appliquer efficacement les techniques d'analyse ou de classification des gènes, il est nécessaire de réduire la dimension des données en sélectionnant des sous-ensembles de gènes les plus intéressants qui garantissent une bonne performance en classification.

Ce travail sert à exploiter les possibilités offertes par les approches bio-inspirées pour le pré-traitement des données issus des biopuces et la sélection des gènes.

Mots clés : Métaheuristiques, optimisation combinatoire, bioinformatique, microréseau d'ADN, sélection des gènes,approches bio-inspirées.

ملخص

التشخيص الطبي هو عنصر مهم جدا في مجال التعرف على الأمراض وعلاجها. تعتبر الرقاقة الحيوية واحدة من التقنيات الحديثة التي تساعدنا على إجراء التشخيص. تجعل من الممكن دراسة مستوى التعبير عن عدة آلاف من الجينات أو مجموعة من الجينات في الوقت نفسه في ظل ظروف مختلفة (الفسيولوجية أو المرضية).

يتم الحصول على البيانات من الرقائق الحيوية عن طريق بروتوكول معقد حيث يمكن لعدة خطوات إدخال ضوضاء في البيانات. بالإضافة إلى ذلك ، يتم وصف هذه البيانات من قبل الآلاف من الصفات (الجينات). من أجل تطبيق تقنيات التحليل الجيني أو تصنيف الجينات بشكل فعال ، من الضروري تقليل حجم البيانات عن طريق اختيار مجموعات فرعية من الجينات الأكثراهمية والتي تضمن أداء تصنيف جيد.

يهدف هذا العمل إلى استغلال الإمكانيات التي توفرها الأساليب المستوحاة من البيولوجيا لمعالجة البيانات الحيوية المسبقة واختيار الجينات.

كلمات مفتاحية: ميتاهيوريستيک ، الاستمثال التوافقي ، المعلوماتية الحيوية ، رقاقة الحمض النووي ، انتقاء الجينات ، النهج المستوحاة من الحيوية

ABSTRACT

Medical diagnosis is a very important element in the field of recognition and treatment of diseases. The biochip (DNA chip) is one of the modern techniques that helps us to make the diagnosis. It makes it possible to simultaneously study the level of expression of several thousand genes or group of genes under different conditions (physiological or pathological).

The data from biochips are obtained from a complex protocol where several steps can introduce noise into the data. In addition, these data are described by thousands of attributes (genes). In order to effectively apply gene analysis or classification techniques, it is necessary to reduce the size of the data by selecting the most interesting subsets of genes that guarantee a good classification performance.
This work aims to exploit the possibilities offered by bio-inspired approaches for preprocessing biochip data and gene selection.

key words : Metaheuristics, combinatorial optimization, bioinformatics, DNA microarray, gene selection, bio-inspired approaches.

TABLE DES MATIÈRES

TABLE DES FIGURES

LISTE DES TABLEAUX

Liste d'Abréviation

ADN	Acide DésoxyriboNucléique.
ARN	Acide ribonucléique.
CEL	Cell Intensity File. fichier d'intensité de cellule.
CDF	Chip description file. fichier de description de puce.
RMA	Robust Multi-array Average. moyenne multi-matrice robuste.
GP	Genetic Programming. Programmation Génetique.
MGGP	Multi-Gene Genetic Programming. Programmation Génetique Multi-Gènes.
RMSE	Root Mean Square Error . racine carrée de l'erreur quadratique moyenne.
R^2	Coefficient de détermination.
MSE	Mean squared error . Erreur quadratique moyenne.
SSE	Sum of squared errors . somme d'erreurs au carré.
MAE	Mean absolute error . L'erreur absolue moyenne.
MAXE	Max error. Erreur Max.

INTRODUCTION GÉNÉRALE

Ces dernières années, les données biologiques ont évolué quantitativement grâce au développement de nouveaux matériaux et techniques pour comprendre l'ADN et d'autres composants des organismes vivants. Les scientifiques se sont tournés vers les outils bioinformatiques pour une meilleure capacité de stockage et d'analyse des données.

La technologie des puces à ADN connaît actuellement une croissance exceptionnelle et suscite un intérêt considérable dans la communauté scientifique en raison de son potentiel à mesurer simultanément le niveau d'expression d'un grand nombre de gènes dans des échantillons de tissus [13]. L'analyse des données de Microarrays peut être utilisée pour identifier de nouveaux médicaments et outils diagnostiques, qui représentent un défi majeur pour l'exploration de données dans différents domaines tels que le regroupement de gènes, la classification des échantillons et la sélection des gènes.

Typiquement, l'ensemble de données de puces à ADN produit des valeurs d'expression pour un petit nombre d'échantillons (habituellement moins de 100) avec plusieurs milliers de gènes, avec seulement un plus petit nombre d'entre eux montre une grande corrélation avec les cas d'étude [14]. Dans une tâche de classification, ces conditions peuvent conduire à surapprentissage, ce qui signifie qu'un classificateur peut facilement montrer la performance d'une fonction de décision qui se comporte très bien avec les données d'apprentissage mais très mal dans les données de test [13] . De plus, pour améliorer la validité du classificateur, il est nécessaire de réduire la dimensionnalité des données en sélectionnant un sous-ensemble de gènes pertinents pour la classification. Ce problème peut être défini comme un problème d'optimisation combinatoire qui consiste à rechercher un sous-ensemble P, de variables (gènes) ayant la plus grande puissance de classification, parmi les N variables disponibles, qui minimise une fonction objective (la corrélation entre les variables) .

Dans ce travail, on a exploité les capacités de la technique du programmation génétique

multigène (MGGP) afin de résoudre le problème de sélection des gènes. Cette technique inspiré de la nature de la théorie de l'évolution de Darwin est défini comme une méthode de modélisation de système non-linéaire qui intègre la programmation génétique standard et les capacités de régression classique.

Ce mémoire comporte quatre chapitres organisés de la manière suivante :

— le premier chapitre est consacré aux concepts et notions biologique, la définition de la bio-informatique, ainsi que, les technologies biopuces, ses phases d'analyse, ses types et ses domaines d'application.

— Le deuxième chapitre expose les approches de sélection des gènes, ses principes et processus de sélection, ainsi que les travaux existants sur leur utilisations au domaine des puces à ADN.

— Le troisième chapitre présente la méthode proposée pour la sélection des gènes, puis, la conception du système à réaliser et son architecture globale et détaillée.

— L'implémentation du système et les expérimentations et les résultats sont abordés dans le quatrième et le cinquième chapitre.

Le mémoire est terminé par une conclusion générale avec les perspectives envisagées.

CHAPITRE

1

BIO-INFORMATIQUE ET PUCES A ADN

1.1 Introduction

Les technologies de puces à ADN dans leur ensemble fournissent de nouveaux outils qui transforment la façon dont les expériences scientifiques sont menées. L'avantage principal des technologies de puces à ADN par rapport aux méthodes traditionnelles est celui de l'échelle : au lieu de mener des expériences basées sur les résultats d'un ou de quelques gènes, les puces à ADN permettent l'interrogation simultanée de centaines ou de milliers de gènes. Dans ce chapitre, nous allons définir cette technologie, ses phases d'analyse, ses types, ses domaines d'application.

1.2 Bio-informatique et génomique

1.2.1 La bio-informatique

Au cours de ces trente dernières années, la récolte de données en biologie a connu un boom quantitatif grâce notamment au développement de nouveaux moyens techniques servant à comprendre l'ADN et d'autres composants d'organismes vivants.

Pour analyser ces données, plus nombreuses et plus complexes aussi, les scientifiques se sont tournés vers les nouvelles technologies de l'information. L'immense capacité de stockage et d'analyse des données qu'offre l'informatique leur a permis de gagner en puissance pour leurs recherches. Et la rencontre entre la biologie et l'informatique, c'est ce qu'on appelle **la bioinformatique**.

Le terme bio-informatique 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 les systèmes biotique. elle est constituée par l'ensemble des concepts et des techniques nécessaires à l'interprétation informatique de l'information biologique [10].

Ses principaux champs d'applications sont :
— la bio-informatique des réseaux,
— la bio-informatique structurale,
— et la bio-informatique des séquences.

Nous nous intéresserons particulièrement à la bio-informatique des séquences qui traite de l'analyse de données issues de l'information génétique contenue dans la séquence de l'ADN ou dans celle des protéines qu'il code [10].

1.2.2 La génomique

La génomique est une discipline récente de la biologie qui s'intéresse à l'étude exhaustive des génomes et en particulier de l'ensemble des gènes, de leur disposition sur les chromosomes, de leur séquence, de leur fonction et de leur rôle [25].

Elle se décompose de deux branches que sont :

Génomique structurelle : qui s'intéresse à l'organisation et position des gènes, la taille du génome, la comparaison des génomes de différents organismes et le séquençage de l'ADN et analyse des séquences [26].

Génomique fonctionelle : qui s'intéresse à la fonction des gènes, l'analyse globale de l'expression génétique . . . etc [26].

1.2.2.1 Terminologie génomique

Afin de mieux appréhender les différentes notions qui seront abordées dans la suite du travail, nous essayons dans cette partie de présenter la terminologie de base de la génomique.

a) 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.

Elle est l'unité de base biologique chez les organismes les plus simples (procaryotes) tels que les bactéries, l'information génétique n'est pas compartimenté dans un noyau vrai mais est libre dans le cytoplasme. Pour les organismes les plus simples (eucaryotes),

l'information génétique est compartimentée dans un noyau (figure 1.1) [28].

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 [10].

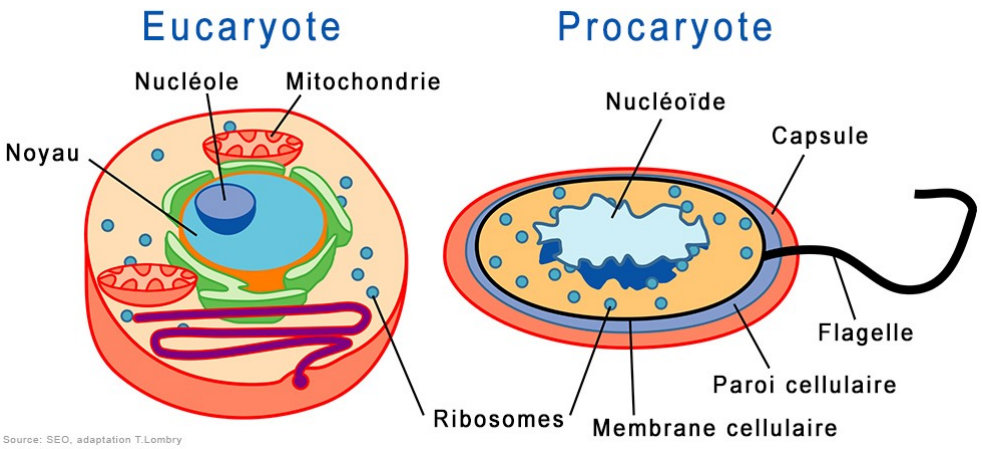


FIGURE 1.1: Composants d'une cellule eucaryote et cellule procaryote.

b) Acide désoxyribonucléique (ADN)

Les molécules d'ADN sont les plus grosses molécules du monde vivant et sont présentent dans tous les organismes vivants.

Une molécule d'ADN est une double hélice composée de deux brins enroulés l'un autour de l'autre On dit que l'ADN est bicaténaire (contrairement à l'ARN, qui est monocaténaire). Chaque brin d'ADN est composé d'une chaine de ddésoxyriboses sur laquelle sont attachés les nucléotides (A, T, C, G) qui codent l'information ; c'est un langage à quatre lettres [24].

Ces nucléotides peuvent être regroupés en deux paires de bases complémentaires par des liaisons hydrogène : A est toujours reliée à T et C à G. Cela signifie par exemple qu'un T est toujours en face d'un A sur l'autre brin. Les bases sont reliées entre elles à l'intérieur d'un brin d'ADN par des sucres des oses, appelés désoxyriboses, et par des acides phosphoriques (figure 1.2) [24].

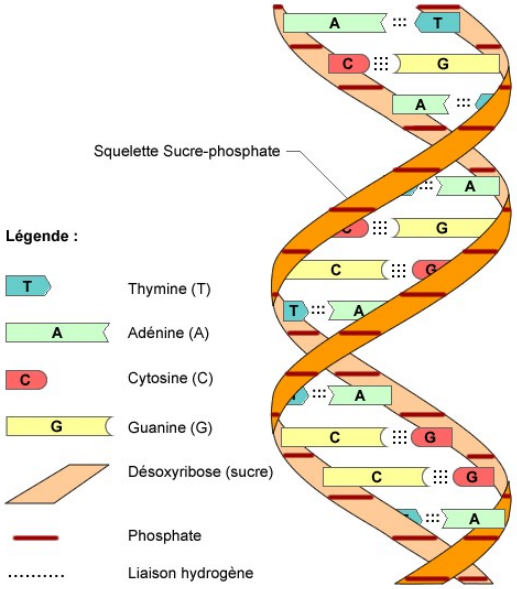


FIGURE 1.2: Structure de l'ADN.

c) Transcriptome

Le gène, unité de base de stockage de l'information génétique, est une petite séquence d'ADN. L'inventaire génique est défini comme l'ensemble de ces gènes, dont le nombre peut varier suivant l'individu, de 6200 chez la levure à une trentaine de milliers chez la souris ou l'homme [28]. L'ensemble du matériel génétique d'un individu ou d'une espèce encodé dans son ADN est appelé alors son génome.

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 [10] car il subira des maturations avant son transfert vers le cytoplasme (passage de membrane nucléaire) ou il sera traduit en protéines par les ribosomes [28].

La traduction interprète chaque triplet de nucléotides (codon) de l'ARN pré-messager en un acide aminé selon le code génétique universel.(figure 1.3)

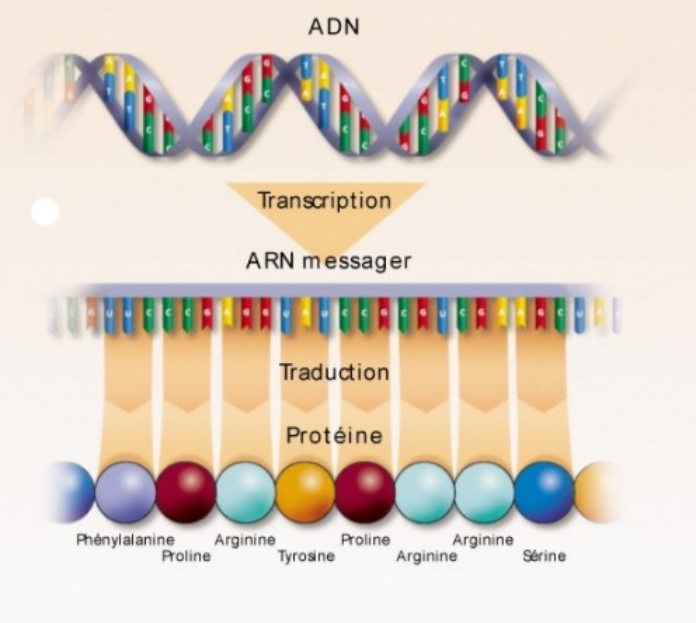


FIGURE 1.3: Synthèse de la protéine .

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 et offre d'autre part beaucoup d'intérêt en termes d'application potentielles [10].

Les méthodes d'analyse du transcriptome global reposent sur la technique des puces à ADN (Biopuces) car elle permet de l'étudier par l'observation simultanée de l'expression de plusieurs milliers de gènes dans une cellule ou un tissu donné, mesurant ainsi les modifications des différents états cellulaires.

1.3 Puce à ADN (DNA Microarray)

1.3.1 Définition et principe

Les puces à ADN reposent sur une technologie pluridisciplinaire intégrant la microélectronique, la chimie des acides nucléiques, l'analyse d'images et la bioinformatique [10]. Dans la définition la plus large, les puces à ADN ou biopuces sont définies comme des technologies versatiles et polyvalentes permettent la mesure simultanée des niveaux d'expression de plusieurs milliers de gènes, voire d'un génome entier, dans des dizaines de conditions différentes, physiologiques ou pathologiques.

L'utilité de ces informations est scientifiquement incontestable car la connaissance du niveau d'expression d'un gène dans ces différentes situations 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 [21].

Toutefois, dans une définition technique plus exacte, une puce ADN (appelée DNA microarray en anglais) est constituée de fragments d'ADN appelées puces immobilisés de manière ordonnée sur un support solide généralement fait de verre, de silicium ou bien de membrane en nylon,. Chaque emplacement de séquence est soigneusement repéré : la position (xi, yi) correspond au gène i. Un emplacement est souvent appelé spot ou sonde [11].

L'hybridation de la puce avec un échantillon biologique des fragments inconnus d'ADN que l'on cherche à identifier (appelées cibles) et qui a été marqué par une substance radioactive ou fluorescente permet de quantifier l'ensemble des cibles qu'il contient ; l'intensité du signal émis est proportionnel à la quantité de gènes cibles qu'il contient.

La fabrication d'une puce à ADN se décompose en trois étapes : la production des sondes (fragments courts d'une séquence d'ADN connus) et leur dépôt sur le support, la production et le marquage des cibles (fragments inconnus d'ADN que l'on cherche à identifier), enfin l'hybridation des sondes avec les cibles. Ces différentes étapes constituent [28] les étapes de base pour la fabrication de toutes les puces, indépendamment de la technologie utilisées. Le schéma suivant donne un aperçu général de cette technique de fabrication (Figure 1.4).

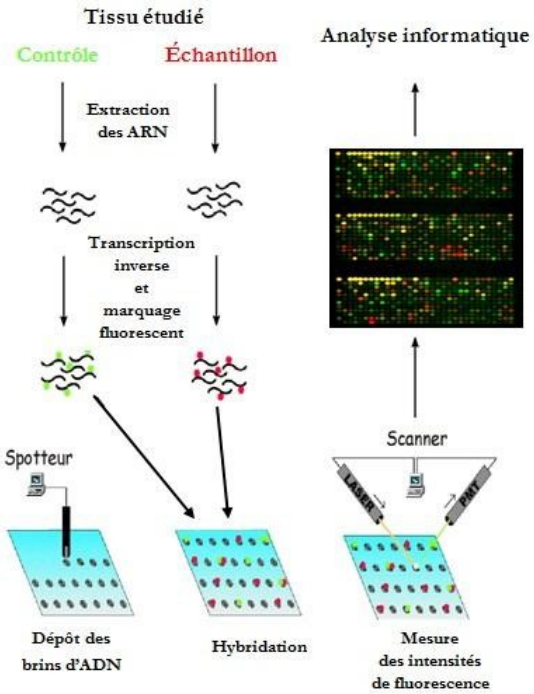


FIGURE 1.4: Schéma général du processus de fabrication d'une puce à ADN [10].

1.3.2 Origines de la technique

Les premières techniques de quantification de l'expression génique dérivent de celles développées pour l'ADN, telles que le Southern blot (1974), basées sur l'hybridation moléculaire entre des cibles d'intérêts et des sondes complémentaires marquées (radioactives ou fluorescentes) [21].

Ainsi, dans les années 70 fut développée la technique du buvardage de Northern, ou Northern blot (1977), consistant a séparer des ARN totaux par électrophorèse, a les transférer sur une membrane de nitrocellulose ou un filtre de nylon, et a employer des sondes (ARN ou ADN) radio marquées pour quantifier les transcrits d'intérêt [21].

Les techniques du RNA dot blot (1979) ou slot blot sont des variantes simplifiées du Northern blot, ou les ARN sont déposes directement sur les membranes sans électrophorèse préalable, ce qui les rend plus adaptées a des comparaisons simultanées entre de multiples conditions, mais ne permet pas d'identifier des différences de longueur entre les messagers produits par ces cellules. Ces méthodes comportent peu d'étapes susceptibles d'introduire des biais, mais les risques d'hybridations croisées, non détectables en absence de migration, peuvent conduire à une détection non spécifique [21].

Alors, on peut dire que l'idée de concevoir des puces à ADN découle de la volonté de développer un outil permettant de quantifier de façon fiable le niveau d'expression d'un grand nombre de gènes simultanément. En plus d'une simplification et transposition à grande échelle du principe du Northern blot, cette technique s'en distingue donc par le fait que les sondes sont fixées et non marquées, alors que les cibles sont marquées et en solution.

1.3.3 Phases d'analyse par puce à ADN

Les différentes phases d'une analyse par puces ADN sont indiquées dans la figure ci-dessous :

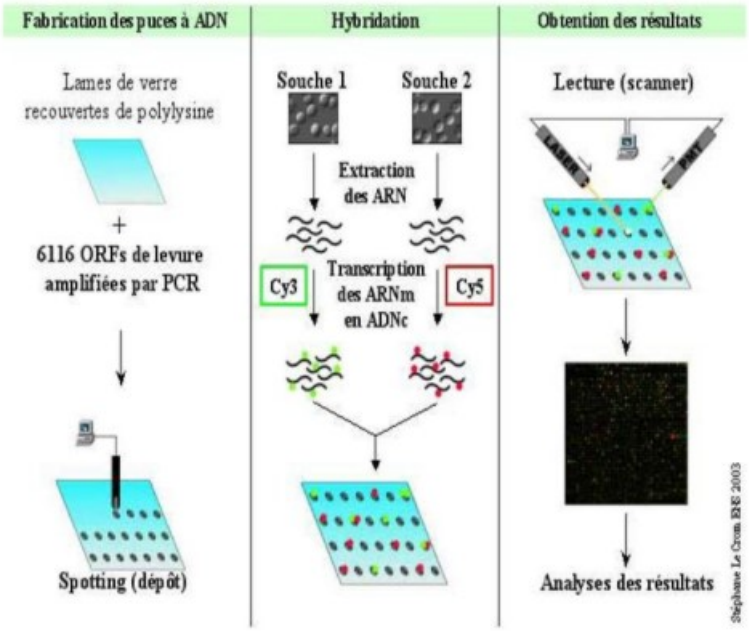


FIGURE 1.5: Les trois phases d'une analyse par puce à ADN.

