**REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE**

**MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET LA RECHERCHE SCIENTIFIQUE**

**UNIVERSITE LARBI BEN M'HIDI-OUM EL BOUAGHI**

**FACULTE DES SCIENCES EXACTES ET DES SCIENCES DE LA NATURE ET DE LA VIE**

**Département de Mathématique et Informatique**

**N°D'orde :……..**

**Série :…….…..**

***MEMOIRE***
***EN VUE DE L'OBTENTION DU DIPLOME DE MAGISTER EN INFORMATIQUE***
***OPTION : Intelligence Artificielle et Imagerie***

***Thème***

# *CLASSIFICATION ET PREDICTION EN BIOINFORMATIQUE :*

*La découverte des biomarqueurs pour le diagnostic du cancer*

**Présenté par :** **Encadré par :**

**Mme SID Karima** **Pr. BATOUCHE Mohamed**

**Soutenu le : 18/06/2014**

**Devant le jury composé de :**

| | | |
|---|---|---|
| **Pr. BENMOHAMMED Mohamed** | **Président** | **Université de Constantine2** |
| **Pr. BATOUCHE Mohamed** | **Rapporteur** | **Université de Constantine2** |
| **Dr. BOUTEKKOUK Fateh** | **Examinateur** | **Université d'Oum El Bouaghi** |
| **Dr. NINI Brahim** | **Examinateur** | **Université d'Oum El Bouaghi** |

**Année universitaire 2013/2014**

# *Remerciement*

*Je remercie « ALLAH » le tout puissant, qui ma donnée la foi, la force et la patience pour aller jusqu'au bout de ce travail*

*Ma reconnaissance va à tous ceux qui m'ont aidé à conduire ce travail à son terme : tout particulièrement, j'aimerais remercier vivement, mon encadreur de thèse, le professeur* ***BATOUCHE Mohamed****,*
*de l'attention et du soutien qu'il a porté à mon travail.*

*Ainsi je tiens à remercier vivement le professeur*
***BENMOUHAMMED Mohamed*** *d'avoir accepté de présider ce jury*
*Je remercie également Dr* ***NINI Ibrahim*** *et Dr* ***BOUTEKKOUK Fateh*** *qui ont participé à examiner ce travail.*

*Finalement, je remercie toutes les personnes qui, directement ou indirectement ont contribué à la réalisation de ce travail*

# *Résumé :*

La découverte de biomarqueurs est l'un des domaines de recherche en bioinformatique. Quelque soit le type de biomarqueurs génomiques, transcriptomiques, protéomiques ou métabolomiques, le défi principal consiste à développer une méthode robuste et performante pour découvrir ces biomarqueurs à partir d'un grand ensemble de données qui peut contenir des données non pertinentes et redondantes. Pour ce faire, les deux techniques, la sélection de caractéristiques et l'apprentissage supervisé (classification) sont utilisées. Les méthodes existantes présentent des faiblesses au niveau de leur complexité très élevée, l'indépendance au classificateur ainsi que l'instabilité. Dans le but de limiter ces inconvénients, nous proposons, dans ce travail, une nouvelle approche stable pour la découverte de biomarqueurs, l'approche proposée est composée de trois étapes : l'étape de clustering basée sur l'approximation d'une couverture de Markov, l'étape de filtrage et l'étape d'optimisation utilisant trois algorithmes d'optimisation les algorithmes génétiques, l'optimisation par l'essaim particulaire et l'algorithme de la sélection clonale en coopération. Les expérimentations ont montré que notre approche est efficace et qu'elle a la capacité de sélectionner un nombre réduit de gènes tout en conservant des taux d'erreur de classification très faible est une stabilité très satisfaisante. Les performances de l'approche proposée sont mises en évidence à travers une comparaison avec d'autres méthodes de la littérature du domaine.