1.3.3.1 Fabrication de la puce

Pour comparer les niveaux d'expression dans deux échantillons biologiques ou deux conditions (référence et pathologique), la première étape consiste en la fabrication de la puce par la production des sondes : Il s'agit d'extraire des séquences d'ADN double brin (sondes) puis de le amplifier par la technique PCR (Polymerase Chain Reaction) et ensuite déposées sur la lame de verre. La zone de dépôt du gène est appelée spot. Les sondes sont dénaturées en un seul brin pour pouvoir être hybridées avec les cibles marquées dans la phase d'hybridation [10].

1.3.3.2 Hybridation

Après les dépositions des sondes sur la lame, la deuxième étape consiste en :

a) La préparation des cibles :

Il s'agit d'extraire les ARNm d'un échantillon biologique à analyser puis leur intégrer **un marqueur fluorescent** (Cy3 et Cy5), c'est-à-dire qu'un échantillon est marqué avec un fluorochrome vert, tandis que le deuxième est marqué avec un fluorochrome rouge ce qui permettra d'évaluer et de quantifier l'appariement sonde/cible.

Le marquage est effectué suite à la transcription inverse de l'ARNm permettant d'obtenir un brin d'ADNc fluorescent.

La qualité de l'extraction est bien sûr primordiale pour la réussite de l'hybridation qui va suivre. Une mauvaise purification peut conduire à une augmentation des bruits de fond sur la lame [11].

b) Hybridation sondes/ cibles

L'hybridation est ensuite réalisée sur une seule puce (simple marquage) ou sur deux puces (double marquage : un échantillon sur chaque puce) [11]. Elle aboutit à la fixation sur chaque gène cible d'une quantité de l'espèce correspondante d'ARNm de la sonde, proportionnelle à son abondance dans l'ARN de départ [2].

La durée oscille entre 10 à 17 heures en milieu liquide à 60 degrés, en fait à cette température un fragment d'ADN simple brin ou d'ARN messager reconnaît son brin complémentaire (ADNc) parmi des milliers d'autres pour former un ADN de double brin (duplex ou double hélice) [11].

Cette étape est suivie d'un lavage du support de culture pour éliminer les cibles non fixées ou fixées non spécifiquement.

1.3.3.3 Obtention des résultats

Suite à l'hybridation, une étape de lecture de la puce permet de repérer les sondes ayant réagi avec l'échantillon testé. Cette lecture est une étape clé [11]. En effet, sa qualité conditionne de façon importante la précision des données et donc, la pertinence des interprétations.

L'acquisition des images est réalisée par lecture des puces au moyen d'un scanner de haute précision. Dans le cas du marquage avec les deux fluorochromes, on obtient une image dont la couleur des spots va du rouge au vert :

— Un spot de couleur verte indique un gène dont le niveau d'expression est plus élevé dans l'échantillon marqué avec le Cy3 que celui marqué avec le Cy5, et inversement pour la couleur rouge [10].

— La couleur jaune indique que le gène est exprimé de manière identique dans les deux échantillons tandis que la couleur noire indique l'absence de signal [10].

Il ne reste plus qu'à analyser les images scanner en trois étapes. La localisation des spots, la segmentation des spots et l'extraction des intensités des signaux Cy3 et Cy5 (Figure 1.6).

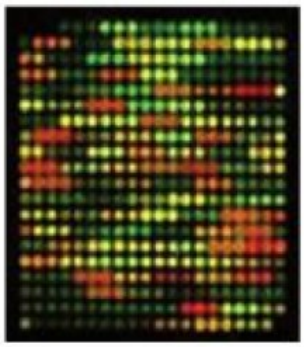


FIGURE 1.6: Résultat d'une image d'un scanner d'une puce à ADN.

1.3.3.4 Analyse des données

Après lavage et acquisition, l'image d'hybridation va être traitées par des logiciels d'analyse qui permettent de mesurer la fluorescence de chaque spot sur la lame (estimant les niveaux d'expression pour chacun des gènes présents sur la puce), mais aussi de relier chaque sonde à l'annotation correspondante (nom de gêne, numéro de l'ADNc utilisé, séquence de l'oligonucléotide, etc.) puis calculer l'intensité de chaque spot [11].

a) Transformation des données : Les données d'intensité sont rarement manipulées sans transformation. L'une des transformations la plus couramment employée est celle qui utilise le logarithme à base deux pour les raisons suivants [11] :

- D'une part, la variation du logarithme des intensités est moins dépendante de la grandeur des intensités,
- et d'autre part, cette transformation permet de se rapprocher d'une distribution symétrique et d'obtenir une meilleure dispersion avec moins de valeurs extrêmes.

L'autre transformation qu'elle peut être appliquée sur ces intensité c'est la normalisation. Elle consiste à éliminer les différences entre les différentes puces liées aux variations de quantité de départ, aux biais de marquage ou d'hybridation et aux variations du bruit de fond [11] afin de dresser pour chaque échantillon (une tumeur, par exemple) un véritable portrait moléculaire qui pourra alors être comparé à celui d'autres échantillons [2].

b) Présentation des données : après les transformations décrites ci-dessus, les données recueillies pour l'étude d'un problème donné sont regroupées sous forme de matrice avec une ligne par gène, et une colonne par échantillon (Table 1.1). Chaque valeur de mij est la mesure du niveau d'expression du i^{ieme} gène dans le j^{ieme} échantillon, où i =1, . . . ,M et j=1, . . . ,N.

Gène id	Echantillon 1	Echantillon 2	....	Echantillon N
Gèene 1	m_{11}	m_{12}	....	m_{1N}
Gène 2	m_{21}	m_{22}	....	m_{2N}
Gène 3	m_{31}	m_{32}	....	m_{3N}
....	....	....	....	.....
Gène M	m_{M1}	m_{M2}	....	m_{MN}

TABLE 1.1: Matrice d'expressions des gènes .

L'analyse comparative de ces données peut être utiliser pour l'identification d'une signature moléculaire (combinaison de plusieurs gènes) afin de définir de nouvelles classes de tumeurs sur la seule base de leur profil d'expression (approches non supervisées) et/ou de caractériser des classes de tumeurs associées à un phénotype d'intérêt (la survie par exemple, approches supervisées) [2] .Elle fait appel à des outils bio-informatiques sophistiqués traitant l'énorme quantité des données produites.

1.3.4 Types des puces à ADN

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

a) Dépôt direct d'ADNc :

Ces puces reposent sur le principe de dépôt direct d'ADNc sur lamelle de verre activée ou de nylon à l'aide d'un micropipetteur robotisé. Grâce à cette technique, chacun des gènes (de fonction connue ou inconnue) est représenté par un seul point sur la lamelle [18].

En général, deux échantillons d'ARN (sous forme d'ADNc obtenus par transcription inverse) sont Co hybridés sur la puce à ADNc. Les deux échantillons marqués par un fluorochrome différent (Cy-3 vert ou Cy-5 rouge) s'hybrident simultanément avec les molécules complémentaires sur la puce. La puce est lue par un scanner afin de mesurer l'intensité du signal lumineux mesurée aux deux longueurs d'ondes correspondant aux différents fluorochromes [17].

Le rapport de fluorescence rouge/vert est ainsi déterminé. Il permet de comparer les taux d'expression relatifs de chacun des gènes pour les deux échantillons d 'ADNc. Un excès du gène X dans l'échantillon marqué en rouge produira un signal rouge au point représentant le gène, un excès du gène Y dans l'échantillon marqué en vert produira un signal vert ; enfin, une expression équivalente du gène Z dans les deux échantillons produira un signal jaune [17].

La société Agilent est l'une des plus grandes industries qui commercialise ce type [10]
.

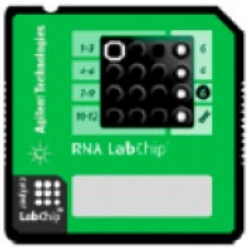


FIGURE 1.7: Puce à ADN d'Agilent Technologies.

b) Les puces à oligonucléotides :

Elles reposent sur le principe de la synthèse insitu d'oligonucléotides par photolithographie. Dans ce procédé, une lumière dirigée sur des sites spécifiques de la puce active la réaction d'oligo-synthèse. On peut synthétiser jusqu'à 300 000 oligonucléotides représentant 30 000 gènes sur une puce d'une surface d'environ 1 cm2 [28].

Une puce à ADN destinée à des études d'expression contient pour chaque gène un ensemble d'oligonucléotides mimant la séquence du gène, souvent choisis dans sa région, réduisant ainsi les risques d'hybridations croisées avec des séquences homologues de ce gène [18].

Contrairement au Puces d'Agilent Technologies, celle produite par ce procédé permet l'hybridation d'un seul échantillon marqué à la fois. L'intensité de l'hybridation est également mesurée par scanner [18].

Les puces les plus couramment utilisées sont les puces de la société Affymetrix [28].

FIGURE 1.8: Puce à oligonucléotides d'Affymetrix.

1.3.5 Domaines d'application

Les puces à ADN permettent des tests plus rapides, plus sensibles et plus spécifiques. Elles sont utiles dans divers domaines très important tels que la recherche biologique (et notamment la génomique fonctionnelle), pharmaceutique, le génotypage, les diagnostics médicaux, les expertises médico-légales, et bien d'autres domaines.

1.3.5.1 Analyse d'expressions des gènes

Les premières puces ont servi à évaluer l'expression simultanée de milliers de gènes dans des systèmes biologiques bien connus, tels que celui du métabolisme respiratoire et de fermentation chez la levure, le cycle cellulaire de la levure et la stimulation de fibroblastes par le sérum. Ces premiers travaux ont permis de valider la technologie [17].

1.3.5.2 Diagnostic médical

L'utilisation de la puce ADN dans le domaine de l'oncologie est réputée. Indépendamment de ceci, elle est employée pour étudier cardiovasculaire, inflammatoire variés, et des maladies infectieuses, ainsi que des troubles psychiatriques.

L'application de la puce ADN dans le domaine médical peut être classée par catégorie dans quatre types [22] :

- **La découverte de l'objectif :**la puce ADN peut être employé pour comparer les tissus/cellules malades aux tissus sains/aux cellules pour trouver les caractéristiques d'une maladie particulière. Ceci aide en trouvant les gènes responsables de cette maladie.
- **La découverte des médicaments et des plombs :**des puces ADN peuvent être employées pour examiner les composés potentiels et pour recenser la toxicité du composé de plomb qui aidera en décidant le médicament correcte pour le patient.
- **Diagnostics et pronostics :**la puce ADN est très utilisée pour connaître la condition de la maladie, type de tumeur et autre factorise important pour le patient.
- **Pharmacogenomics et theranostics :**la technique de puce ADN peut être employée pour décider la demande de règlement et le traitement d'un patient sur la base de son renivellement génétique. Elle peut également aider dans des effets secondaires de réglage des médicaments.

1.3.5.3 L'identification et l'expertise médico-légale

Le but est l'identification humaine dans le cadre d'enquêtes policières ou judiciaires. Les analyses sur le terrain étant très souvent complexes ainsi que la confidentialité et le respect de la procédure judiciaire assez lourdes, il sera souhaitable d'avoir sur les lieux d'enquête des systèmes portables d'analyse de l'ADN [28].

1.3.5.4 L'environnement

Les secteurs de la défense et de l'environnement font partie des diverses applications des puces à ADN, notamment pour la détection rapide et à bas coût de substances organiques, principalement des agents pathogènes dilués dans l'environnement [10].

1.3.5.5 L'identification d'espèces animales dans l'alimentation humaine et animale

Aujourd'hui, la qualité et la sécurité alimentaire sont garanties grâce à l'identification des espèces animales dans l'alimentation humaine ou animale. Les industriels de l'industrie agroalimentaire peuvent apporter la preuve aux consommateurs ou aux instances réglementaires de ce qu'ils ont dans leur assiette ! Les techniques actuelles permettent de répondre à cette question : « Cette espèce est-elle dans mon produit ? » [28].

1.3.6 Banques des données génomiques

aujourd'hui, Les banques de données sont devenues indispensables pour sauvegarder et structurer les informations issues des expériences de biologie et plus particulièrement des données générées par les différentes technologies de puces à ADN [17].

La MGED (Microarray Gene Expression Data Society) 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 [10].

L'intérêt principal de ces banques de données généralistes est la mise à disposition des jeux de données aux communautés de chercheurs en bio-informatique, mathématiques et statistiques pour le développement de nouvelles méthodologies d'analyse [17].

Parmi les banques de données publiques, les banques de données d'expression de gènes sont particulièrement importantes et intéressantes en termes de partage des connaissances et des données d'expression de gènes (notamment issues des expériences de puces à ADN) au niveau de la communauté scientifique internationale.

La MGED (Microarray Gene Expression Data Society) 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) [10].

Le standard MIAME [10] requiert que les informations suivantes soient fournies pour les publications basées sur les expériences de puce à ADN :

- Les données brutes résultant de l'analyse de l'image de chaque puce (fichiers CEL).
- Les données finales après le prétraitement qui est la matrice d'expression des gènes .
- Les informations essentielles à propos de l'annotation de l'échantillon et des facteurs expérimentaux.
- Le plan expérimental incluant les relations entre échantillons, puces et fichiers de données.
- Une description de la conception de la puce (information sur les sondes et leurs numéros dans la base de données d'où elles proviennent).

— Les protocoles de traitement expérimentaux des données.

Les deux principales banques de données généralistes pour le dépôt des données d'expression de gènes sont :

— **GEO (Gene Expression Omnibus) :** 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). 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) [10].
— **ArrayExpress :** une base de données publique d'expérience de puce à ADN et de profils d'expression des gènes établie en 2002 à l'European Bioinformatics Institute (EBI).

Ces entrepôts sont d'une importance grandissante puisque, aujourd'hui, la majorité des journaux scientifiques requièrent, pour toutes publications dans le domaine des puces à ADN, le dépôt des données d'expression dans au moins une des banques de données publiques conforme au standard international MIAME [17].

1.4 Conclusion

Dans le domaine de la cancérologie, la technologie des puce à ADN connaît aujourd'hui un essor considérable, puisque les techniques d'analyse de données permettent d'en déduire la fonction d'un gène, de créer des modèles de prédiction et des outils de diagnostic.

Cependant, la construction d'un modèle de prédiction peut s'avérer très lourde, du fait que les données issues des puces à ADN sont obtenues à partir d'un protocole complexe où plusieurs étapes peuvent introduire du bruit dans les données. De plus ces données sont décrites par des milliers d'attributs (gènes), pour un très petit nombre de cas d'études.

CHAPITRE

2

SELECTION D'ATTRIBUTS

2.1 Introduction

Chercher à réduire la dimensionnalité d'un ensemble de données devient de plus en plus indispensable en raison de la multiplication des données. Plusieurs paramètres peuvent influer sur les performances d'un système de classification dont les caractéristiques extraites, à partir des entités afin de représenter ces dernières, peuvent être considérées parmi les paramètres les plus importants.

La réduction de la dimensionnalité, via la sélection d'attributs, est l'une des étapes les plus importantes dans le traitement de données [16] qui sert à faciliter la visualisation et la compréhension des données, réduire l'espace de stockage nécessaire, réduire le temps d'apprentissage et d'utilisation, identifier les facteurs pertinents et améliorer la précision du module de classification.

Dans ce chapitre, nous allons présenter le problème de sélection d'attribut, détailler les étapes de son processus et montrer la nécessité de son utilisation dans le domaine des puces à ADN.

2.2 La sélection d'attributs

Au cours de la dernière décennie, la motivation pour l'application des techniques de sélection de caractéristiques (Feature Selection) est passée d'un exemple illustratif à un véritable prérequis. L'augmentation du nombre de ces variables (caractéristiques) qui modélisent un problème donné introduit des difficultés à plusieurs niveaux comme la

complexité, le temps de calcul ainsi que la détérioration du système de résolution en présence de données bruitées [3].

2.2.1 Définition de sélection d'attributs

La sélection d'attributs, de caractéristiques ou feature selection en anglais, est un problème difficile qui a été étudié depuis les années 70.
C'est une technique de recherche qui consiste à choisir parmi un ensemble d'attributs de grande taille un sous-ensemble d'attributs pertinents et intéressants pour des objectifs et des critères du système fixés au paravent [12] : Les données d'entrée du processus sont constituées par l'ensemble initial de variables qui forment l'espace de représentation et l'ensemble des données d'apprentissage du problème étudié.

Dans la littérature, le problème de sélection de caractéristiques a été généralement défini comme suit :

Soit $F = \{F_1; F_2...F_N\}$ un ensemble de caractéristiques de taille N où N représente le nombre total de caractéristiques étudiées. Soit E une fonction qui permet d'évaluer un sous-ensemble de caractéristiques. Nous supposons que la plus grande valeur de E soit obtenue pour le meilleur sous-ensemble de caractéristiques. L'objectif de la sélection est de trouver un sous-ensemble $\acute{F}(\acute{F} \subseteq F)$ de taille $\acute{N}(\acute{N} \subseteq N)$ tel que [3] :

$$E(\acute{F}) = \max E(Z), Z \subseteq F$$

Où $|Z| = \acute{N}$ et $\acute{N}$ est, soit un nombre prédéfini par l'utilisateur ou soit contrôlé par une des méthodes de génération de sous-ensembles.

Dans ce contexte, les principales motivations de la sélection d'attributs sont les suivantes :

1. Utiliser un sous-ensemble plus petit permet d'améliorer la classification si l'on élimine les attributs qui sont source de bruit. Cela permet aussi une meilleure compréhension des phénomènes étudiés.

2. Des petits sous-ensembles d'attributs permettent une meilleure généralisation des données en évitant le sur-apprentissage.

3. Une fois que les meilleurs attributs sont identifiés, les temps d'apprentissage et d'exécution sont réduits et en conséquence l'apprentissage est moins coûteux/

2.2.2 Processus de sélection d'attributs

Le processus de sélection de variables se décompose de la manière suivante [19] :

— A partir de l'ensemble initial des variables, le processus de sélection détermine un sous-ensemble de variables qu'il considère comme les plus pertinentes.

— Le sous-ensemble est ensuite soumis à une procédure d'évaluation. Cette dernière permet d'évaluer les performances et la pertinence du sous-ensemble .

— En fonction du résultat de la procédure d'évaluation, un critère d'arrêt du processus détermine si le sous-ensemble de variables peut être soumis à la phase d'apprentissage. Si tel est le cas, le processus de sélection s'arrête, sinon, un autre sous-ensemble de variables est généré.

Ce processus peut être schématisé comme suit :

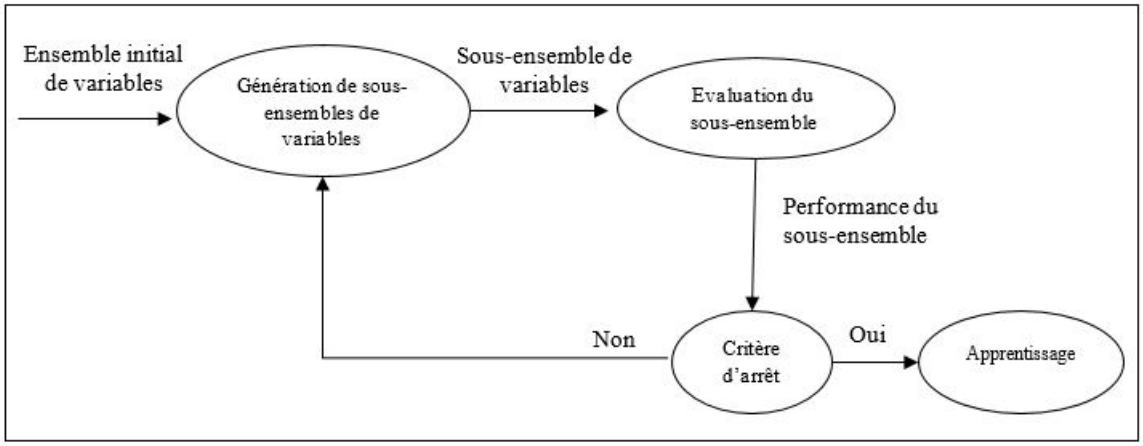


FIGURE 2.1: Processus de sélection d'attributs [19].

Alors, on peut dire qu'il y a quatre étapes de base pour une méthode typique de sélection de Caractéristique :

— Une procédure de génération,
— Une fonction d'évaluation,
— Un critère d'arrêt,
— Une procédure de validation.

2.2.2.1 La procédure de génération

C'est une procédure de recherche permettant, à chaque itération, de générer un sous-ensemble d'attributs qui va être évalué lors de la seconde étape de la procédure de sélection. Cette procédure de génération peut soit commencer avec un ensemble vide d'attributs, soit avec l'ensemble de tous les attributs, soit avec un sous-ensemble d'attributs choisis aléatoirement [15].

Dans les deux premiers cas, les attributs sont itérativement ajoutés (Forward selection) ou retirés (Backward selection). Dans le troisième cas, soit on ajoute, ou on retire des

attributs comme dans les deux premiers cas, soit un nouveau sous-ensemble d'attributs est créé de manière aléatoire à chaque itération (Random generation).

Trois différentes stratégies de génération ont été proposées dans la littérature : la génération complète, séquentielle et aléatoire.

a. La génération complète

Les approches regroupées, dans cette catégorie, effectuent une recherche complète du sous-ensemble optimal [16] au sens d'exécution de la fonction d'évaluation utilisée.

Cette stratégie de recherche garantit de trouver le sous-ensemble optimal. Le problème majeur de cette approche est que le nombre de combinaisons croît exponentiellement en fonction du nombre de caractéristiques. Pour un ensemble de N caractéristiques, et quand N devient grand, les 2N combinaison possibles rendent la recherche exhaustive impossible [3].

Différentes fonctions heuristiques peuvent être utilisées afin de réduire l'espace de recherche sans pour autant compromettre les chances de trouver le sous-ensemble optimal. Il s'agit d'utiliser un processus de backtracking permettant de revenir en arrière si la sélection s'engage dans une mauvaise direction de génération [15].

b. La génération séquentielle

Le principe des procédures de génération séquentielle est d'ajouter ou de supprimer un ou plusieurs attributs au fur et à mesure des itérations. On distingue alors deux approches de génération séquentielle [15] :

— L'approche de type Forward ou Ascendante : cette approche part d'un ensemble vide d'attributs auquel, à chaque itération, un ou plusieurs attributs sont ajoutés.

— L'approche de type Backward ou Descendante : c'est l'approche inverse, elle part de l'ensemble total des attributs. Chaque itération permet de supprimer un ou plusieurs attributs.

Les algorithmes avec une génération séquentielle sont simples à implémenter et rapides dans la production des résultats, l'espace de recherche utilisé est de l'ordre O(N2). Dans ces algorithmes, on abandonne l'exhaustivité et on risque ainsi de perdre les sous-ensembles optimaux [16].

c. La génération aléatoire

Par rapport aux deux premières catégories, l'utilisation de la génération aléatoire dans la procédure de sélection de caractéristiques est la plus récente. Elle parcourt au hasard l'ensemble des 2^N sous-ensembles candidats, le sous-ensemble courant n'est alors pas issu

d'une augmentation ou diminution d'attributs du sous-ensemble précédent. Cela permet de ne pas arrêter la recherche lorsque la fonction d'évaluation d'un sous-ensemble atteint un optimum local. Un nombre maximal d'itérations est imposé afin que les temps de calcul restent raisonnables [15].

2.2.2.2 La fonction d'évaluation

Typiquement, une fonction d'évaluation tente de mesurer la capacité de discrimination d'une caractéristique ou d'un sous-ensemble pour distinguer les différentes classes, dont l'optimalité d'un sous-ensemble est relative à la fonction d'évaluation utilisée.

Les méthodes utilisées pour évaluer un sous-ensemble de caractéristiques dans les algorithmes de sélection peuvent être classées en trois catégories principales : "filter", "wrapper" et "embedded".

a. Approche par filtre (Filter Approch)

Le principe de cette approche consiste à évaluer chaque attribut pour lui assigner **un score de pertinence** selon des mesures qui reposent sur les propriétés des données d'apprentissage. Ce score permet un classement des attributs afin de sélectionner les attributs les mieux classés c'est-à-dire les plus pertinents [12].

L'évaluation se fait généralement indépendamment d'un classificateur. Les méthodes qui se basent sur ce modèle pour l'évaluation des caractéristiques, utilisent souvent une approche heuristique comme procédure de génération [3]. La procédure du modèle "filter" est illustrée par la figure suivante :

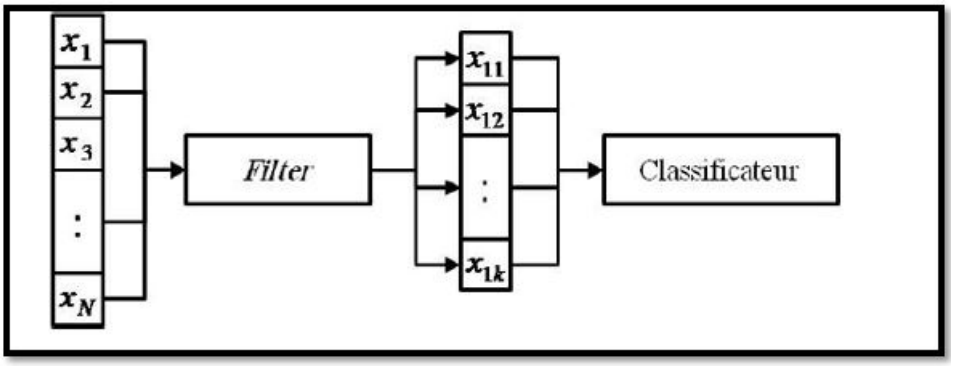


FIGURE 2.2: Processus du modèles Filtre [3].

soit $X = \{x_k | x_k = (x_{k1}, x_{k2}, ...x_{kn}), k = 1, 2, ...m\}$ un ensemble de m exemples d'apprentissage dans un espace de représentation comportant n caractéristiques. Soit $Y = \{y_k, k = 1, 2, ...m\}$ ou y_k représente l'étiquette de la classe de l'exemple x_k. si $x^i = (x_{1i}, x_{2i}, ...x_{mi})$ représente la i^{eme} caractéristique$(i = 1, 2, ..., n)$ alors le but d'une

méthode d'évaluation "Filter" est de calculer un score pour évaluer le degré de pertinence de chacune des caractéristiques (x^i).

Ces méthodes sont rapides, plus générales et moins coûteuses en temps de calcul, ce qui leur permet d'opérer plus facilement avec des bases de données de très grandes dimensions. Cependant, comme elles sont indépendantes de l'étape de classification, elles ne permettent pas de garantir que le meilleur taux de classification soit obtenu dans l'espace retenu [15].

b. Approche enveloppe (Wrapper Approch)

Le principal inconvénient des approches "Filter" est le fait qu'elles ignorent l'influence des caractéristiques sélectionnées sur la performance du classificateur à utiliser par la suite.

Cette approche utilise l'algorithme d'apprentissage comme une fonction d'évaluation, elle définit donc la pertinence des attributs par l'intermédiaire d'une prédiction de la performance du système final [16]. La procédure du modèle "wrapper" est illustrée par la figure (2.3) :

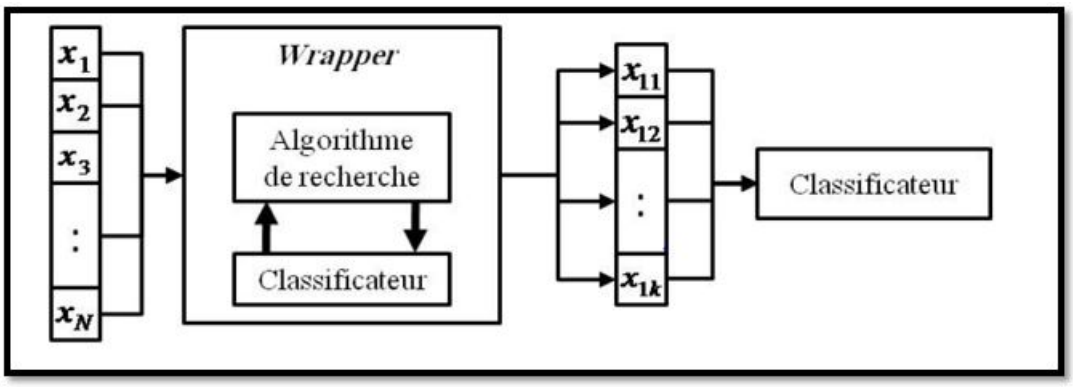


FIGURE 2.3: Processus du modèles Enveloppe [3].

Les sous-ensembles de caractéristiques sélectionnés par cette méthode sont bien adaptés à l'algorithme de classification utilisée, mais ils ne sont pas forcément valides si on change le classificateur. La complexité de l'algorithme d'apprentissage rend les méthodes "wrapper" très couteuses en temps de calcul [3]. Le problème de la complexité de cette technique rend impossible l'utilisation d'une stratégie de recherche exhaustive. Par conséquent, des méthodes de recherche heuristiques ou aléatoires peuvent être utilisées.

En général, pour diminuer le temps de calcul et pour éviter les problèmes de surapprentissage, le mécanisme de validation croisée est fréquemment utilisé.

L'approche filtre est plus rapide que l'approche Wrapper en terme de génération de résultats. Cependant, cette dernière à l'avantage de fournir généralement des résultats

plus pertinents pour la classification [16]. La Figure (2.4) présente deux modèles généraux des approches Filtre et Wrapper pour la sélection de caractéristiques :

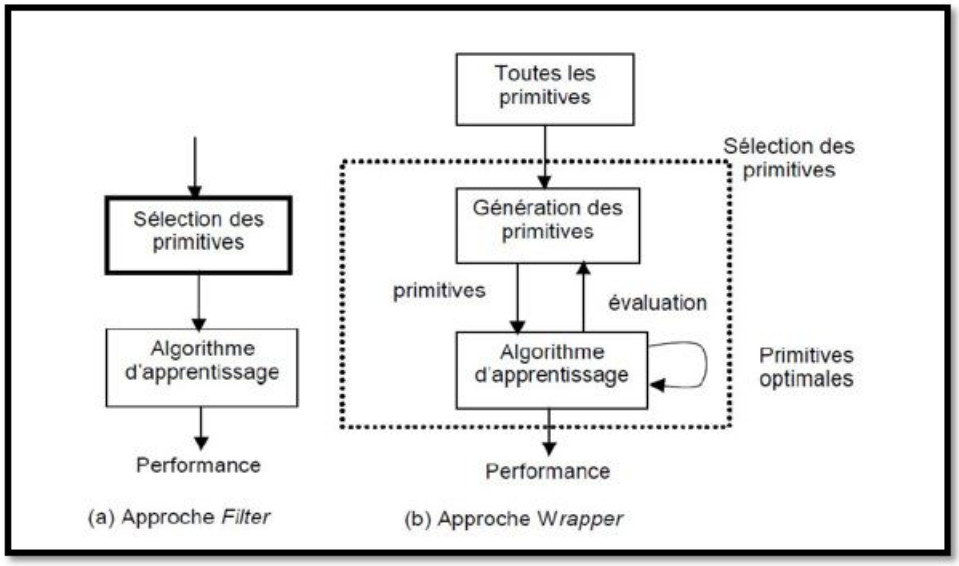


FIGURE 2.4: L'approche Filtre et Enveloppe [16].

c. Approche hybride (Embedded Approch)

es algorithmes hybrides ont été plus récemment proposés par rapport aux précédents (filtre et wrapper). Cette approche incorpore la sélection de variables lors du processus d'apprentissage.

La différence avec les méthodes enveloppe est que le classificateur sert non seulement à évaluer un sous ensemble candidats mais aussi à guider le mécanisme de sélection [12] ; la base d'apprentissage dans l'approche enveloppe est divisée en deux parties : une base d'apprentissage et une base de validation pour valider le sous-ensemble de caractéristiques sélectionnées, en revanche, les méthodes intégrées peuvent se servir de tous les exemples d'apprentissage pour établir le système de classification [3]. Un tel mécanisme peut être trouvé, par exemple, dans les algorithmes de type SVM, ou dans les arbres de décisions [3].

L'avantage de ces méthodes est que le processus de recherche est guidé par des informations intéressantes fournies par le classifieur, ce qui rend ces méthodes plus efficaces que les méthodes enveloppes [12]. Un autre avantage de ces méthodes est leur rapidité par rapport aux approches "Wrapper" parce qu'elles évitent que le classificateur recommence de zéro pour chaque sous-ensemble de caractéristiques [3].

En effet, la sélection d'un sous-ensemble d'attributs optimal est toujours relative au critère utilisé car l'emploi différents critères ne permettent pas de sélectionner le même sous-ensemble d'attributs optimal. Différentes mesures d'évaluation ont été proposées pour évaluer un attribut ou un sous-ensemble d'attributs dans un contexte de sélection. Elles

peuvent être classées en cinq mesures distinctes :

a. Mesure d'information

Ces mesures permettent d'estimer le gain d'information d'une caractéristique. Le gain d'information de la caractéristique X est défini comme la différence entre l'incertitude a priori et celle a posteriori, c'est-à-dire la caractéristique X est préférée à Y, si le gain d'information de X est plus grand de celui de Y [16].

b. Mesure de distance

Les mesures de distance sont aussi nommées mesures de séparabilité, divergence ou de discrimination. Un attribut ou un sous-ensemble d'attributs est sélectionné s'il permet une meilleure séparabilité et cohérence des classes [15]. Une variable X est préférée à Y, si X introduit une plus grande différence, entre les probabilités conditionnelles des deux classes, que Y [16].

c. Mesure d'erreur de classification

L'attribut ou les sous-ensembles d'attributs considérés sont évalués en fonction de la qualité de la classification obtenue en utilisant ces attributs. Le sous-ensemble d'attributs le plus discriminant est celui pour lequel le taux d'erreur de classification est le plus faible [15].

d. Mesures de dépendance

Également appelée mesure de corrélation ou mesure de similarité. Elle permet de mesurer la capacité d'une caractéristique à prédire la valeur d'une autre [16]. La caractéristique X est préférée à Y, si la corrélation d'une variable X avec une classe C est plus importante que la corrélation de la variable Y avec C.

e. Mesures de consistance

Les mesures de consistance cherchent à évaluer si l'attribut (ou le sous-ensemble d'attributs) étudié contient les informations nécessaires à la discrimination des classes [15].

2.2.2.3 Le critère d'arrêt

Les itérations du processus de sélection de caractéristiques continuent à s'exécuter jusqu'à ce qu'un critère d'arrêt soit rempli. La procédure de génération et la fonction d'évaluation peuvent influencer sur le choix d'un critère d'arrêt [16] :

a. Les critères d'arrêt basés sur la procédure de génération

— Le nombre de caractéristiques sélectionnées est égal à un nombre prédéfini.

— Un nombre prédéfini d'itérations est atteint.

b. Le critère d'arrêt basé sur la fonction d'évaluation

— L'ajout (ou suppression) d'une caractéristique ne produit pas un meilleur sous ensemble.

— Un sous-ensemble optimal de caractéristiques est obtenu à partir de certaines fonctions d'évaluation.

Par exemple, lorsque les méthodes par filtre sont utilisées, le critère d'arrêt couramment utilisé est basé sur l'ordre des caractéristiques, rangées selon certains scores de pertinence et une fois les caractéristiques ordonnées, celles qui ont les scores les plus élèves seront choisies et utilisées par un classificateur [3].

D'autre part, si on utilise l'approche "enveloppe" ou l'approche "hybride", les taux de bonne classification obtenus par les différents sous-espaces sont comparés pour mesurer le gain d'information. On peut ainsi décider d'arrêter la procédure de sélection dès que ce taux diminue ou alors dès qu'il atteint un certain seuil [16].

2.2.2.4 La procédure de validation

La procédure de validation n'est pas une étape du processus de sélection de caractéristiques lui-même, mais une méthode de sélection d'attributs (en pratique) doit être validée. Elle sert à tester la validité des sous-ensembles de caractéristiques sélectionnées avec la réalisation de différents tests, et la comparaison des résultats obtenus avec les résultats précédents, ou avec des résultats d'autres méthodes de sélection de caractéristiques [16].

L'ensemble des données est généralement divisé en deux sous-ensembles distincts : le sous ensemble d'apprentissage constitué des prototypes des classes (données avec leurs labels) et le sous-ensemble de test dont on ne connait pas les labels de classes de ses données [15]. Selon la répartition des données entre ces deux sous-ensembles, il existe différentes approches de validation, on peut citer :

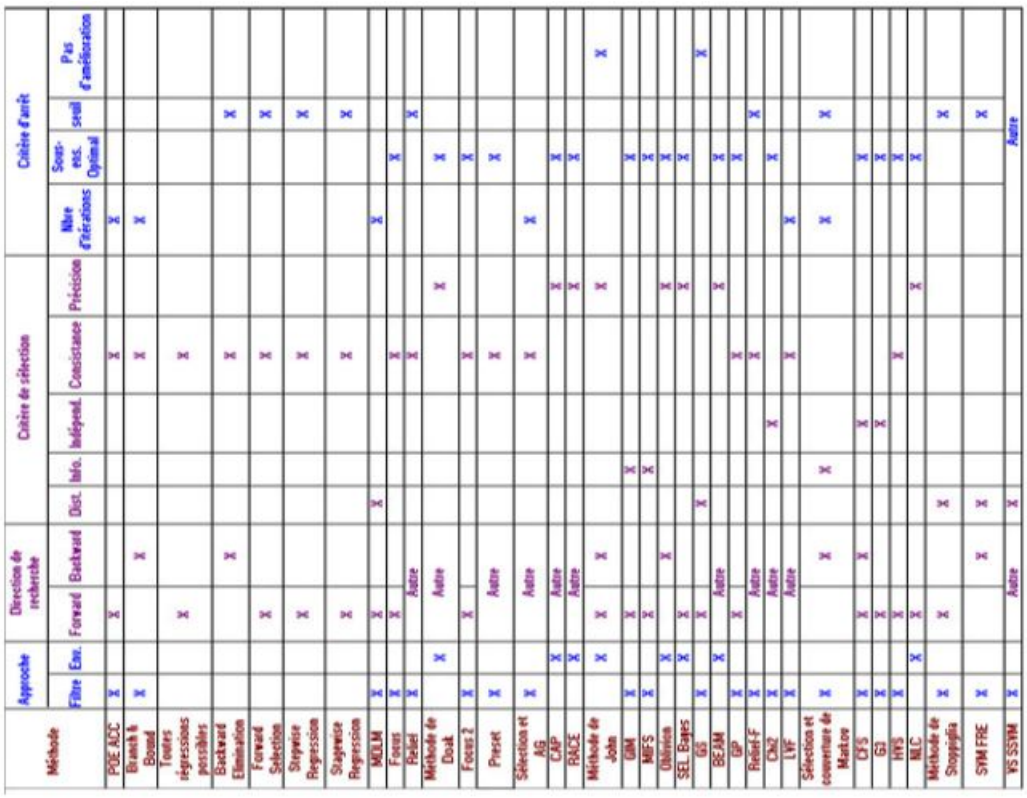
a- La méthode Hold Out : les données sont divisées en deux sous-ensembles : le sous-ensemble d'apprentissage et le sous-ensemble de test dans des proportions $1/2$ pour chacun de ses deux sous-ensembles ou $2/3$ pour l'ensemble d'apprentissage et $1/3$ pour l'ensemble de test.

b- La méthode de resubstitution : l'ensemble d'apprentissage est utilisé comme ensemble de test.

c- La méthode V-validation croisée : l'ensemble des données est partitionné en V parties de tailles à peu près égales. Nous réalisons ainsi V fois la procédure de validation et à chaque fois une des parties constitue l'ensemble test et les V-1 parties restantes sont réunies pour former l'ensemble d'apprentissage.

Une fois que les données sont divisées en un ensemble d'apprentissage et un ensemble de test, l'erreur de classification est mesurée sur l'ensemble de test en utilisant les prototypes de classes de l'ensemble d'apprentissage.

Afin de récapituler l'ensemble des méthodes de sélection, la table 2.1 présente l'ensemble des méthodes ainsi que leurs caractéristiques principales, à savoir le type d'approche, la méthode de génération, l'approche d'évaluation, le critère d'évaluation et le critère d'arrêt :



Méthode	Approche		Direction de recherche		Critère de sélection						Critère d'arrêt				
	Filtre	Env.	Forward	Backward	Dist.	Info.	Indépend.	Consistance	Précision	Nbre d'itérations	Sous-ens. Optimal	seuil	Pas d'amélioration		
POE ACC	X		X							X					
Branch & Bound	X			X				X			X				
Toutes régressions possibles			X					X							
Backward Elimination				X				X				X			
Forward Selection			X					X				X			
Stepwise Regression			X					X				X			
Stagewise Regression			X					X				X			
MDLM	X		X		X					X					
Forest	X		X					X			X				
Relief	X			Autre				X				X			
Méthode de Douk		X		Autre					X		X				
Focus 2	X		X					X			X				
Pinset	X			Autre				X			X				
Sélection et AG	X			Autre				X		X					
CAP		X	Autre						X		X				
RACE		X	Autre						X		X				
Méthode de Jobs		X	X	X					X				X		
GINI	X		X			X					X				
MIFS	X		X			X					X				
Uninion		X		X					X		X				
SEL Bayes		X	X						X		X				
GS	X		X		X								X		
BEAM		X	Autre						X		X				
GP	X		X					X							
Relief-F	X		Autre				X	X				X			
Chi2	X	X	Autre								X				
LVF	X	X	Autre				X	X		X					
Sélection et couverture de Markov	X			X		X				X		X			
CFS		X	X	X			X				X				
GI	X	X	X				X				X				
HYS	X	X	X					X			X				
NLC		X	X						X		X				
Méthode de Stoppiglia	X		X		X							X			
SVM FRE	X			X	X							X			
WSSVM	X		Autre		X									Autre	

TABLE 2.1: Récapitulatif des méthodes de sélection d'attributs [19].

2.3 Sélection d'attributs pour les puces à ADN

Comme on a déjà montré dans le chapitre précédent, les puces à ADN a permis de mesurer les niveaux d'expression d'un grand nombre de gènes simultanément dans une seule expérience ce que permet aux chercheurs d'avoir un aperçu complet pour découvrir précisément quels gènes sont exprimés dans un tissu spécifique dans diverses conditions.