***Mots clefs:*** *bioinformatique, biomarqueur, sélection de caractéristiques, classification, approximation d'une couverture de Markov, les algorithmes d'optimisation, optimisation multiobjectifs.*

# *Abstract:*

Biomarker discovery is one of the research areas in bioinformatics. Whatever type of biomarkers, genomic, transcriptomic, proteomic, or metabolomic, the main challenge is to develop a robust and fast method to discover these biomarkers from a large dataset that may contain irrelevant and redundant data. For this reason, the two techniques, the feature selection and supervised learning (classification) are used. Existing methods have weaknesses in their high complexity, independence to classifier and instability. In order to limit these disadvantages, we propose in this work a new stable approach to biomarker discovery, the proposed approach is consisted of three steps: clustering step based on Markov Blanket approximation, filtering step and optimization using three optimization algorithms genetic algorithms, particle swarm optimization and the clonal selection algorithm in cooperation. Experiments have shown that our approach is effective and has the ability to select a small number of genes while maintaining very low error rate classification with very satisfactory stability. The performances of the proposed approach are highlighted through a comparison with other methods in the literature in the field.



***Keywords*:** *bioinformatic, biomarker, feature selection, classification, Markov Blanket optimization, optimization algorithms, multiobjectives optimization.*

# ملخص

اكتشاف العلامات البيولوجية هي واحدة من مجالات البحث في المعلوماتية الحيوية. اي نوع من المؤشرات الحيوية جينية ، تروسكريبتومية ، بروتنية او ايضية ، فان التحدي الرئيسي هو تطوير طريقة قوية وسريعة لاكتشاف هذه المؤشرات الحيوية من مجموعة كبيرة من البيانات التي قد تحتوي على بيانات ليس لها صلة وزائدة عن الحاجة. للقيام بذلك ، تم استخدام التقنيتين : اختيار الميزة و تقنيات التعلم (تصنيف). الطرق الموجودة لديها نقاط ضعف في تعقيداتها العالية ، استقلاليتها عن المصنف وعدم الاستقرار .من اجل الحد من هذه العيوب ، نقترح في هذا العمل ، طريقة جديدة لاكتشاف العلامات البيولوجية ، تتالف الطريقة المقترحة من ثلاث خطوات :خطوة التجميع على اساس التقريب لغطاء ماركوف ، خطوة التصفية وخطوة التحسين الامثل باستخدام ثلاثة خوارزميات: الخوارزميات الجينية ، تحسين سرب الجسيمات والاختيار النسيلي. وقد اظهرت التجارب ان طريقتنا فعالة ولديها القدرة على اختيار عدد صغير من الجينات مع الحفاظ على نسبة الخطا في التصنيف منخفضة جدا ونسبة استقرار مرضية للغاية. ويتم إثبات فعالية الطريقة المقترحة من خلال المقارنة مع الطرق الاخرى الموجودة في هذا المجال.

***الكلمات المفتاحية :***المعلوماتية الحيوية، العلامات البيولوجية ، اختيار الميزة ، التصنيف ، التقريب لغطاء لماركوف ، خوارزميات التحسين ، التحسين المتعدد الاهداف.

# *Table des matières*

# *Liste des figures*

# *Liste des tableaux*

# *Introduction générale*

## 1. Introduction

Ces dernières années, la recherche en biologie et tout particulièrement en génétique a connu un formidable essor et continue sur sa lancée. Les avancées réalisées sont considérables. Cependant si à l'époque de Gregor Mendel, il suffisait de faire certains croisements entre divers espèces de pois comestibles pour faire de grandes avancées dans le domaine de la génétique, de nos jours les chercheurs utilisent d'autres techniques qui fournissent une très importante somme d'informations que nous ne pouvons traiter sans l'aide de l'informatique. A tel point, qu'il y a un peu plus d'une dizaine d'années, une nouvelle discipline a été créée nommée *la bioinformatique*, située au carrefour de *la biologie moléculaire*, des mathématiques, des statistiques et de l'informatique.