2.3.1 Motivations

Développer des modèles de prédiction utilisant des profils d'expression génique est assez difficile à cause de la grande dimensionnalité de données issues de ces puces due au nombre élevé de descripteurs (gènes) pour un nombre limité d'échantillons [12]. Parmi tous ces gènes, beaucoup sont non pertinents, insignifiants ou redondants par rapport au problème discriminant étudié [31]. Pour éviter cette «malédiction de la dimensionnalité», la sélection des gènes joue un rôle crucial dans l'analyse des puces à ADN.

La sélection de gènes pertinents pour la classification des échantillons est une tâche courante dans la plupart des études d'expression génique, où les chercheurs tentent d'identifier le plus petit ensemble possible de gènes capables d'atteindre de bonnes performances prédictives (par exemple, utilisation diagnostique en pratique clinique [5].

a- Du point de vue de l'analyse discriminante : nous avons quelques raisons d'effectuer la sélection des gènes : [30]

- Premièrement, parmi l'ensemble important de gènes, beaucoup pourraient être non pertinents, insignifiants ou redondants à un problème discriminant spécifique. Des études ont montré qu'un petit sous-ensemble de gènes pourrait être suffisant pour un problème biologique particulier.
- Deuxièmement, la sélection de gènes réduit le volume de données et facilite la manipulation et l'analyse des données de biopuces.
- Troisièmement, la réduction du nombre de gènes réduit la demande pour un grand nombre d'échantillons d'apprentissage, car la performance d'un classificateur de motifs dépend en partie du rapport entre le nombre d'échantillons et le nombre de caractéristiques. La collecte d'un nombre assez important d'échantillons dans la technique des biopuces est coûteuse, longue et même impossible.

b- Du point de vue des biologistes : l'importance de la sélection des gènes réside dans sa contribution à la compréhension des maladies et des fonctions de gènes particuliers, et à la conception d'expériences de biopuces à des fins de diagnostic clinique et de pronostic.

2.3.2 Travaux existants

Dans le contexte de l'analyse des données d'expression génique, plusieurs approches de sélection de gènes ont été publiées.

- Golub et al. [8] et Furey et al. [7] ont utilisé un score individuel de classement des gènes, facteur de pondération, pour effectuer la sélection des gènes avant la classification.

- Li et al. [20] ont proposé une méthode hybride de l'algorithmes génétique et K plus proche voisin, pour l'évaluation des gènes et la classification des échantillons. L'idée principale de GA / KNN est de trouver un grand nombre de sous-ensembles optimaux ou quasi-optimaux et d'évaluer l'importance des gènes pour la classification en examinant la fréquence des appartenances aux gènes dans ces sous-ensembles.

- Guyon et al. [9] ont introduit un algorithme d'élimination récursive des fonctions récursives (RFE), dans lequel les caractéristiques sont successivement éliminées lors de l'apprentissage d'une séquence de classificateurs de machines vectorielles de support (SVM).

- Xin Zhou ET K. Z. Mao [30] ont introduit un nouveau critère, appelé LS Bound, pour aborder le problème de sélection des gènes. Cette méthode peut être considé-

rée comme une hybridation entre les méthodes par filtre et enveloppe. D'une part, la mesure LS Bound est dérivée de la procédure leave-one out des LS-SVM.

— Ramón Díaz-Uriarte et al. [6] ont étudié la forêt aléatoire pour la classification des données de biopuces (y compris les problèmes multi-classes) et proposons une nouvelle méthode de sélection des gènes dans les problèmes de classification basés sur la forêt aléatoire.

2.4 Conclusion

Dans ce chapitre, nous avons présenté les aspects fondamentaux du domaine de sélection d'attributs qu'il devient de plus en plus un sujet de recherche plus actif et indispensable en raison de la multiplication des données. Dans la section suivante, on va introduire notre méthode utilisée pour la réduction de dimensionnalité des données des puces à ADN basée sur la sélection des gènes les plus pertinents.

CHAPITRE

3

PROGRAMMATION GENETIQUE MULTI-GENE ET CONCEPTION

3.1 Introduction

Dans le chapitre précédent, nous avons présentés les différentes méthodes de sélection des gènes, et notre problème posé de sorte que le nombre de variables(gènes) est très élevé. Nous avons concentré notre étude sur la sélection des variables pour la construction d'un bon prédicteur avec un nombre minimal des variables les plus pertinents. Dans ce chapitre nous présentons notre méthode que nous avons choisi pour résoudre notre problème de sélection des gènes, puis la conception de notre travail. D'abord la conception générale puis la conception détaillée en spécifiant les différents éléments composant notre travail et précisant son fonctionnement.

3.2 Programmation génétique multi-gène (MGGP)

3.2.1 Programmation génétique

La programmation génétique est l'une des techniques les plus célèbres, développée en 1992, qui sert à résoudre les problèmes de régression symbolique. Le modèle de prédiction est développé à la base d'un apprentissage adaptatif sur un certain nombre de cas de données fournies. Il imite l'évolution biologique des organismes vivants et utilise les principes des algorithmes génétiques (AG) [23].

Dans l'analyse de régression traditionnelle, l'utilisateur doit spécifier la structure du modèle, alors que dans GP, la structure et les paramètres du modèle mathématique évo-

luent automatiquement. Il fournit une solution sous la forme d'une structure arborescente ou sous forme d'une équation compacte utilisant l'ensemble de données donné [23].

Le modèle GP est composé de nœuds, ce qui ressemble à une structure arborescente, et donc, il est également connu sous le nom d'arbre GP.

Les nœuds sont les éléments provenant d'un ensemble fonctionnel ou d'un ensemble de terminaux.

Un ensemble fonctionnel peut inclure des opérateurs arithmétiques (+, * , ou -), des fonctions mathématiques (sin (.), Cos (.), Tan (.) Ou ln (.)), Des opérateurs booléens (AND, OR, NOT , etc.), les expressions logiques (IF, ou THEN) ou toute autre fonction appropriée définie par l'utilisateur.

Le jeu des terminaux comprend des variables (comme X_1, X_2, X_3, etc.) ou des constantes (comme 3, 5, 6, 9, etc.) ou les deux. Les fonctions et les terminaux sont choisis au hasard pour former un arbre GP avec un nœud racine et les branches s'étendant de chaque nœud fonction aux nœuds terminaux comme il est montré dans la Figure (3.1).

L'arbre GP, comme elle est illustrée dans la la Figure(3.1), présente une expression mathématique : tan ($6.5X_2$ / X_1) telle que :

— les variables : X_1, X_2 et le constant 6.5 constituent les nœuds terminaux,

— les opérateurs arithmétiques x, / et la fonction mathématique *tan* , constituent les nœuds fonctionnels.

— Le nœud fonctionnel de départ (tan) à partir duquel la branche des autres nœuds commence avec l'arborescence GP est appelé nœud racine [23].

3.2.1.1 Principe

a) Population initiale

Dans la première étape de la programmation génétique, un certain nombre d'arbres GP sont générés en sélectionnant au hasard des fonctions et des terminaux définis par l'utilisateur. Ces arbres GP forment la population initiale [23].

b) Reproduction

Dans la deuxième étape du GP, une partie de la population initiale est sélectionnée et copiée à la génération suivante et cette procédure est appelée reproduction [23].

c) Croisement

Dans une opération de croisement, deux arbres GP (Parent1 et Parent2) sont sélectionnés aléatoirement de l'ensemble des individus sélectionné pour l'opération de croisement [23].

Un nœud de chaque arbre est sélectionné de manière aléatoire, les sous-arbres sous les nœuds sélectionnés sont échangés et deux descendants (descendant1 et descendant2) sont générés [23].

Un exemple d'opération de croisement est illustré à la Figure (3.2) [23].

d) Mutation

Dans l'opération de mutation, un arbre GP est d'abord sélectionné au hasard dans la population, tout nœud de l'arbre est remplacé par un autre nœud de la même fonction ou du même ensemble de terminaux [23].

Un nœud de fonction ne peut remplacer qu'un nœud de fonction et le même principe s'applique aux nœuds terminaux [23].

Un exemple d'opération de mutation est montré sur la figure (3.3) dans laquelle le nœud fonctionnel "/" de l'arbre GP représentant une expression mathématique : tan (X_1 / X_2) est remplacé par un autre nœud fonctionnel, "x" et ainsi, un nouveau Arbre GP représentant une expression mathématique : tan (X_1 x X_2) est produit [23].

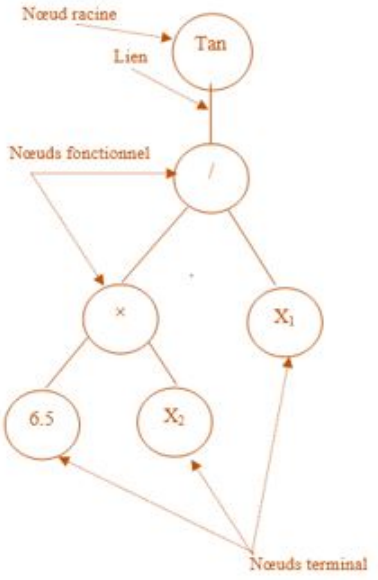


FIGURE 3.1: Example d'arborescence GP : Tan($6.5X_1/X_2$) [23].

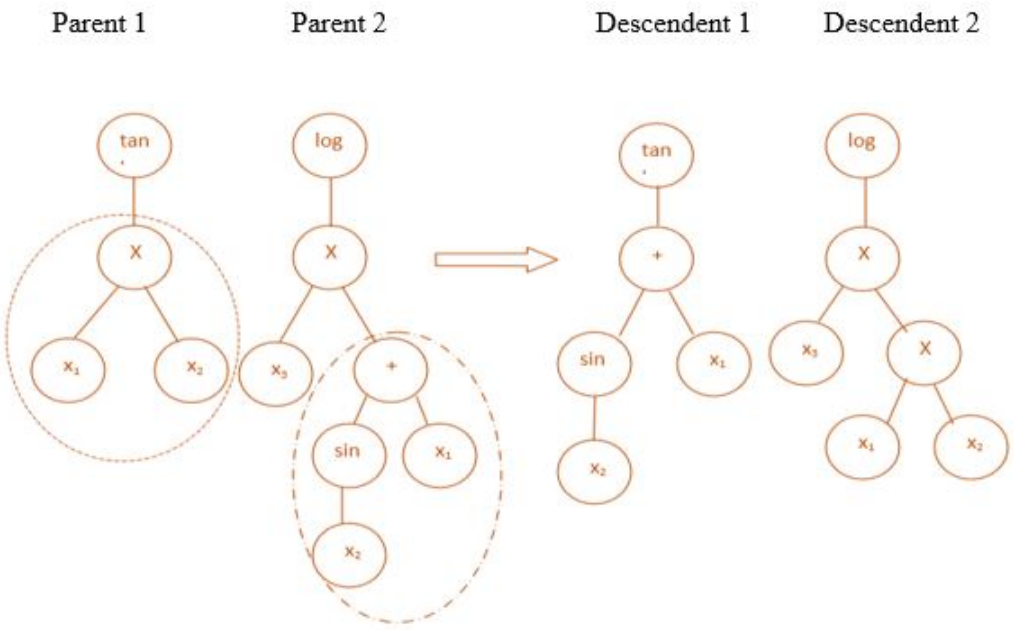


FIGURE 3.2: Opération de croisement dans GP [23].

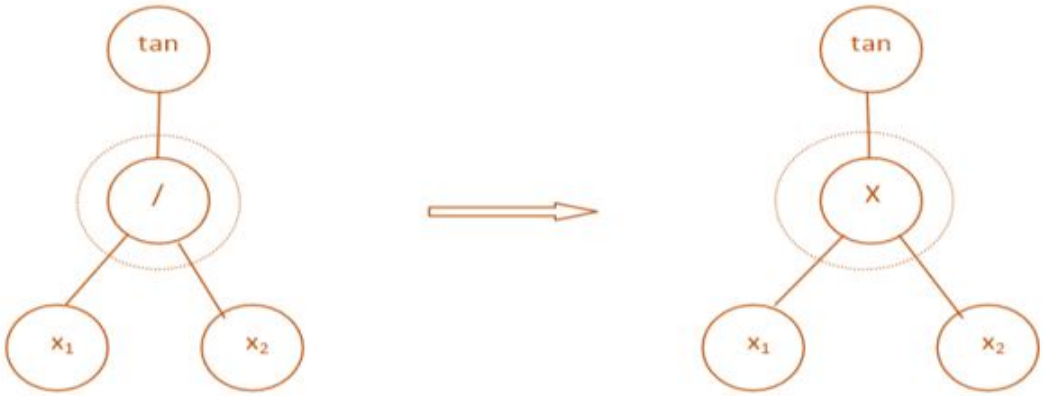


FIGURE 3.3: Opération de mutation dans GP [23].

3.2.2 Régression symbolique multi-gène

3.2.2.1 Régression symbolique naïve

Souvent, GP est utilisée pour faire évoluer une population d'arbres, dont chacun est interprété directement comme une équation mathématique symbolique qui prédit un (N) vecteur de sorties / réponses y, où N est le nombre d'observations de la variable de réponse y. La matrice d'entrée correspondante X est une matrice de données (NM) où M est le

nombre de variables d'entrée.

En général, seulement un sous-ensemble des variables M sera sélectionné par GP pour former les modèles. Dans la régression symbolique naïve, la $i^{\text{ème}}$ colonne de X comprend les N valeurs d'entrée pour la variable $j^{\text{ème}}$ et est désignée par la variable d'entrée X_j. La figure (3.4) illustre la régression symbolique naïve [27].

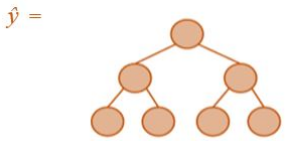


FIGURE 3.4: Régression symbolique naïve [27].

3.2.2.2 Régression symbolique linière

Pour améliorer l'efficacité de la régression symbolique, un terme de biais (offset) b_0 et un terme de pondération b_1 peuvent être utilisés pour modifier la sortie de l'arbre afin qu'elle soit mieux ajustée.

Les valeurs de ces coefficients sont déterminées par la méthode des moindres carrés et, pour tout arbre donné, la prédiction est garantiée au moins aussi bonne que la prédiction naïve. Il sera presque toujours meilleur (le seul cas où ce n'est pas le cas $b_0 = 0$ et $b_1 = 1$) [27].

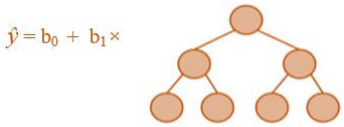


FIGURE 3.5: Régression symbolique linière [27].

Par conséquent, la prédiction de y est donnée par :

$$\hat{y} = b_0 + b_2\mathbf{t} \tag{3.1}$$

Où t est le vecteur (N * 1) des sorties de l'arbre GP sur les données d'apprentissage. Cela peut être aussi écrit comme suit :

$$\hat{y} = \mathbf{D}b \tag{3.2}$$

Où b est un vecteur (2 * 1) comprenant les coefficients b_0 et b_1, D est une matrice (N * 2) où la 1ère colonne est une colonne de 'uns' (ceci est utilisé comme entrée de biais) et le La 2ème colonne est la sortie de l'arbre t.

L'estimation optimale des moindres carrés linéaires de b est calculée à partir de y et D en utilisant l'équation normale des moindres carrés connue (3.3) où DT est la matrice transposée de D [27].

$$\hat{y} = (D^T D)^{-1} D^T y \tag{3.3}$$

3.2.2.3 Régression symbolique multi-gène

Une généralisation de l'approche précédente consiste à utiliser les arbres G pour prédire les données de réponse y. Là encore, il existe un coefficient de décalage b_0 et maintenant les coefficients b_1, b_2, ..., b_G sont utilisés pour mettre à l'échelle la sortie de chaque arbre/gène [27].

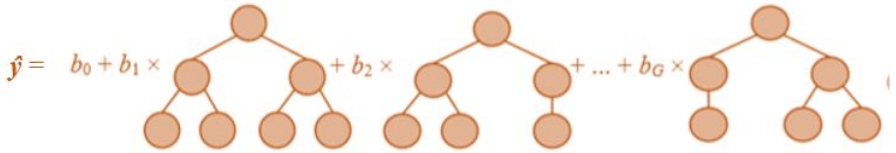


FIGURE 3.6: Régression symbolique multi-gène [27].

La prédiction des données d'entraînement y est donnée par :

$$\hat{y} = b_0 + b_1 t_1 + b_2 t_2 + ... + b_G t_G \tag{3.4}$$

Où t_i est le vecteur (N * 1) des sorties de IIème arbre/gène dans l'individu multi-gène.

Ensuite, définissez G comme une matrice de réponse de gêne (N * (G + 1)) comme suit :

$$G = [1 t_1 t_2 ... t_G] \tag{3.5}$$

Où le 1 se réfère à une colonne (N * 1) de uns utilisées comme entrée de biais.

Alors l'équation (3.4) peut être réécrit comme :

$$\hat{y} = \mathbf{G}b \tag{3.6}$$

L'estimation des moindres carrés des coefficients b_0, b_1, b_2, ..., b_G est formulés comme un vecteur ((G + 1) 1) qui peut être calculée, à nouveau, à partir des données d'apprentissage comme suit :

$$\hat{y} = (G^T G)^{-1} G^T y \tag{3.7}$$

3.2.3 Programmation Génétique multi-gène (MGGP)

MGGP (ou régression symbolique) est une variante de la GP, conçue pour développer un modèle mathématique empirique,sous forme d'une combinaison linéaire pondérée d'un certain nombre d'arbres GP [23].

Chaque arbre (gène) représente des transformations non linéaires d'ordre inférieur des variables d'entrée . Le terme "Multi-gène" se réfère alors à la combinaison linéaire de ces gènes [23].

La figure (3.6) montre un exemple de modèle MGGP où la sortie est représentée comme une combinaison linéaire de deux gènes (Gene 1 et Gene 2), développés en utilisant quatre variables d'entrée (X_1, X_2, X_3, X_4).

Chaque gène est un modèle non linéaire car il contient des termes non linéaires (sin (.) / log (.)).

Les coefficients linéaires (poids) c_1 et c_2, de Gene 1 et Gene 2 respectivement, et le biais (c_0) du modèle sont obtenus à partir des données d'apprentissage en utilisant une analyse de régression statistique (méthode des moindres carrés ordinaires) [23].

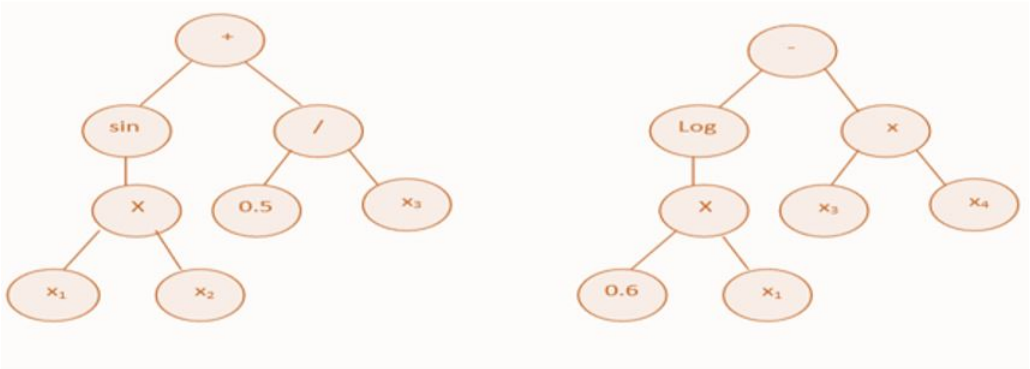


FIGURE 3.7: Un exemple de modèle GP multi-gènes [23].

3.2.3.1 Principe

Après la définition de certains paramètres, le processus de MGGP contient deux étapes : initialisation et évolution. Dans la première étape, une population initiale d'individus est créée de façon aléatoire, à partir des fonctions et des variables définies par l'utilisateur. Dans la deuxième étape, un processus de quelques étapes sera répété jusqu'à atteindre des critères spécifiques : calculer la valeur de fitness de chaque individu, sélectionner les meilleurs comme parents, reproduire de nouveaux individus par des opérateurs génétiques (croisement, mutation et sélection) et enfin remplacer les parents les plus faibles par les plus forts (voir Figure 3.8).

Il y a quelques mécanismes spéciaux de croisement de MGGP, qui permettent l'échange de gènes entre individus [23].

a) Croisement de 2 points de haut niveau :

Une opération de croisement à deux points de haut niveau permet l'échange de gènes entre deux individus parents dans la population et peut être expliquée par un exemple où le premier parent a quatre gènes [G1, G2, G3, G4] et les seconds trois gènes [G5, G6, G7] avec Gmax comme 5. Deux points de croisement sont choisis de manière aléatoire pour chaque parent et les gènes entourés de points de croisement sont désignés par ....

$[G_1,\{G_2,G_3, G_4\}]$, $[G_5, G_6, \{G_7\}]$

Les gènes entourés par les points de croisement sont échangés et ainsi, deux individus descendant sont créés comme indiqué ci-dessous :

$[G_1,\{G_7\}]$,$[G_5, G_6,\{G_2, G_3, G_4\}]$

Si l'échange de gènes conduit à un individu contenant plus de gènes que G_{max}, les gènes sont sélectionnés au hasard et éliminés jusqu'à ce que l'individu contienne des gènes G_{max} [23].

b) Croisement de bas niveau

Le croisement de sous-arbre GP standard est appelé croisement de bas niveau. Dans cette opération, d'abord un gène est choisi au hasard parmi chacun des individus parents (deux quelconques) dans l'ensemble des individus sélectionné pour l'opération de croisement, puis un échange de sous-arbres sous des nœuds arbitrairement sélectionnés de chaque gène est effectué. Les arbres résultants remplacent les arbres parents chez les individus parents par ailleurs inchangés, qui produisent ensuite des individus de descendance pour la génération suivante sans aucune délétion de gènes [23].

c) Mutation

De même, MGGP fournit des méthodes de mutation pour les gènes comme :(i) mutation de sous-arbre, (ii) substitution d'un nœud d'entrée aléatoirement nœud d'entrée sélectionné, (iii) substituer une constante sélectionnée aléatoirement avec une autre constante générée aléatoirement (iv) réglage de la constante sélectionnée au hasard [23].

3.2.3.2 Paramètres

MGGP nécessite de définir certains paramètres, tels que : nombre d'itération, le nombre maximum de gènes autorisés, dans un individu (G_{max}), profondeur maximale d'arbres (D_{max}), les probabilités d'événement de croisement et de mutation et méthode de sélection, les probabilités d'événement de croisement et de mutation et méthode de sélection er aussi les fonction mathématique (+, -, cos, sin ...) qui peuvent représenter les nœuds de sous-arbres.

La capacité du modèle à développer par MGGP est affectée par la sélection de ces paramètres de contrôle. Le nombre d'individus dans la population est fixé par la taille de la population. Le nombre de générations est le nombre de fois que l'algorithme est utilisé avant la fin de l'exécution. La taille de la population et le nombre de générations dépendent souvent de la complexité des problèmes. L'augmentation de la valeur G_{max} et de la valeur D_{max} augmente la valeur de fitness des données d'entraînement, tandis que la valeur de fitness des données de teste diminue, ce qui est dû au sur-apprentissage des données d'apprentissage. La capacité du modèle développé diminue.

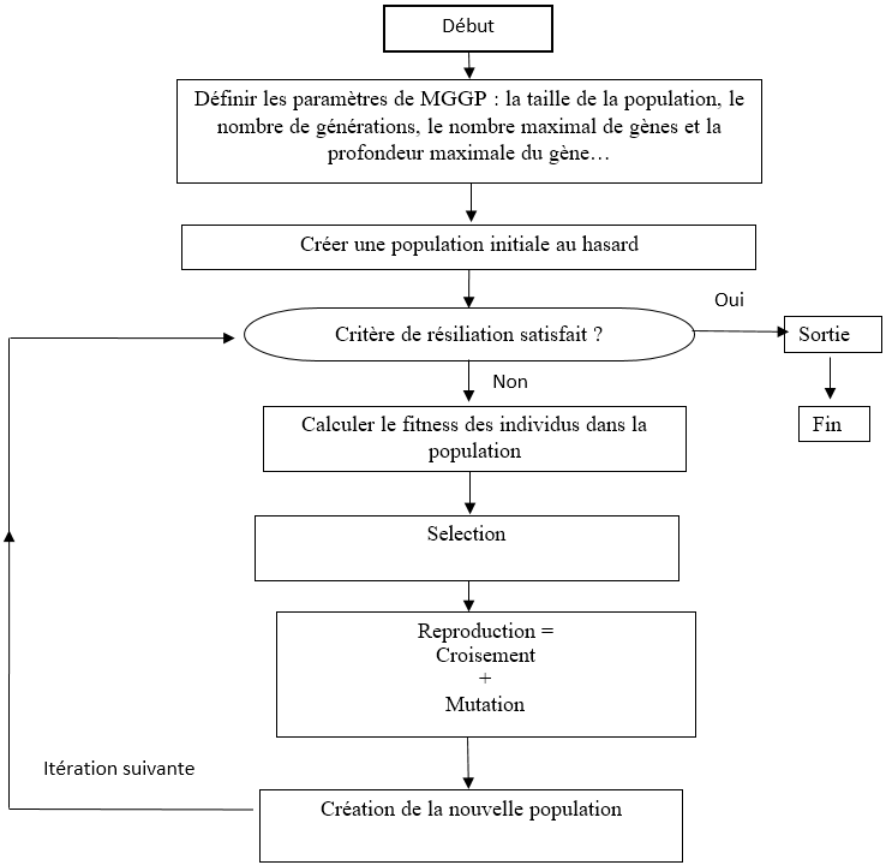


FIGURE 3.8: Un diagramme pour une procédure de programmation génétique multi-gènes [23].

3.3 Conception

3.3.1 Conception globale

Globalement, notre système est composé de trois modules principales :

1. **Module Prétraitement :** Les données utilisés dans notre travail sont des données brutes résultant de l'analyse de l'image des puces (fichiers CEL). Une phase de prétraitement est alors nécessaire pour qu'on peut les utiliser dans notre problème.

2. **Module de sélection des gènes :** ce module représente le coeur de notre travail, dont, on a esseyé de résoudre le problème de la grande dimensionnalité des données biopuces par la sélection des gènes pertinents, en exploitant les capacités de la MGGP.

3. **Module de visualisation des résultats :** ce module nous permet de présenter les résultats d'application de la MGGP pour la sélection des gènes .

L'architecture générale de notre système est présentée dans la figure suivante :

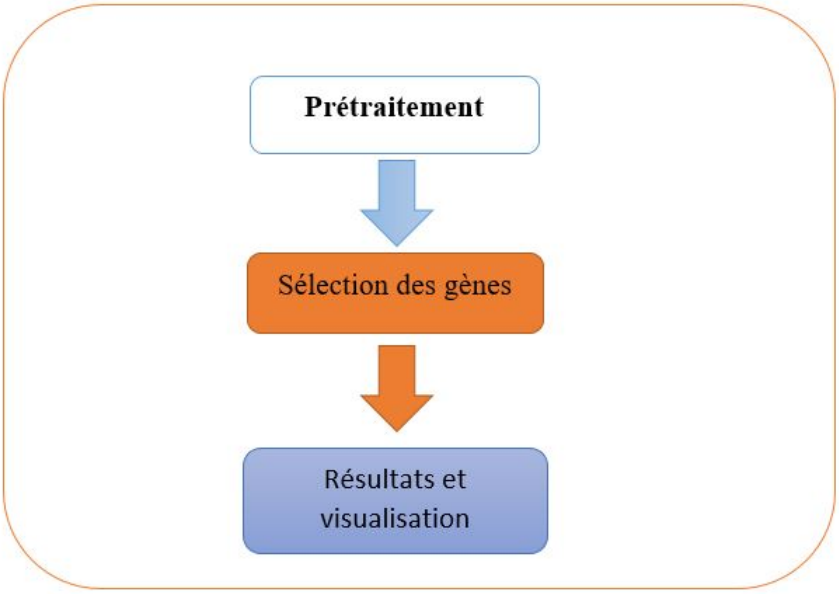


FIGURE 3.9: Architecture globale de notre travail.

3.3.2 Conception détaillée

3.3.2.1 Prétraitement

i) Fichier CEL

Les données brute de puce à ADN sont stockées dans un fichier à l'extension CEL.On obtient dans ces fichiers une quantité énorme d'informations. On a pour chaque gène : la moyenne des intensités de tous les pixels sur la zone correspondante au gène, la médiane de ces intensités, l'écart-type de ces intensités et le nombre de pixels dans la zone considérée. Certaines informations additionnelles telles que l'identifiant associant une sonde est stocké dans un fichier **CDF**.

ii)Étape de prétraitement (transformation des données)

L'étape de transformation des données passe par 2 étapes :

a) Correction du bruit de fond (Background Correction)

- Estimation de bruit global (Constant Background Correction) par l'utilisation de la moyenne ou la médiane d'intensité des gènes.
- Estimation du bruit local (Local Background Correction) par l'utilisation les pixels se trouvant près du spot.
- Estimation du bruit de fond (Morphological Opening) par l'utilisation des méthodes non-linéaire.

Nous avons utilisé la 1ère méthode dans notre travail qui est la plus utilisée.

b) Normalisation
pour recentrer la distribution des données et la rendre symétrique,Nous avons appliqué une transformation logarithmique à base deux.

Le résultat de prétraitement une matrice d'expression des gènes filtrer et normaliser qu'elle nous permet de sélectionner les gènes les plus pertinents.

Gène id	Echantillon 1	Echantillon 2	....	Echantillon N
Gèene 1	m_{11}	m_{12}	....	m_{1N}
Gèene 2	m_{21}	m_{22}	....	m_{2N}
Gèene 3	m_{31}	m_{32}	....	m_{3N}
....	....	....	....	.....
Gèene M	m_{M1}	m_{M2}	....	m_{MN}

TABLE 3.1: Matrice d'expression des gènes normaliser et filtrer .

3.3.2.2 Sélection des gènes

Dans notre travail, nous avons utilisé, MGGP pour sélectionner le meilleur modèle de sélection d'attributs/gènes. Pour cela, l'algorithme MGGP est exécuté sur les données d'apprentissage (BA) et vérifié sur des données de test (BT). Le modèle optimal est sélectionné en fonction sa simplicité ainsi que de sa performance sur les données d'apprentissage.

L'application du l'algorithme MGGP, pour résoudre le problème de sélection d'attributs/gènes, nécessite la description des données d'apprentissage, la représentation d'une solution solutions et des fonctions objectif à optimiser. Le processus de sélection des gènes est décrit selon le schéma présenté dans la figure 3.11.

a) Données d'apprentissage
Dans notre travail, nous nous somme intéressée aux jeux de données de puces à ADN décrivant le niveau d'expression de gènes mesuré sur des tissus réparties en deux classes (tissus normaux et tissus cancéreux).

Les échantillons normaux de la classe1 possèdent un étiquette y=1, et ceux de la classe 2 (échantillons cancéreux) possèdent un étiquette y=-1.

b) Représentation d'une solution

Une solution Y dans notre problème sous la forme d'une régression linéaire où Y désignée comme une structure d'arbres, également appelée gènes, reçoit les variables d'entrées

X et tente de prédire la variable de sortie Y.

Chaque prédiction de la variable de sortie Y est formée par une sortie pondérée de chacun des arbres / gènes dans la solution plus un terme de biais.

Chaque solution Y peut contenir (de manière aléatoire) entre 1 et G_{max} sous-arbres/gènes, et la hauteur de chaque sous-arbre/gène est entre 1 et D_{max}. Les feuilles des sous-arbres sont les numéros des gènes (attributs) et les nœuds sont les opérateurs mathématiques (-, +,...).

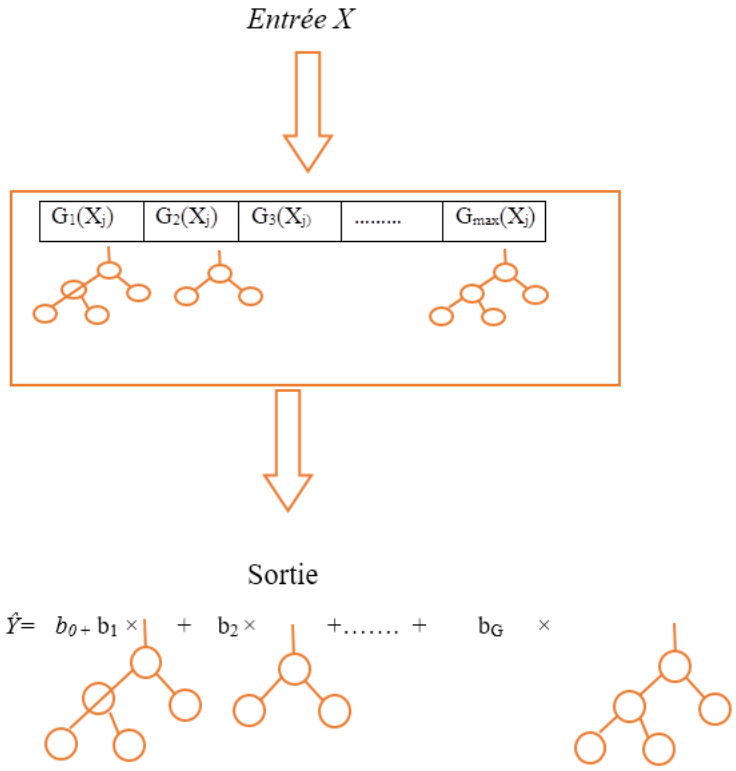


FIGURE 3.10: Présentation d'une solution = individu [1,Gmax] gènes (structure d'arbres) [23].

c) Choix de fonction objectif

Deux fonctions objectif (f_1 et f_2) sont considérées, pour sélectionner le meilleur modèle de sélection d'attributs/gènes : la complexité du modèle et sa performance sur la base d'apprentissage. La complexité du modèle f_1 est définie comme la somme des nœuds de tous les sous-arbres d'un arbre. Minimiser cette complexité f_1 est utilisé pour évaluer sa simplicité.

La performance du modèle, sur la base d'apprentissage, est déterminée par la maximisation de sa qualité d'ajustement qui est représentée par le coefficient R^2. Ou bien la minimisation de f_2 :

$$f_2 = 1 - R^2 \tag{3.8}$$

Où R^2 est le coefficient de détermination : une mesure de la qualité de la prédiction d'une régression linéaire.

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3.9}$$

Où : n c'est le nombre total d'exemples, y_i la valeur de l'exmple i, $\hat{y}_i$ la valeur prédite correspondante et $\bar{y}$ la moyenne des exemples.

Donc, nous avons un résultat de 2 fonction objectif à optimiser :

1) Minimisation du critère statistique (3.8)
2) Minimisation de la complexité de modèle (un nombre minimal des nœuds).

d) Création d'une population initiale

La création d'une population initiale se fait par la génération aléatoire de N solutions (modèles). Pour chaque modèle, les fonctions représentant les nœuds sont choisies, de façon aléatoire, parmi un ensemble des opérateurs défini par l'utilisateur. La même chose, pour les numéros de gènes (attributs) représentant les feuilles sont choisis de façon aléatoire parmi l'ensemble des numéros de gènes existe dans la base.

e) Evaluation et sélection des modèles

Nous avons basé, sur le principe de algorithme d'optimisation multi-objectif NSGA-2 (Non dominated Sorting Genetic Algorithm2) [4] pour l'évaluation et la sélection des modèles (solutions) parents. Dans ce cas les solutions sont classées selon leurs rangs (fronts F_i) de dominance. Les meilleures solutions vont se retrouver dans le ou les premiers fronts.

Une nouvelle population de parents (P_t+1) est formée en ajoutant les premiers fronts au complet (premier front F_1, second front F_2, etc.) tant que ceux-ci ne dépassent pas la moitié de la taille de population (N/2).

Si le nombre d'individus présents dans (P_t+1) est inférieur à (N/2), une procédure de crowding est appliquée sur le premier front suivant (F_i), non inclus dans (P_t+1). Le but de cet opérateur est d'insérer les (N/2 - |P_t+1|) meilleurs individus qui manquent dans la population (P_t+1) [4].

Une fois que la première moitié des individus appartenant à la population (P_t+1) sont identifiés, la moitié restante de la population P_t+1 est créée par la sélection de tournoi de crowding.

f) Reproduction

La nouvelle population d'enfants est produite par l'opérateur de croisement et suivi par l'operateur de mutation selon les principes décrits dans les sections : a, b et c .

g) Critère d'arrêt

Afin d'arrêter la procédure de la MGGP, on a choisi d'utiliser un critère d'arrêt basénsur un nombre prédéni d'itérations.

h) Evaluation du modèle obtenu

Pour évaluer les performances de généralisation de modèle de résultat de MGGP, nous avons séparer l'ensemble des données à un ensemble d'apprentissage et un ensemble de test. La qualité du modèle est alors jugée par sa capacité de réduire **l'erreur de test** RMSE et maximise **le coefficient de détermination(corrélation)** R^2.

Pour évaluer le résultat obtenu (modèle obtenu par MGGP), nous avons utilisé une mesure statistique P-value.

pour prouver l'importance des gènes sélectionnés, nous avons utilisé la fréquence de ces gènes dans les meilleurs modèles dans la population que nous avons développée.

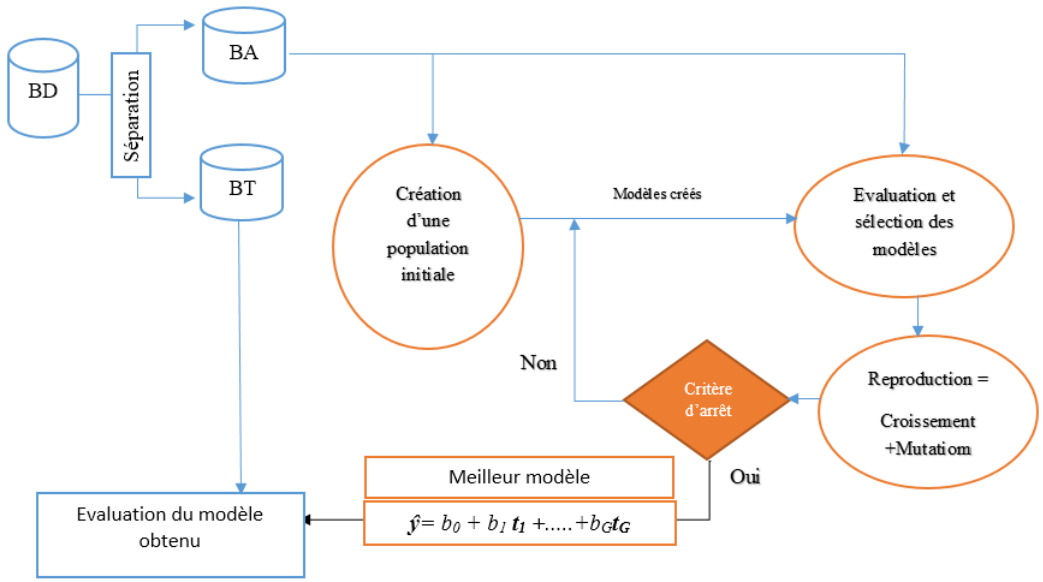


FIGURE 3.11: Processus de Sélection des gènes.

3.3.2.3 Résultats et visualisation

Apres la phase de selection des genes, nous avons presenté notre resultat obtenu tels que les noms des ses genes, les propriétés statistiques du MGGP avec une visualisation de ses propretiés statistiques permettant un bon analyse de nos résultats.

3.4 Conclusion

Dans ce chapitre, nous avons décrit le principe et le fonctionnement de la méthode de MGGP sur laquelle notre projet se base, puis nous avons présenté le processus de sélection de gènes par la méthode MGGP avec une description détaillée de ses composants.

Dans le chapitre suivant, nous passerons à l'implémentation du processus de sélection de gènes, en présentant les différents détails de sa réalisation.

CHAPITRE

4

IMPLEMENTATION

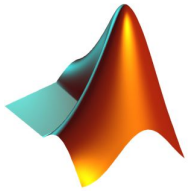
4.1 Introduction

Dans ce chapitre, nous allons présenter l'environnement de travail, le langage de programmation et les outils que nous avons utilisé pour construire le système. Puis nous allons présenter le jeu de donnée utilisé dans notre travail,un jeu de donnée publique qui est utilisé dans de nombreux travaux concernant l'analyse des données de puces à ADN. Et enfin nous allons présenter interface de notre application.

4.2 Environnement et outils de développement

Pour développer notre application et valider notre proposition, nous avons utilisé le langage Matlab et l'environnement Matlab. Pour la transformation et la normalisation des données nous avons utilisé la boîte à outils *Bioinformatics Toolbox* et particulièrement la catégorie *Microarray Analysis*. Pour la sélection des gènes et la visualisation des résultats nous avons utilisé la plateforme GPTIPS 2.

4.2.1 Environnement de développement



Matlab est un logiciel de calcul numérique commercialisé par la société MathWorks1. Il a été initialement développé à la fin des années 70 par Cleve Moler. Matlab est abréviation de MATrix LABoratory. Il est avant tout un programme de calcul matriciel. Il est un langage pour le calcul scientifique, l'analyse de données, leur visualisation, le développement d'algorithmes. Son interface propose, un environnement de développement intégré (IDE) pour la programmation d'applications. Le logiciel peut être complété par de multiples toolboxes, c'est-à-dire des boîtes à outils. Celles-ci sont des bibliothèques de fonctions dédiées à des domaines particuliers [29].

4.2.2 Outils utilisés

4.2.2.1 Bioinformatics Toolbox

Bioinformatics Toolbox est une boite à outils intégrer dans le Matlab, elle fournit des algorithmes et des applications pour le séquençage de nouvelle génération (NGS), l'analyse de puces à ADN, la spectrométrie de masse et l'ontologie génique. En utilisant les fonctions de la boîte à outils, vous pouvez lire des données génomiques et protéomiques à partir de formats de fichiers standard tels que SAM, FASTA, CEL et CDF, ainsi que de bases de données en ligne telles que NCBI Gene Expression Omnibus et GenBank®. La boîte à outils fournit également des techniques statistiques pour détecter les pics, imputer des valeurs pour les données manquantes et sélectionner des caractéristiques [1].

Nous nous sommes intéressés à l'analyse de puces à ADN, ç'est pour cela nous avons choisis la catégorie *Microarray Analysis* de *Bioinformatics Toolbox*.