L'objectif principal de la bioinformatique consiste à traiter des données biologiques à l'aide des outils informatiques afin d'extraire une nouvelle bioinformation qui peut être utilisée par les spécialistes du domaine pour prendre certaines décisions. Les sources fondamentales de ces données sont des technologies à haut débit de la biologie moléculaire comme les techniques à puce (biopuces), la technique d'électrophorèse, les PCR…etc. En effet, il est aujourd'hui démontré que les données produites par ces technologies et qui sont connues sous le nom des données *omiques* seront plus importantes que tout ce qui a été produit dans le passé, notamment dans le domaine médical lors d'examen des patients. Ces données peuvent être utilisées comme support de décision médicale, dans la littérature, nous trouvons régulièrement la notion *d'aide au diagnostic*. La quantité des données obtenues par ces technologies est très grande (la haute dimensionnalité) avec un petit nombre des échantillons pour des raisons de coût, cette situation est de plus en plus fréquente avec la technique des puces à ADN. Les puces à ADN sont des techniques qui reposent sur le principe de complémentarité des brins de la double hélice d'ADN, et elles permettent de mesurer simultanément sur une seule puce l'expression de centaines, voire des milliers jusqu'à dizaines de milliers de gènes transcrits. Une des tâches de la bioinformatique consiste à analyser cette grande quantité de données (niveaux d'expression des gènes) pour identifier un petit sous ensemble de gènes exprimés différentiellement, qui représentent *le bon*

*biomarqueur* d'une certaine maladie comme le cancer et qui est utilisé immédiatement comme un moyen de diagnostic.

## 2. Motivation et objectif du travail :

Le biomarqueur est un nouveau concept connu dans la recherche biomédicale, il désigne une signature biologique qui indique la présence ou l'absence d'une maladie, les effets thérapeutiques, le type d'une maladie, leur stade…etc. Ce concept est très fréquent dans la cancérologie où il joue un rôle très important dans le diagnostic. L'identification ou bien la découverte de biomarqueurs à partir d'une grande dimension de données d'une expérience de puce à ADN (les niveaux d'expression des milliers de gènes), où le nombre des échantillons est petit repose sur les techniques issues de l'intelligence artificielle notamment les technique d'apprentissage automatique. Au point de vue d'apprentissage automatique, la découverte de biomarqueurs peut être définie comme un problème de sélection des caractéristiques pour une tâche de classification, dont le but est de trouver un petit ensemble de gènes(biomarqueurs) qui explique le mieux la différence entre les échantillons malades et les échantillons témoins. La difficulté principale est la masse de données disponibles qui rend le processus de découverte très complexe, une étape de sélection de caractéristiques constitue alors un module important intégré à ce processus complexe, l'objectif de cette sélection consiste à réduire la dimensionnalité. Cette réduction rend, d'une part, beaucoup plus facile la gestion des données et d'autre part, aide à mieux comprendre les résultats fournis par un système basé sur ces caractéristiques. Les approches de sélection des caractéristiques sont généralement de trois types : filter, wrapper et embedded. Les approches filter utilisent des mesures statistiques comme l'information mutuelle, le coefficient de corrélation …etc., calculées sur les caractéristiques afin de filtrer les plus informatives. Cette étape est généralement réalisée avant d'appliquer un algorithme de classification. Ces méthodes de filtrage présentent des avantages au niveau de leur efficacité calculatoire, mais elles ne tiennent pas en compte les interactions entre les caractéristiques, ce qui conduit à sélectionner des caractéristiques comportant des informations redondantes. De plus, ces méthodes ne tiennent absolument pas compte des résultats d'une méthode de classification. Dans les approches wrapper la sélection est toujours liée à la classification ce qui permet d'obtenir des résultats plus performants. La plupart des approches wrapper sont définies comme des hybrides d'un algorithme d'optimisation tels que les algorithmes génétiques, l'optimisation par l'essaim particulaire, l'algorithme de la sélection clonale…etc. et un algorithme de classification (GA/SVM,

GA/KNN, PSO/SVM…etc.), mais l'inconvénient majeur dans ce type d'approche est sa complexité très élevée. Pour les approches embedded, la sélection est liée à la classification mais avec une complexité réduite comme la méthode SVM-RFE basée sur l'élimination récursive des caractéristiques qui ont les poids les plus faibles, ces poids sont liés au système de classification SVM.

Quel que soit l'approche de sélection utilisée pour développer un algorithme de sélection, le critère de *la robustesse* ou de *la stabilité* est très important c'est-à-dire le résultat de notre algorithme de sélection doit être une signature robuste/stable et performante. Ces dernières années, la stabilité est devenue un sujet d'intérêt dans les recherches biomédicales, la bonne signature biologique ne doit pas être particulière aux données disponibles et elle doit être exportable aux autres jeux de données traitant le même problème de classification. La stabilité d'un algorithme de découverte de biomarqueurs est importante pour les chercheurs qui veulent valider encore leurs conclusions en appliquant les biomarqueurs découverts sur de nouvelles données. Dans la littérature, deux techniques sont proposées pour améliorer la stabilité d'un algorithme de sélection, la première basée sur le clustering où elle consiste à grouper les caractéristiques selon leur pertinence, ensuite une sélection est faite sur ces groupes et enfin les résultats de sélection sont agrégés, la deuxième basée ensemble inspirée de l'ensemble de classificateurs, elle a le même principe que la première mais à la place du groupement nous trouvons un sous échantillonnage aléatoire de la base initiale.