Microarray Analysis est une catégorie de *Bioinformatics Toolbox*, elle est largement utilisé pour l'analyse de données sur microarray, y compris la lecture, le filtrage, la normalisation et la visualisation des données de microarray [1].

Nous avons utilisé la fonction de affyrma de *Microarray Analysis*.

Affyrma est un algorithme utilisé pour créer une matrice d'expression à partir de données CEL. Les valeurs d'intensité brutes sont corrigées en arrière-plan, transformées et normalisées en log_2.

Paramètre d'entrée d'Affyrma

CELFile : c'est le nom de fichier CEL peut être l'une des valeurs suivantes : vecteur des caractères spécifiant un seul nom de fichier CEL,'*' qui lit tous les fichiers CEL dans le dossier en cours, " qui ouvre la boîte de dialogue Sélectionner les fichiers CEL à partir de laquelle on sélectionne les fichiers CEL.

CDFFile : c'est le nom de fichier CDF peut être l'une des valeurs suivantes : vecteur des caractères spécifiant un seul nom de fichier CDF,' ' qui ouvre la boîte de dialogue Sélectionner le fichier CDF à partir de laquelle on sélectionne le fichier CDF.

CELPathValue :c'est le chemin et le dossier où les fichiers spécifiés dans CELFiles sont stockés.

CDFPathValue :c'est le chemin et le dossier où le fichier spécifié dans CDFFile est stocké.

MedianValue :on Spécifie l'utilisation de la médiane des valeurs classées au lieu de la moyenne pour la correction du bruit de fond, le choix est TRUE ou False (par défaut).

OutputValue :on Spécifie le choix de fonction de normalisation : log_2, log, log_{10}, linear, @functionname.

4.2.2.2 GPTIPS

GPTIPS est une plate-forme logicielle libre et open source basée sur MATLAB pour l'exploration de données symbolique (SDM). Il utilise une variante multigénique de la méthode d'apprentissage automatique de la programmation génétique (MGGP), inspirée par la biologie, comme moteur du processus de découverte automatique de modèles.

GPTIPS 2 est la dernière version de GPTIPS. Une variante simplifiée du mécanisme de croisement de gènes de haut niveau MGGP est proposée. De plus, des nouvelles fonctionnalités remarquables de cette version (a) fournit de nombreuses méthodes pour visualiser les propriétés des modèles symboliques(b) met l'accent sur la génération des bibliothèques graphiquement navigables des modèles qui sont optimaux en termes de surface de compromis de Pareto de la performance et de la complexité du modèle(c) utilisant une nouvelle analyse de visualisation centrée sur le gène pour atténuer la météorisation horizontale et réduire la complexité dans les modèles de régression symbolique multigénique [27].

a) Fichier de configuration

Dans ce fichier, les différents paramètres utilisés dans GPTIPS peuvent être spécifie par l'utilisateur.la liste des paramètres les plus utilisés est dans la table 4.1.

Paramètre	Description
gp.runcontrol.pop_size	Taille de population
gp.runcontrol.num_gen	Nombre de génération
gp.selection.elite_fraction	Elitisme valeur entre 0 et 1
gp.nodes.functions.name	Nom des fonctions utiliser
gp.operators.mutation.p_mutate	Probabilité de mutation d'arbre GP
gp.operators.crossover.p_cross	Probabilité de croisement d'arbre GP
gp.treedef.max_depth	Hauteur maximal de chaque arbre
gp.genes.max_genes	Nombre maximal de gènes par individu
gp.userdata.xtrain	Donnée d'entrainement
gp.userdata.xtest	Donnée de test
gp.userdata.ytrain	Y souhaiter d'entrainement
gp.userdata.ytest	Y souhaiter de test
gp.selection.tournament.p_pareto	Pareto tournoi

TABLE 4.1: Paramètres de GPTIPS .

```
function gp = ma_config(gp)

%run control
gp.runcontrol.pop_size = 500;
gp.runcontrol.num_gen=200

%selection
gp.selection.tournament.size = 20;
gp.selection.tournament.p_pareto = 0.3;
gp.selection.elite_fraction = 0.3;

%genes
gp.genes.max_genes = 4;
gp.treedef.max_depth = 1;

% train
str=importdata('p.txt');
d=str;
c=importdata('lib.txt');
c=c.';

[train, test] = crossvalind('holdOut',c,0.4);

d1=d(train,:);
c1=c(train,:);
x=d1;
y=c1;

gp.userdata.ytrain = y;
gp.userdata.xtrain = x;
%test
d2=d(test,:);
c2=c(test,:);
xx=d2;
yy=c2;
gp.userdata.ytest = yy;
gp.userdata.xtest = xx;
%function nodes
gp.nodes.functions.name = {'plus','tanh'};
end
```

FIGURE 4.1: Exemple de fichier de configuration.

b) Fonctions de GPTIPS

GPTIPS propose des fonctions permettant l'exécution le processus de MGGP et d'analyser les résultats des exécutions et certaines fonctions spécifiquement destinées à la régression symbolique multigène.une liste des fonctions est dans la table 4.2.

Fonction	Description
rungp	Le fichier de configuration peut être exécuté en utilisant la fonction rungp. La sortie de cette fonction est une structure gp contient le résultat de l'exécution.
summary	tracer la meilleure RMSE (valeurs en log) et la RMSE moyenne au cours d'exécution.
runtree	La génération des Diagrammes de dispersion de poids des gènes et de corrélation entre y et $\hat{y}$.
popbrowser	La présentation de Front de Pareto des modèles en terme (1-R^2) et la complexité des modèles.
gpmodel2sym	Extraction de modèle comme un objet mathématique symbolique

TABLE 4.2: Fonctions de GPTIPS.

4.3 Système de sélection des gènes proposé

Nous avons développé une application implémentant la proposition présenté dans le chapitre trois. L'application est composée à trois parties :

— Prétraitement (transformation des données).
— Sélection des gènes.
— Visualisation des résultats.

4.3.1 Ensemble des données utilisées

D'après nos recherches, nous avons utilisé des jeux de données publics, accessibles et utilisés dans des nombreux travaux concernant l'analyse des données de puces à ADN.

4.3.1.1 Cancer de la prostate

Cet jeu de données présente le niveau d'expression de 12600 gènes mesuré sur 102 Tissus (50 tissus normaux et 52 tissus cancéreux).

4.3.1.2 Cancer du poumon

Dans cet jeu de données, le niveau d'expression de 12600 gènes est mesuré sur 207 Tissus (17 tissus normaux et 190 tissus cancéreux).

4.3.1.3 Sarcome

Dans cet jeu de données, le niveau d'expression de 22283 gènes est mesuré sur 23 Tissus (15 tissus normaux et 8 tissus cancéreux).

Pour une description complète de ces jeux de données consulter l'adresse :

http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi

Les fichiers CEL de ces jeux de données sont organisés comme il est présenté dans la figure (4.2).

FIGURE 4.2: Exemple des données utilisé (CEL).

4.3.2 Prétraitement des données

Le module de prétraitement présenté dans la figure (4.3) nous permettons de transformer et normaliser les données utilisés, par l'utilisation de la fonction Affyrma de la boite outils Bioinformatics. Le résultat de cette fonction est montré dans la figure (4.4).

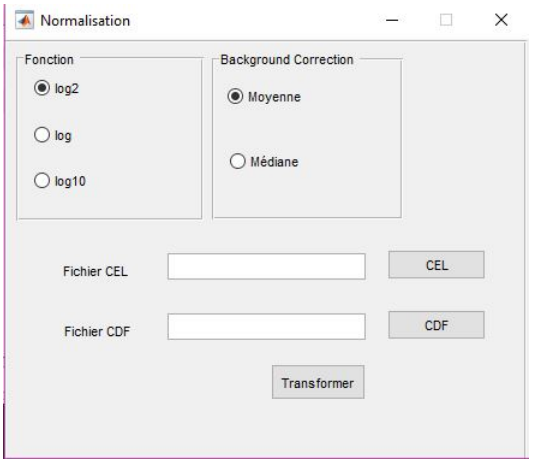


FIGURE 4.3: Interface du module de prétraitement.

	CL2001032212AA	CL2001032213AA	CL2001032214AA	CL2001032216AA	CL2001032227AA
AFFX-MurIL2_at	5.6147	5.8557	5.2494	5.4717	5.4212
AFFX-MurIL10_at	4.7569	4.9489	4.3067	4.4734	4.3944
AFFX-MurIL4_at	3.2623	4.1402	3.126	3.3234	3.1907
AFFX-MurFAS_at	4.1078	4.5543	3.8287	3.9093	4.0356
AFFX-BioB-5_at	4.2063	4.5537	4.0612	4.8949	4.7577
AFFX-BioB-M_at	4.1087	4.2101	3.9226	4.2386	3.9414
AFFX-BioB-3_at	3.8553	3.7783	3.415	3.7452	3.4373
AFFX-BioC-5_at	5.2633	5.2736	4.7337	5.1339	4.7238
AFFX-BioC-3_at	4.6224	4.7358	4.245	4.6798	4.2669
AFFX-BioDn-5_at	3.4355	4.2308	3.1433	3.4659	3.1632
AFFX-BioDn-3_at	9.0635	9.1214	8.6672	9.1451	9.0533
AFFX-CreX-5_at	3.3428	3.3686	3.1755	3.3716	3.2586
AFFX-CreX-3_at	5.3794	5.5425	4.9774	5.3002	4.9339
AFFX-BioB-5_st	4.8504	4.9349	4.52	5.1838	4.8635
AFFX-BioB-M_st	4.5776	4.7643	4.2107	4.587	4.216
AFFX-BioB-3_st	5.9324	6.1391	5.5797	5.8331	5.6235
AFFX-BioC-5_st	5.4393	5.964	5.3511	5.6492	5.3823
AFFX-BioC-3_st	3.7289	4.1509	3.6168	4.0574	3.6382
AFFX-BioDn-5_st	4.966	5.6188	4.8518	5.2761	4.8839
AFFX-BioDn-3_st	6.2993	6.1746	5.7476	6.1401	5.879

FIGURE 4.4: Résultat de prétraitement.

Ces données peuvent être visualiser à travers l'interface d'affichage présentée dans la figure (4.5). Cette interface permet de présenter :

— la base sélectionnée,
— le nombre des gènes,
— et des exemples.

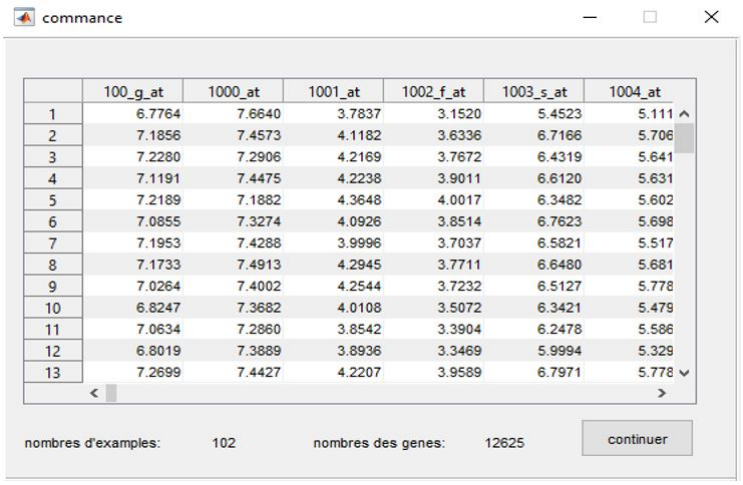


FIGURE 4.5: Interface d'affichage.

Une fois les données sont prétraités, On doit les charger à travers l'interface d'accueil, afin de commencer le module principale de notre système.

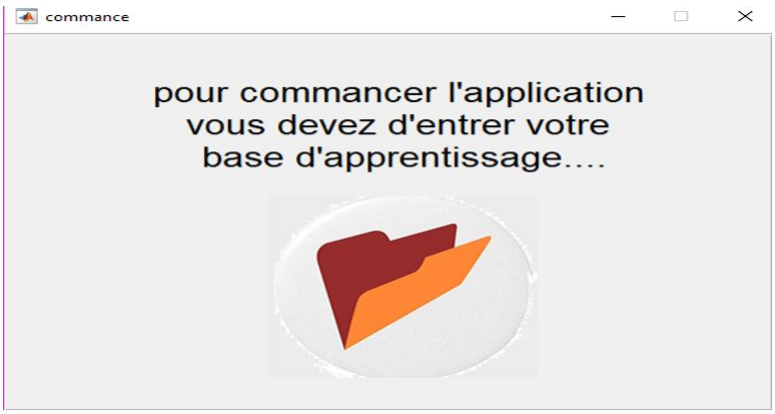


FIGURE 4.6: Interface d'accueil.

4.3.3 Sélection des gènes

Une fois les données sont chargés, l'interface présentée dans la figure (4.7) nous permettons de :

— Choisir le mode de test (utilisation la base d'entrainement, charger une base de test ou le mode Holdout).
— Faire l'opération de sélection des gènes.
— Afficher le résultat de sélection (le meilleur modèle, les noms des gènes sélectionnés).
— Présentation des mesures d'évaluation et le Temps d'exécution.

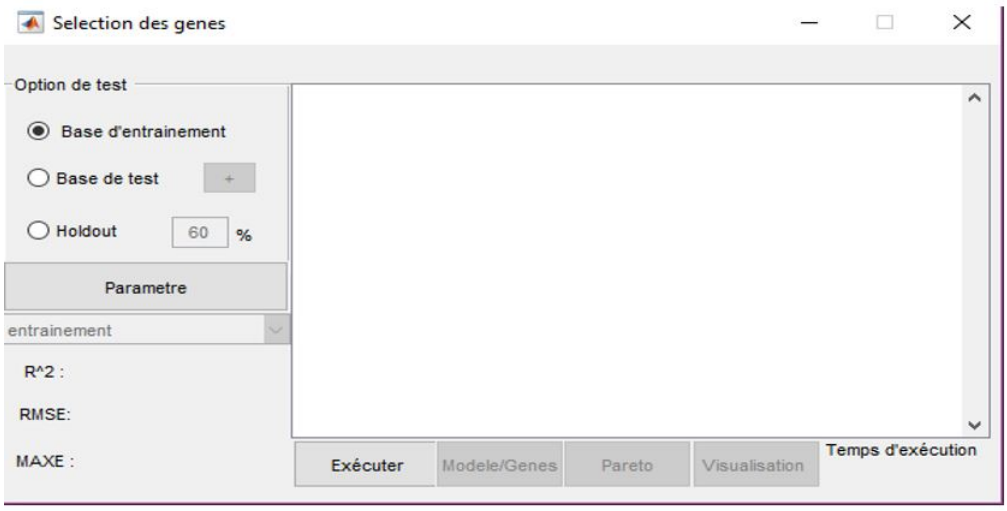


FIGURE 4.7: Interface principale de sélection.

Le réglage des paramètres de controle de MGGP est assurer à travers la fenêtre des paramètres présentée dans la figure suivante :

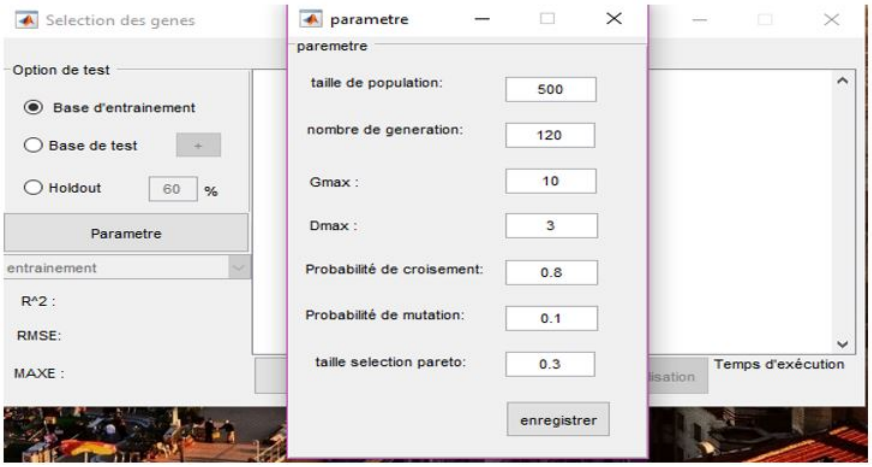


FIGURE 4.8: Interface des paramètres.

4.3.4 Visualisation des résultats

La visualisation des résultats d'application de MGGP est garantiée par les fenêtres suivantes :

4.3.4.1 Visualisation du processus de sélection

Cette interface montre un résumé de l'exécution du processus de sélection des gènes en terme de l'évolution de RMSE (valeur log) avec le nombre de générations.

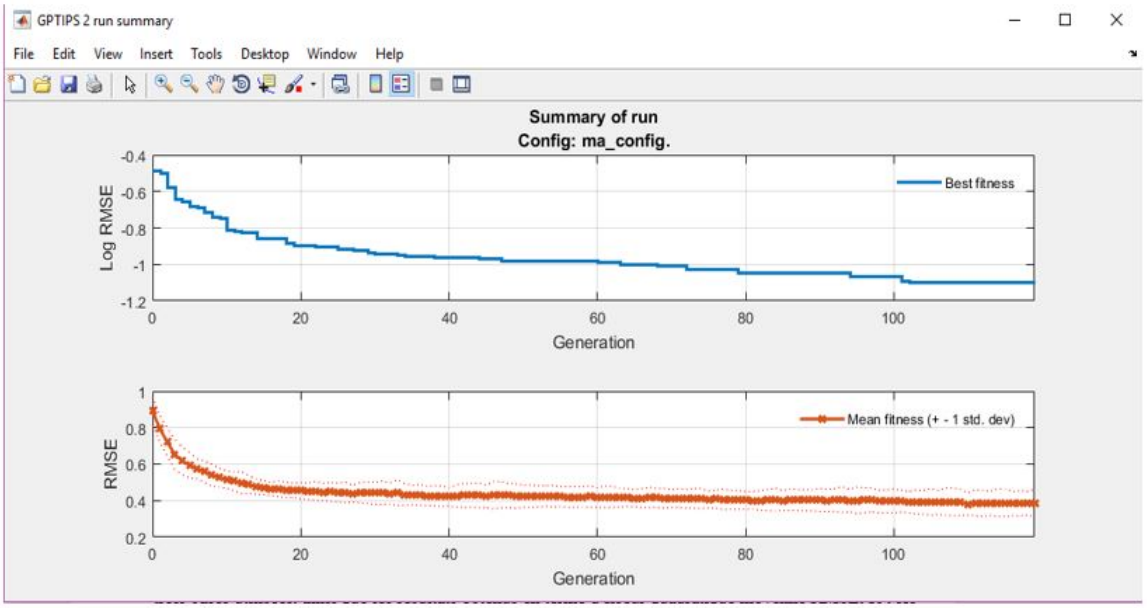


FIGURE 4.9: Interface de resumé de l'exécution.

4.3.4.2 Présentation des Modèles évolués

Cette interface montre la Population des modèles évolués en termes Pareto de complexité et de fitness 1-R^2.

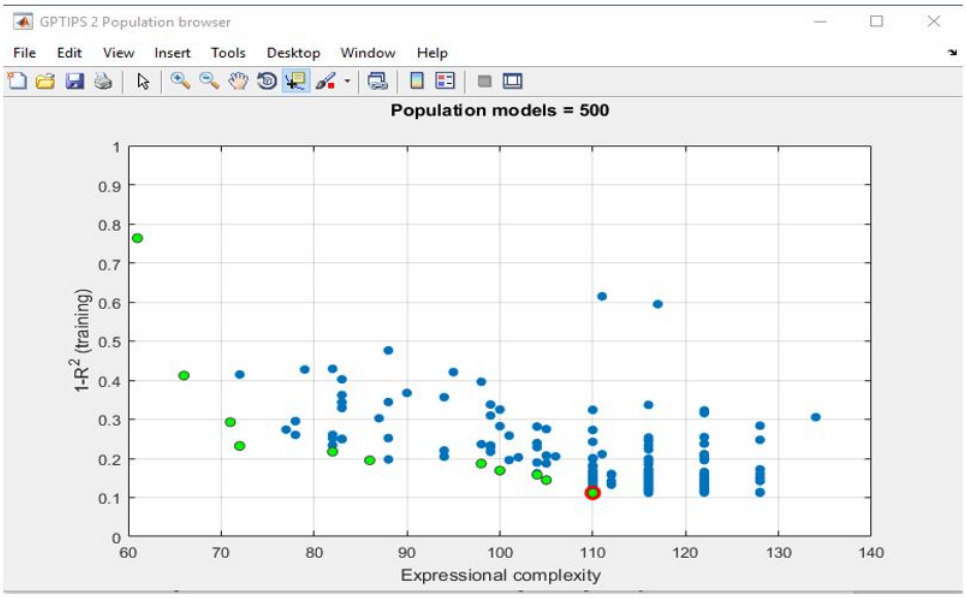


FIGURE 4.10: Interface de presentation les modèles évolués en termes Pareto(complexité/fitness).

4.3.4.3 Signification des gènes constituants les modèles développés

Cette interface montre les poids des gènes et la signification statistique de chacun des gènes de modèle développé en terme valeur P.