Dans ce travail, notre objectif consiste à développer une nouvelle approche stable pour la découverte de biomarqueurs qui exploite efficacement les avantages des approches de sélection Filter, wrapper et embedded, la technique basée groupe pour améliorer la stabilité, l'aspect multiobjectif fourni par les algorithmes d'optimisation et tend à optimiser les deux aspects suivants :

→ La stabilité, fournie une signature biologique robuste,

→ Les performances, fournie une signature performante (une signature de petite taille et qui minimise le taux d'erreur d'une classification).

## 3. Organisation du mémoire :

Outre la partie introductive et la conclusion générale, le travail est organisé en six chapitres. Le premier chapitre est consacré à la biologie moléculaire qui est la base de la bioinformatique. Dans la première section de ce chapitre, nous expliquons quelques notions et concepts liés à la biologie moléculaire afin de comprendre tous ce qui suit dans notre travail,

dans la deuxième section de ce chapitre nous détaillons le principe de quelques techniques les plus connues de cette discipline et qui représentent la source d'une grande quantité de données biologiques que la bioinformatique vise à analyser et interpréter.

Le deuxième chapitre est dédié à la bioinformatique. Nous présentons dans ce chapitre un état de l'art en matière de cette discipline, nous commençons par l'historique et définitions, le stockage de la bioinformation pour terminer avec les défis majeurs du domaine.

Dans le troisième chapitre, nous nous focalisons sur l'un des domaines de recherche en bioinformatique nommé la découverte de biomarqueurs qui est à la base du diagnostic des maladies. Nous présentons dans les premières sections de ce chapitre le concept de biomarqueur, sa définition et ses différents types, ensuite, nous expliquons la technique des puces à ADN comme une technique de découverte dans la cancérologie pour clôturer par une explication en détail du processus de découverte à partir des données d'expression génique.

Le quatrième chapitre est consacré aux techniques d'apprentissage automatique utilisées pour découvrir les biomarqueurs. Dans ce chapitre, nous montrons la relation entre la sélection de caractéristique et la classification pour identifier les biomarqueurs. Ce chapitre est organisé en deux sections principales, la première concerne la technique de la sélection de caractéristiques et la deuxième pour la technique de classification.

Le cinquième chapitre présente les algorithmes d'optimisation pour la sélection de caractéristique, nous détaillons dans ce chapitre trois algorithmes d'optimisation, les algorithmes génétiques, l'optimisation par l'essaim particulaire et l'algorithme de la sélection clonale et nous expliquons le processus de sélection par ces algorithmes.

Le dernier chapitre introduit une nouvelle approche de sélection que nous proposons, nous présentons en détails cette approche, le processus général de sélection et ses différentes étapes. Nous validons l'approche proposée par une phase d'expérimentation sur différentes bases et nous montrons l'efficacité de notre méthode pour sélectionner un nombre minimal de gènes avec un taux d'erreur réduit et une stabilité très satisfaisante. Finalement, nous terminons ce chapitre par une comparaison entre notre approche et d'autres approches représentatives de la littérature.

Enfin, la conclusion générale présente une synthèse des contributions apportées ainsi que les pistes définissant des perspectives possibles pour des travaux futurs.