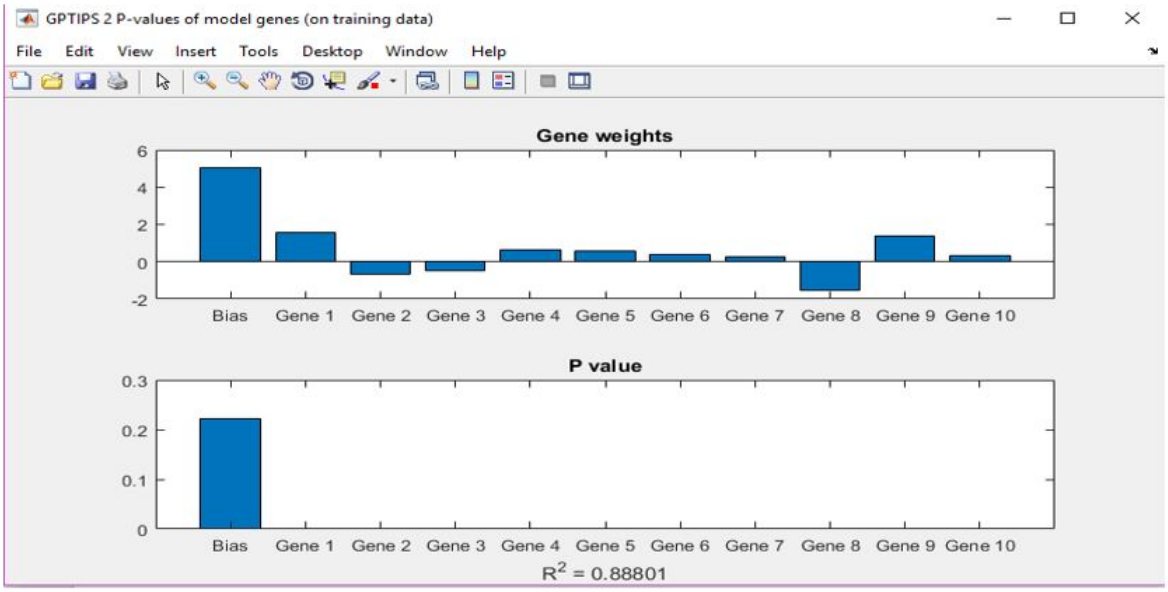


FIGURE 4.11: Interface de visualisation 1(Propriétés statistique).

4.3.4.4 Association entre valeurs réelles et prédites

Cette interface montre une indication visuelle du degré d'association les valeurs prédites $\hat{y}$ et réelles y.

FIGURE 4.12: Interface de visualisation 2(degree d'association les valeurs prédites $\hat{y}$ et réelles y).

4.4 Conclusion

Dans ce chapitre, Nous avons représenté l'implémentation de notre système : L'environnement et les outils de développement. Dans le chapitre suivant, nous avons présenté les expérimentations effectuées sur les trois bases utilisées, ainsi que les résultats obtenus en terme d'erreur quadratique moyenne RMSE, R^2, les diagrammes de dispersion, les histogrammes et quelque mesures statistiques.

CHAPITRE

5

EXPEREMENTATION ET RESULTATS

5.1 Introduction

Dans ce chapitre, nous allons expliquer les expérimentations que nous avons appliqué sur la méthode proposée et les résultats obtenus, en utilisant un jeu de donnée publique qui est utilisé dans de nombreux travaux concernant l'analyse des données de puces à ADN.

5.2 Expérimentations et résultats

5.2.1 Détermination des meilleurs Paramètres

Afin de déterminer les meilleurs paramètres de MGGP, plusieurs expérimentations sont faites. Les paramètres présentés dans le tableau (5.1) sont jugés comme les meilleurs aprés son application sur les trois bases : (a)le cancer de prostate, (b) du poumon et (c) du sarcome.

Paremetre	Valeur		
	(a)	(b)	(c)
Taille de population	500	500	500
Nombre de génération	200	200	200
G_{max}	8	8	8
D_{max}	6	7	6
données d'apprentissage	90%	90%	70%
données de test	10%	10%	10%
Probabilité de croisement	0.8	0.8	0.8
Probabilité de mutation	0.1	0.1	0.1
Fonction	$+, -$	$+, -$	$+, -$
Tournoi Pareto	0.3	0.3	0.3

TABLE 5.1: Valeurs optimales des paramètres de contrôle de MGGP.

5.2.2 Analyse des résultats

5.2.2.1 Meilleurs modèles

Aprés l'exécution de la technique MGGP par les paramètres cités au dessous, les modèles optimaux sélectionnés pour chaque expérimentation sont présentés dans les figures suivantes :

ŷ=2197*x283 - 0.3899*x630 - 0.3503*x1191 - 0.2197*x2459 + 0.2197*x2569 + 0.2331*x2622 - 0.2197*x3918 - 0.5942*x4038 + 0.09378*x4177 - 0.2331*x4225 - 0.09378*x4595 - 0.09378*x5097 - 0.2197*x5329 - 0.3095*x5482 - 1.798*x6102 + 0.2795*x6633 + 0.3503*x6694 - 0.2197*x6751 - 0.3503*x6869 - 0.5942*x6960 + 0.6693*x7161 - 0.3899*x7713 - 0.6096*x8168 - 0.2331*x8367 - 0.4992*x8748 - 0.09378*x8848 + 0.2197*x8961 - 0.1171*x9062 - 0.2197*x9514 - 0.3503*x9521 - 0.3899*x9649 - 0.2331*x9654 - 0.2197*x9813 + 0.2331*x9934 - 0.4904*x9939 + 0.8891*x10121 - 0.3899*x10331 + 0.5942*x10435 + 0.559*x10448 - 0.3899*x10533 + 0.2197*x10637 + 0.2795*x11098 + 0.5942*x11359 + 0.09378*x11484 + 1.245*x12236 - 0.1171*x12289 + 0.559*x12332 + 0.2197*x12333 - 0.3899*x12438 + 18.99

(a)

ŷ=0.2117*x90 + 0.2117*x267 - 0.09122*x274 + 0.09122*x1251 + 0.09121*x1270 - 0.2117*x1293 + 0.1838*x1621 + 0.6361*x1830 - 0.2117*x2013 - 0.2117*x2050 - 0.1873*x2745 + 0.09611*x2865 - 0.1838*x2903 - 0.1838*x2995 + 0.09122*x3321 + 0.3029*x3430 - 0.09121*x4002 - 0.3029*x4224 - 0.09122*x4279 - 0.09121*x4361 - 0.3422*x4472 - 0.09611*x4593 - 0.09261*x4696 - 0.09611*x5569 + 0.1549*x5594 - 0.2117*x5614 + 0.09122*x5781 + 0.1549*x5801 - 0.02419*x6150 + 0.2461*x6305 + 0.2117*x6376 + 0.3339*x6634 + 0.09121*x6962 + 0.09122*x7259 - 0.06366*x7449 + 0.09611*x7556 - 0.3029*x7558 - 0.1838*x7911 + 0.09121*x8112 - 0.09122*x8275 - 0.09122*x8286 + 0.3029*x8934 + 0.2117*x9843 + 0.2424*x10592 - 0.1873*x10886 - 0.1824*x11507 - 0.2785*x11546 - 0.4578*x12103 - 0.09121*x12116 + 0.09122*x12282 + 1.057

(b)

ŷ=0.1884*x247 + 0.05729*x402 - 1.753*x573 - 0.1884*x1396 + 0.3903*x2108 - 0.1731*x2801 - 0.05729*x2857 - 0.1731*x2880 + 0.1884*x3801 - 0.4572*x4171 + 0.1884*x4827 - 0.1884*x4866 + 0.1884*x5102 - 0.1884*x5455 - 0.1884*x5799 + 0.3903*x7493 - 0.1731*x8731 + 0.1884*x9533 + 0.1884*x9613 - 0.05729*x10035 + 0.1884*x11187 - 0.05729*x11775 - 0.2457*x12524 + 0.1884*x12531 + 0.1884*x13262 - 0.4572*x14234 - 0.1731*x14463 + 0.1731*x15359 + 0.05729*x15599 + 0.1884*x15808 + 0.1884*x17267 - 0.1884*x18009 - 0.1884*x18133 + 0.09123*x18527 + 0.09123*x18548 + 0.1884*x18595 - 0.1485*x19105 + 0.1884*x19994 + 0.1884*x20075 + 1.171*x21591 + 0.05729*x21747 + 7.027

(c)

FIGURE 5.1: Meilleurs modèles obtenus.

5.2.2.2 Evolution du Fitness au cours des générations

Au cours d'exécution, la valeur meilleure et moyenne de fitness se variée vis-à-vis le nombre des génération. Cette variation est présentée dans la figure (5.2).

On peut constater que les valeurs de RMSE diminuent avec l'augmentation du nombre de générations. La meilleure RMSE a été trouvée à la 199^{éme} génération (RMSE =0.2114 pour le cancer de la prostate, RMSE=0.15367 pour le cancer du poumon et RMSE=0.0027 pour le sarcome).

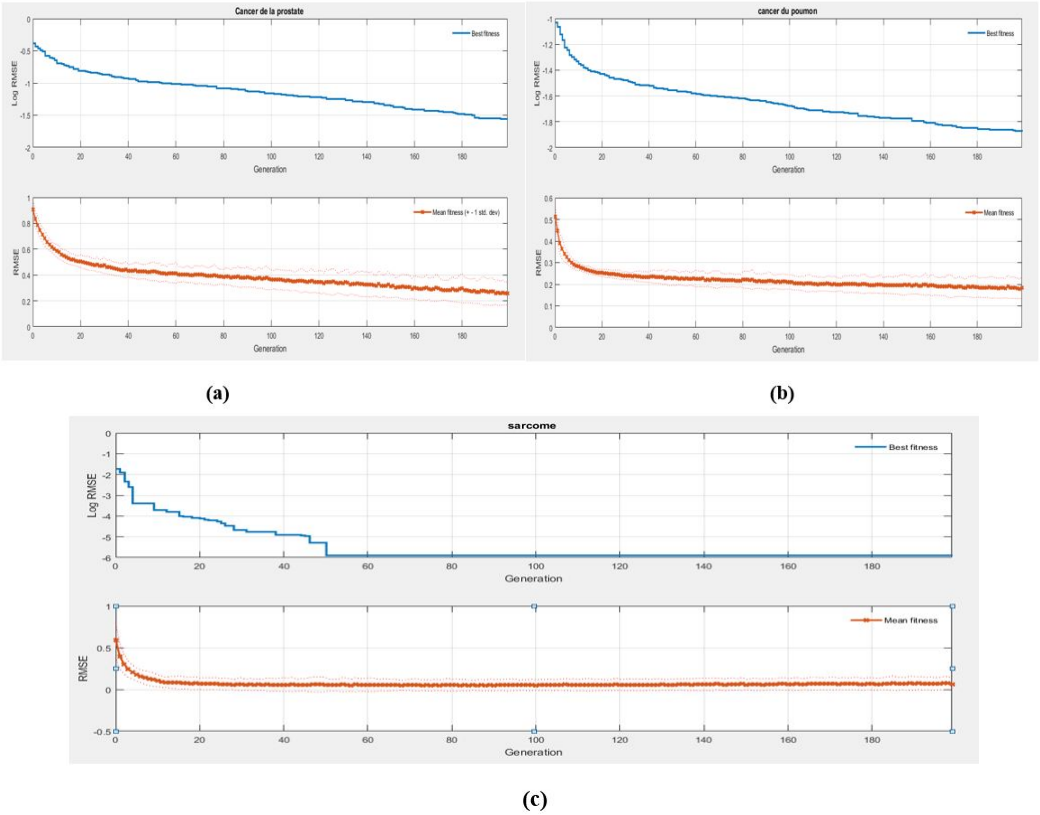


FIGURE 5.2: Variation du meilleure et moyenne RMSE avec le nombre de génération.

5.2.2.3 Qualité des modèles constitués

Les figures 5.3 présente la population des modèles évolués en termes de complexité et de valeur de tness(1-R^2). Les modèles optimaux qui fonctionnent relativement bien et sont beaucoup moins complexes(moins de nœuds) sont représentés dans la population comme des cercles vert,Le meilleur modèle de la population est représenté par un cercle rouge.

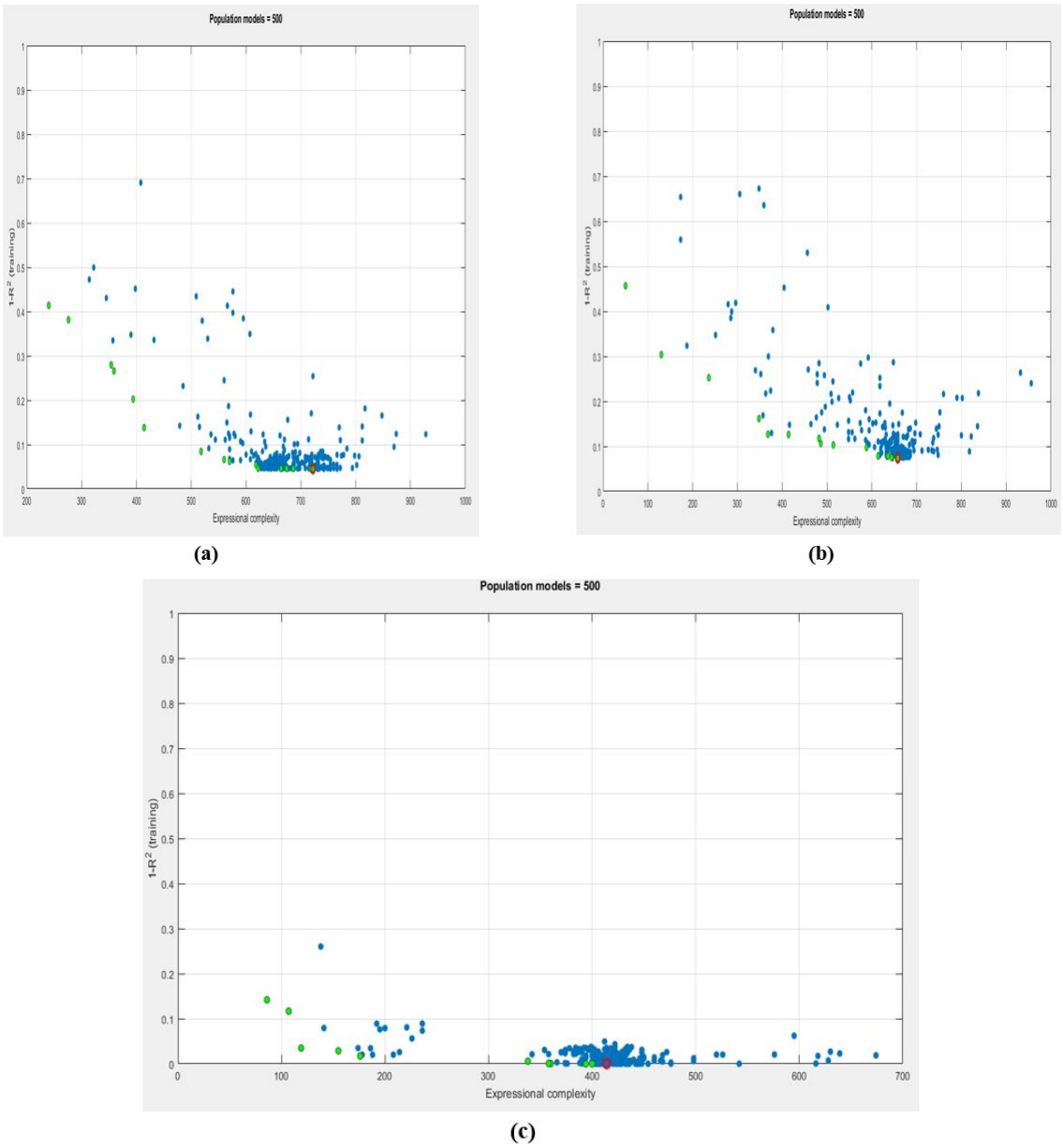
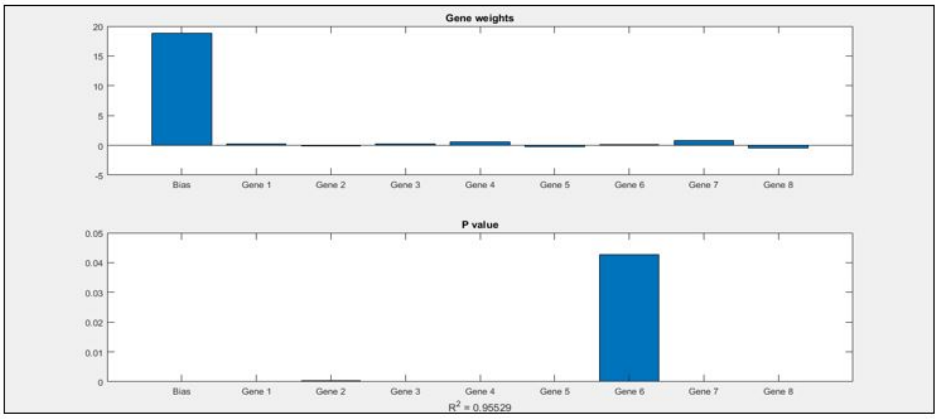


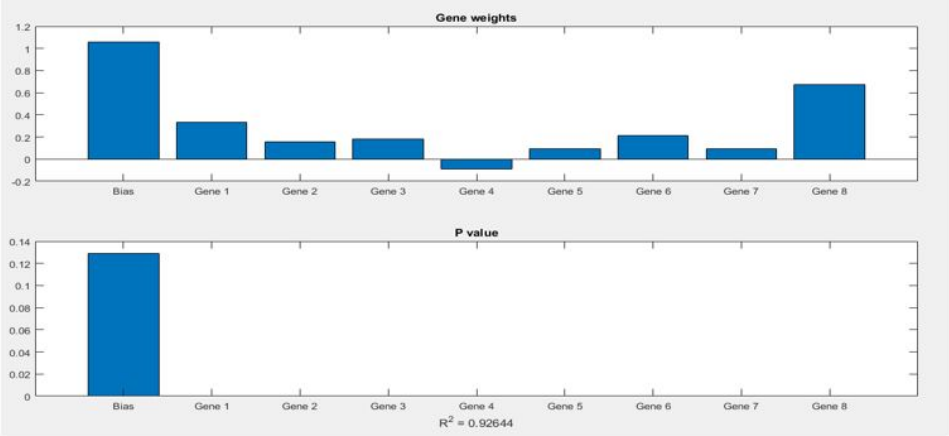
FIGURE 5.3: Population des modèles évolués en termes Pareto de complexité et de fitness.

5.2.2.4 Importances des gènes sélectionnés

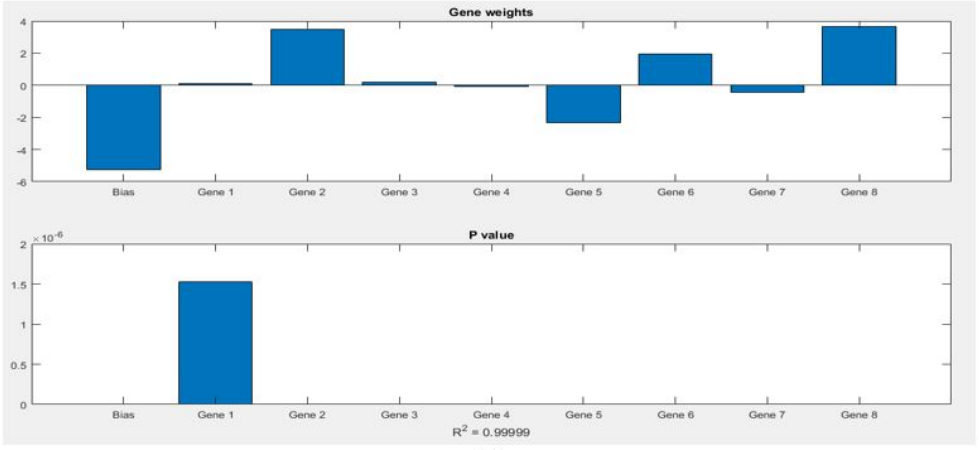
nous avons utilisé la mesure P-value pour identifier le degré de signification des gènes sélectionnés dans les modèles optimaux.Les gènes sont jugés plus significatifs si leurs valeurs de p sont mnimales (Figure 5.4).



(a)



(b)



(c)

FIGURE 5.4: Propriétés statistiques du meilleur modèle MGGP évolué.

5.2.2.5 Evaluation des modèles de prédiction

Les résultats présentés dans la table (4.2), montrent l'efficacité de notre méthode pour résoudre le problème de sélection des gènes, dont, pour chacun des modèles de prédiction, nous avons calculé les métriques d'évaluation suivants :R^2, RMSE, MSE, SSE, MAE et MAXE sur les données d'apprentissage et aussi sur les données du test.

	Mesures d'évaluation											
	R^2		RMSE		MSE		SSE		MAE		MAXE	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
(a)	0.9553	0.722	0.2114	0.5272	0.0447	0.2780	4.1111	2.7796	0.1639	0.4979	0.5225	0.7374
(b)	0.9264	0.7106	0.1537	0.2802	0.0236	0.0785	2.9591	6.4379	0.1133	0.2031	0.5563	1.2812
(c)	0.9999	0.8212	0.0028	0.3987	7.6113 10^{-6}	0.1590	1.2939 $10^{-6]}$	0.9538	0.0023	0.3247	0.0059	0.7007

TABLE 5.2: Valeurs des mesures d'évaluation obtenus

Nous remarquons que les valeurs des métriques d'évaluation montrées dans le tableau 5.2 sont très intéressantes pour les trois bases utilisé(apprentissage et test),car les valeurs R^2 sont proche de 1 et les valeurs de RMSE ,MSE et MAE proche de 0 et les valeurs de SSE(somme d'erreurs au carré) dans le pire des cas 6.4379 (la base du poumon (test) et dans le meilleur des cas 1.2939 * 10^{-6} (la base de sarcome (apprentissage)). le même chose, nous avons obtenu une bonne corrélation entre les classes réelles et prédites comme il est montré la figure suivante :

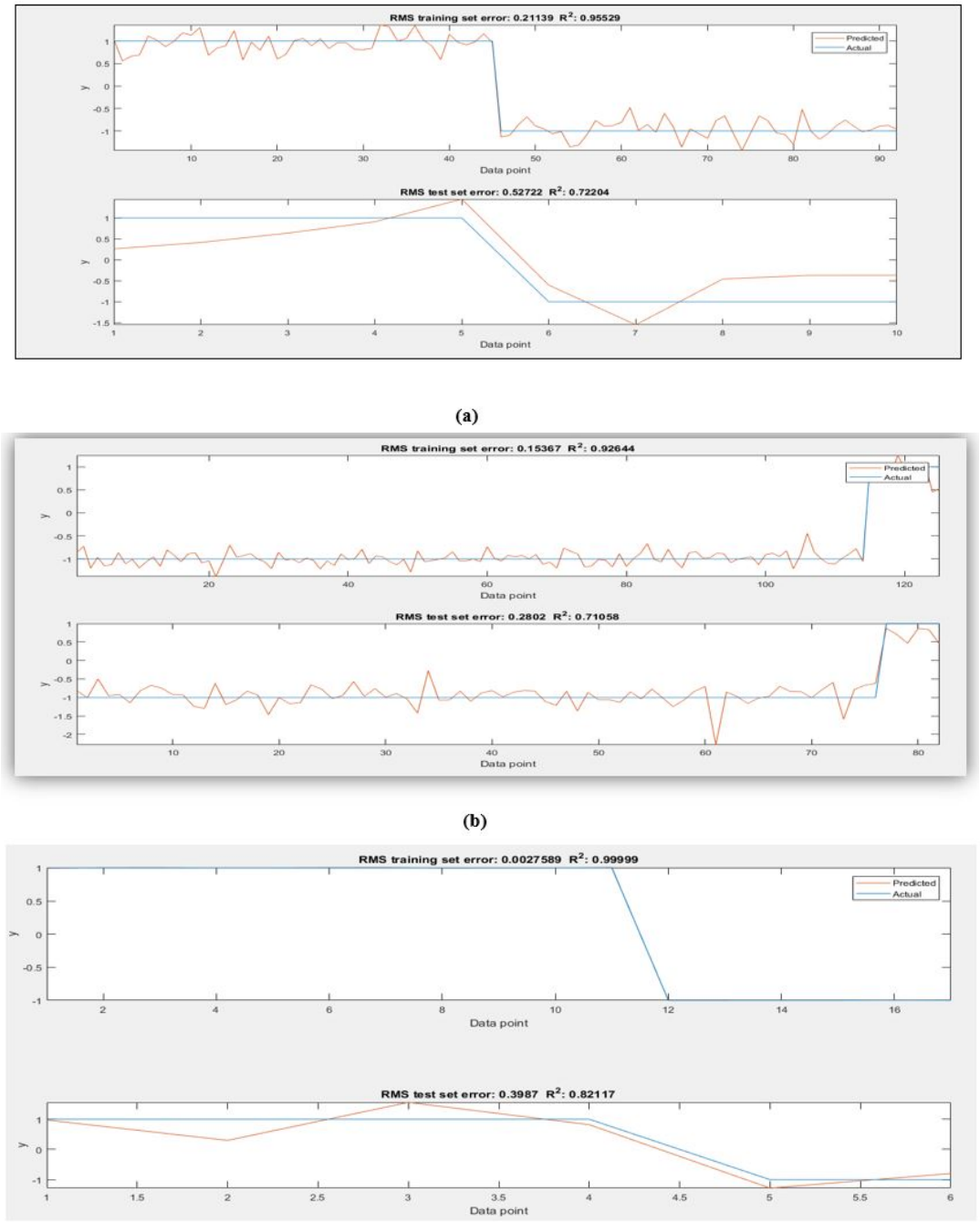


FIGURE 5.5: Corrélation entre les valeurs prédites ($\hat{y}$) et réelles (y).

5.2.2.6 Gènes sélectionnées

La table (5.2) représente les noms des gènes sélectionnés et utilisés par les Modèles MGGP développés pour diagnostiquer les maladies de cancérologie, de sorte que le nombre des gènes sélectionnés était 49 gènes pour le cancer de la prostate,50 gènes pour le cancer du poumon et 41 gènes pour le sarcome.

	Noms des gènes sélectionnés
Cancer de la prostate (49 gènes sélectionnés)	'768_at'- '40349_at' - '36569_at' - '38665_at' - '39845_at' - '40993_r_at' - '40025_at' - '37093_at' -'36802_at' - '2073_s_at' - '36629_at' - '39430_at' - '38976_at' - '725_i_at' -'672_at' - '35278_at' - '340_at' - '39423_f_at' - '38876_at' - '38090_at'- '36686_at' -'36043_at' - '1260_s_at' - '39720_g_at' -'769_s_at' - '32545_r_at' - '40536_f_at' - '32436_at' - '35429_at' -'33881_at' - '36893_at' -'41252_s_at' - '40336_at' - '38287_at' - '39840_at' - '39562_at' - '34185_at' - '32598_at'- '35048_at' - '41376_i_at' - '38764_at' -'34137_at' - '34551_at' - '40233_at' - '37639_at' - '39557_at' - '40433_at' - '876_at'
Cancer du poumon (50 gènes sélectionnés)	'33176_at'- '37604_at'- '39045_at'- '36535_at'-'619_s_at'- '37997_r_at'- '41776_at' -'41695_at'-'32947_at'- '35730_at' - '36447_at' - '33910_at'-'37464_at'- '32082_at' - '37132_at'-'37689_s_at'-'40092_at'-'40387_at'- '1857_at'- '1175_s_at'-'41087_at'- '35019_at'- '34829_at'-'585_at' -'41464_at'-'37094_at'- '32242_at'- '31859_at'- '36734_at'- '35705_at'- '38308_g_at'-'31329_at'-'33966_at'-'40275_at'- '35990_at'- '31506_s_at'- '38856_at'-'32244_at'- '34003_at'- '35042_at'- '444_g_at'-'38508_s_at' - '41382_at'- '34640_at'- '35363_at'- '1224_at'-'36515_at'- '40868_at' -'31513_at'-'35352_at'
Sarcome (41 gènes Sélectionnés)	'218622_at'- '219201_s_at'- '218644_at'-'221690_s_at'- '215445_x_at'-'208696_at '- '202811_at' -'202732_at'- '214544_s_at'- '209500_x_at'- '205035_at'- '218691_s_at' -'204760_s_at'-'211207_s_at'-'218104_at'-'48659_at'- '205388_at'- '218228_s_at'- '212599_at'- '34206_at'-203734_at'-'220090_at'-'201327_s_at'-'212606_at'- '217362_x_at'- '209583_s_at'- '205732_s_at'-'204799_at'-'215895_x_at'- '220171_x_at'-'213341_at'-'202788_at'- '215686_x_at'-'221847_at'- '40359_at'- '210010_s_at'-'211839_s_at'-'207431_s_at'- '202039_at'- '214315_x_at'- '204104_at'

TABLE 5.3: Noms des gènes sélectionnés

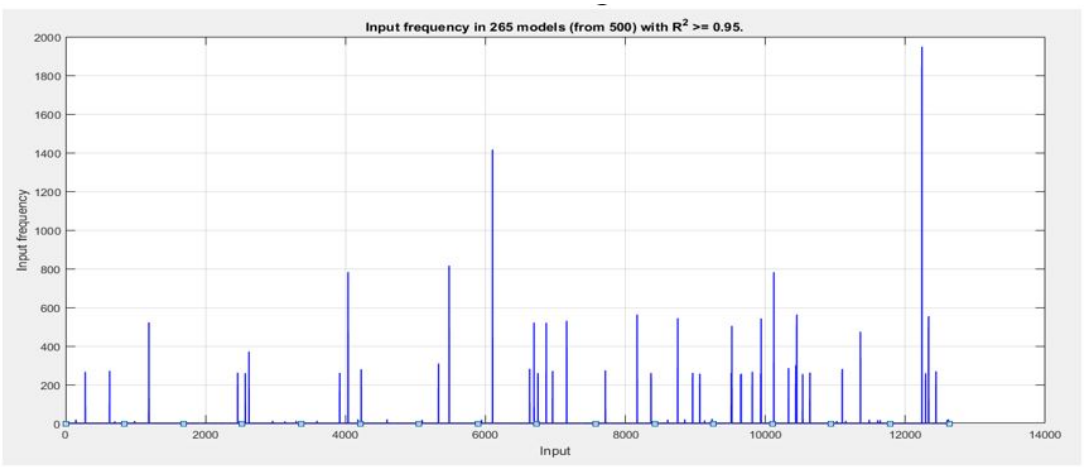
5.2.2.7 L'influence des gènes sélectionnées

La figure (5.6) montre la fréquence des gènes dans les meilleurs modèles dans les populations. Les fréquences montrés dans la figure correspondent au degré ou à la mesure dans laquelle les variables prédictives (x_i) influencent les variables de réponse (y). Plus la fréquence d'une variable prédictive est grande, plus l'influence est grande (c'est-à-dire la variable la plus importante).D'aprés ces résultat nous constatons que :

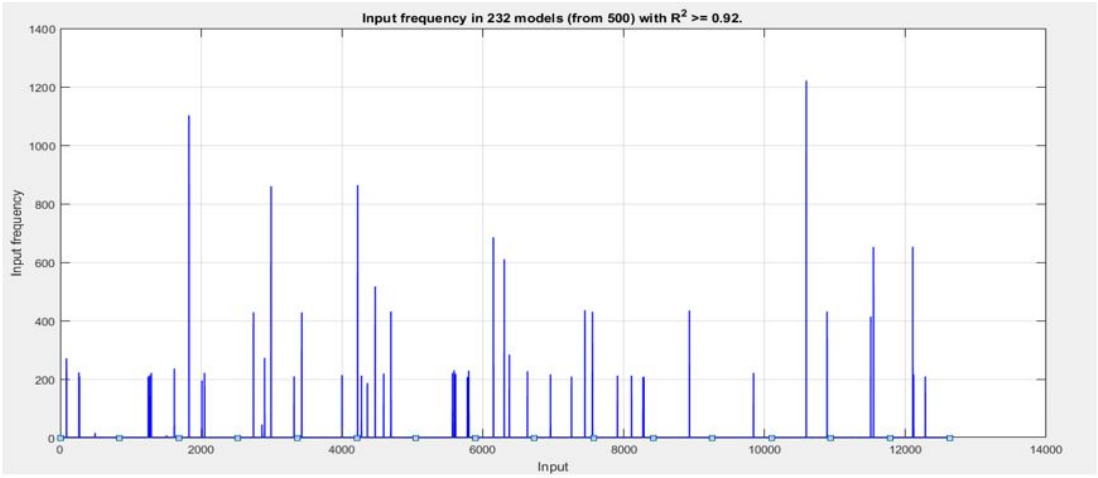
Sur les 49 attributs de la base de cancer de la prostate utilisés par MGGP,l'attribut '36043_at' est le facteur le plus influant sur la variable de réponse (y) suivie par '35429_at' et '340_at'.

Sur les 50 attributs de la base de cancer du poumon utilisés par MGGP,l'attribut '32947_at' est le facteur le plus influant sur la variable de réponse (y) suivie par '41087_at' et - '37464_at'.

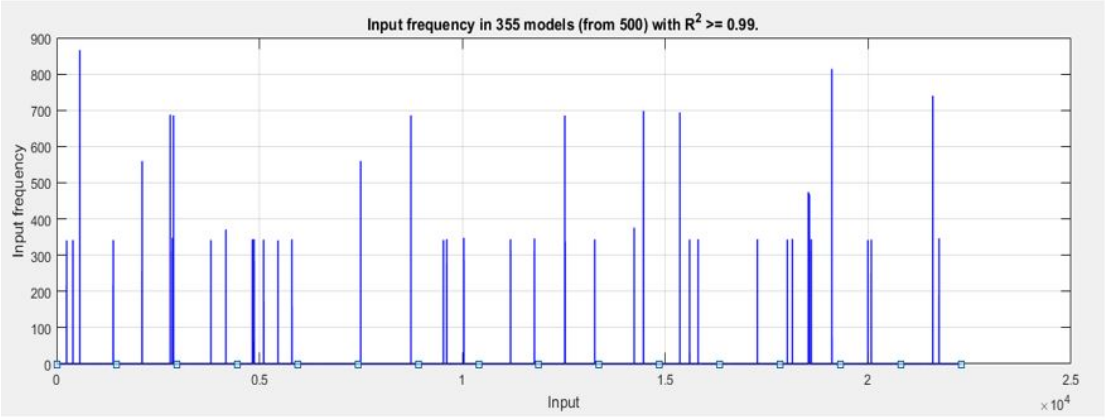
Sur les 41 attributs de la base de sarcome utilisés par MGGP,l'attribut '48659_at' est le facteur le plus influant sur la variable de réponse (y) suivie par '221690_s_at' et '202732_at'.



(a)



(b)



(c)

FIGURE 5.6: Fréquences des gènes dans les meilleurs modèles.

5.2.3 Evaluation des résultats

Pour garantir la qualité des sous ensembles trouvés par MGGP,nous avons fait une classification par trois methodes svm, regression logistique et l'arbre de decision en utilisant l'outil weka pour valider notre resultats.la table 5.4 presente le résultat de la classification.et les sous ensembles trouvés sont valides si nous changeons de méthode d'induction,

	SVM			regression logistique			arbre de decision		
	i	ii	iii	i	ii	iii	i	ii	iii
(a)	94.1176%	97.0297%	67.0211%	93.1373%	97.0297%	59.4059%	83.3333%	89.1089%	55.4455%
(b)	98.032%	99.0291%	91.254%	89.8551%	99.5146%	90.7767%	97.1014%	96.1165%	91.7476%
(c)	100%	100%	86.3636%	95.4545%	95.4545%	86.3636%	78.2609%	90.90%	86.3636%

TABLE 5.4: Taux de classification par svm,regression logistique,arbre de decision utilisant : i= gènes avant la sélèction , ii=gènes selectionnés , iii= meilleur 3 gènes frequents.

5.2.4 Conclusion

Les résultats obtenus dans les différentes bases montrent l'efficacité des paramètres de contrôle de MGGP que nous avons déterminé par expérimentations.

Dans le cadre de la sélection des gènes,L'erreur quadratique moyenne RMSE, R^2, les diagrammes de dispersion et les histogrammes montrent clairement l'efficacité des modeles déterminés par MGGP.

Pour cela, nous pouvons conclure que le modèle MGGP permettant la réduction du nombre d'attributs et la sélection du sous-ensemble optimal des gènes ce qui conduira à diagnostiquer les maladies de cancérologie efficacement et surmonter le problème de sur-apprentissage des données.

La même analyse peut être étendue pour prendre en charge d'autres maladies de cancer tel que cancer de sein, cancer du cerveau, cancer du colon . . . etc.

CONCLUSION GÉNÉRALE

Ce mémoire s'articule autour d'un problème dans le domaine du bio-informatique et la reconnaissance et traitement des maladies celui de la sélection des gènes ou la réduction du dimension des données de la technique du biopuce, ce problème constitue une plateforme essentielle pour plusieurs tâches complexes de la bio-informatique.

Pour remédier ce problème, Nous avons proposé une méthode basée sur des métaheuristiques d'optimisation et plus particulièrement sur la méthode MGGP. Dans cette méthode, un individu désigné comme une structure d'arbres (gènes) reçoit ensembles des variables entrée et tente de prédire une variable de sortie Y et caractérisé par les paramètres de contrôle G_{max} et D_{max}. Ses deux paramètres jouent un rôle important dans la détermination du nombre des gènes sélectionnés.

Pour valider notre proposition et pour tester la possibilité de l'utilisation la méthode MGGP et ses propriétés de réduire les attributs(gènes) et trouver le sous-ensemble optimal, nous avons utilisé trois bases de plus de 12600 gènes. La qualité des modèles et l'influence des gènes sélectionnés sur la qualité des modèles est déterminé par le RMSE et R^2.

Pour les perspectives et les travaux futur, nous proposons des idées qui peuvent améliorer et généraliser notre système de sélection des gènes telles que :

- L'utilisation d'autres types de prétraitement des données pour adapter les données au type d'analyse souhaité.
- généraliser notre système pour prendre en considération :
  - le traitement des données classées en multi-classes(multi-label).
  - l'utilisation d'autres bases des données concernent d'autres problèmes.
- faire une étude sur les paramètre de contrôle de MGGP pour améliorer les performances du système.

BIBLIOGRAPHIE

[1] Bioinformatics toolbox. https://www.mathworks.com/help/bioinfo/index.html. Accessed : 2018-06-01.

[2] F Bertucci. Profils d'expression génique et puces à adn dans le cancer du sein : choix du patient, choix du protocole. In *Cancer du sein*, pages 267–276. Springer, 2006.

[3] Hassan CHOUAIB. *Sélection de caractéristiques : méthodes et applications*. PhD thesis, Université Paris Descartes, 2011.

[4] Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and TAMT Meyarivan. A fast and elitist multiobjective genetic algorithm : Nsga-ii. *IEEE transactions on evolutionary computation*, 6(2) :182–197, 2002.

[5] Ramón Díaz-Uriarte and Sara Alvarez De Andres. Gene selection and classification of microarray data using random forest. *BMC bioinformatics*, 7(1) :3, 2006.

[6] Ramón Díaz-Uriarte and Sara Alvarez De Andres. Gene selection and classification of microarray data using random forest. *BMC bioinformatics*, 7(1) :3, 2006.

[7] Terrence S Furey, Nello Cristianini, Nigel Duffy, David W Bednarski, Michel Schummer, and David Haussler. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10) :906–914, 2000.

[8] Todd R Golub, Donna K Slonim, Pablo Tamayo, Christine Huard, Michelle Gaasenbeek, Jill P Mesirov, Hilary Coller, Mignon L Loh, James R Downing, Mark A Caligiuri, et al. Molecular classification of cancer : class discovery and class prediction by gene expression monitoring. *science*, 286(5439) :531–537, 1999.

[9] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1) :389–422, Jan 2002.

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

[11] 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.

[12] 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.

[13] Edmundo Bonilla Huerta, Béatrice Duval, and Jin-Kao Hao. A hybrid ga/svm approach for gene selection and classification of microarray data. In *Workshops on Applications of Evolutionary Computation*, pages 34–44. Springer, 2006.

[14] Indu Jain, Vinod Kumar Jain, and Renu Jain. Correlation feature selection based improved-binary particle swarm optimization for gene selection and cancer classification. *Applied Soft Computing*, 62 :203–215, 2018.

[15] Mariam Kalakech. *Sélection semi-supervisé d'attributs : Application à la classification de textures couleur*. PhD thesis, 2011.

[16] MENGHOUR Kamilia. *Approches Bio-inspirées pour la Sélection d'Attributs*. PhD thesis, Université Badji Mokhtar-Annaba, 1945.

[17] Mohammed Khabzaoui. *Modélisation et résolution multi-objectifs des règles d'association : application à l'analyse de données biopuces*. PhD thesis, Lille 1, 2006.

[18] P LEE and TJ HUDSON. La puce à adn en médecine et en science. *MS. Médecine sciences*, 16(1) :43–49, 2000.

[19] Gaëlle Legrand. *Approche méthodologique de sélection et construction de variables pour l'amélioration du processus d'extraction des connaissances à partir de grandes bases de données*. PhD thesis, Lyon 2, 2004.

[20] Leping Li, Clarice R Weinberg, Thomas A Darden, and Lee G Pedersen. Gene selection for sample classification based on gene expression data : study of sensitivity to choice of parameters of the ga/knn method. *Bioinformatics*, 17(12) :1131–1142, 2001.

[21] Nicolas Mattiuzzo. *Etude du transcriptome kératinocytaire au cours du programme de différenciation de l'épiderme*. PhD thesis, Université de Toulouse, Université Toulouse III-Paul Sabatier, 2009.

[22] M. Tomislav Meštrović. Application de puce adn, 2017.

[23] PRADYUT KUMAR Muduli. *Evaluation of liquefaction potential of soil using genetic programming*. PhD thesis, 2013.

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

[25] Maude Pupin. La génomique. , CRISTAL, Université de Lille, 1994.

[26] Philippe Reymond. Introduction à la génomique fonctionnelle. , Département de Biologie moléculaire et végétale, Université de Lausanne, 2007.

[27] Dominic P Searson. Gptips 2 : an open-source software platform for symbolic data mining. In *Handbook of genetic programming applications*, pages 551–573. Springer, 2015.

[28] Nguyen Hoai Tuong. Puces à adn. *Ecole Polytechnique, Université de Lille*, 1 :14, 2008.

[29] Jérome Cadieux Yassine Ariba. Manuel matlab. , Icam de Toulouse.

[30] Xin Zhou and KZ Mao. Ls bound based gene selection for dna microarray data. *Bioinformatics*, 21(8) :1559–1564, 2004.

[31] Xin Zhou and David P Tuck. Msvm-rfe : extensions of svm-rfe for multiclass gene selection on dna microarray data. *Bioinformatics*, 23(9) :1106–1114, 2007.