## I.1. Introduction :

Tous les organismes vivants, des plus simples aux plus complexes, sont composés de cellules qui présentent des caractéristiques communes en termes de structure mais également de fonctionnement. Dans la cellule nous trouvons des molécules de petite taille et des autres qui ont une grande taille nommées les macromolécules. Trois types de macromolécules fondamentales sont impliqués dans cette unité cellulaire du vivant : les ADNs, les ARNs et les protéines. Schématiquement, la séquence des éléments (nucléotides ou les acides aminés) qui composent ces macromolécules constitue la représentation minimale permettant de décrire l'information qu'elles contiennent. La biologie moléculaire (parfois abrégée bio. mol.) c'est une discipline scientifique qui s'intéresse à étudier la cellule au niveau moléculaire c'est-à-dire étudier la structure de ces macromolécules, leur reproduction, leurs interactions, leur expression…etc. à l'aide d'un ensemble d'outils et de techniques. Ces techniques permettent de dupliquer (comme le clonage moléculaire), séparer (comme l'électrophorèse) et quantifier (comme les biopuces) ces éléments afin d'obtenir un dictionnaire descriptif de nos génomes. Dans ce chapitre nous présenterons la biologie moléculaire en trois sections principales. Dans ces sections nous nous focaliserons sur quelques notions liées à cette discipline. Ensuite, nous détaillerons le fonctionnement de quelques techniques les plus connues (connues comme les techniques de la biologie moléculaire).

## I.2. Définition :

La biologie moléculaire est une discipline scientifique ancienne. Elle a commencé avec l'évolution de la vie. Peu de temps après la cellule a été reconnue comme l'unité de base de la vie et sa structure a été déchiffrée (**la théorie cellulaire**), il est devenu clair que les constituants cellulaires ne sont que des composés chimiques. Bien que, certains de ces composés sont des substances chimiques simples comme les sucres et de nombreuses autres petites molécules, certaines autres molécules sont très grandes ayant une structure chimique complexe comme les acides nucléiques (ADN, ARN), des protéines, des glucides et des lipides complexes ; Ces molécules ont souvent une très grande masse moléculaire et sont relativement difficiles à synthétiser par voie chimique. Ceux-ci sont donc appelés des macromolécules. La biologie moléculaire s'intéresse principalement à étudier la cellule au niveau moléculaire c'est-à-dire la compréhension des interactions entre les différents systèmes moléculaire d'une cellule, y compris les interactions entre ces macromolécules (l'ADN, l'ARN et la biosynthèse des protéines) ainsi que l'apprentissage de la façon dont ces interactions sont

réglementées. Selon Michel Morange[1], la biologie moléculaire est « l'ensemble des techniques et découvertes qui ont permis l'analyse moléculaire des processus les plus intimes du vivant, de ceux qui en assurant la pérennité et la reproduction » **[101]**.

La naissance de la biologie moléculaire moderne liée à la découverte de l'ADN et leur structure en double hélice en 1953. Ensuite, cette science a connu un essor très rapide selon les avancées réalisées dans la découverte des molécules cellulaire.

La biologie moléculaire est principalement en croisement avec trois disciplines la biologie, la chimie et la physique, notamment à la génétique, la biochimie et la biophysique. Mais il ne faut pas confondre entre la génétique, la biochimie et la biologie moléculaire **[113]**.

- *Biochimie* c'est la discipline scientifique qui étudie les réactions chimiques ayant lieu dans les cycles métaboliques au sein de la cellule.
- *La génétique* est la science de l'hérédité qui étudie les caractères héréditaire et l'effet des différences génétiques sur les organismes.
- Tandis que, *la biologie moléculaire* quant à elle, s'intéresse à la structure, au fonctionnement, la composition des macromolécules cellulaire, la réplication, la transcription et la translation de matériel génétique (ADN).

Enfin, *la biophysique* c'est la discipline qui s'intéresse à étudier les phénomènes biologique et les structure des macromolécules en utilisant des théories et des techniques de la physique.

## I.3. Les notions biologiques :

Comme déjà vu, la biologie moléculaire a commencé après la théorie cellulaire et elle s'intéresse à étudier cette unité au niveau moléculaire et particulièrement les macromolécules d'ADN, ARN et protéine. Dans ce qui suit nous expliquerons ces notions et les processus biologiques appropriés au sein de la cellule.

### I.3.1. Le dogme central de la biologie moléculaire:

Le dogme central de la biologie moléculaire a été initialement introduit par Francis Crick en 1957. Le dogme indique que, une fois une séquence d'information est passée à la protéine ne peut pas être transférée à une acide nucléique ou une protéine. La description originale de Crick du dogme (Crick 1958) était **[22]**:

[1] Professeur de biologie à l'Ens et à l'Université Paris 6, Directeur du Centre Cavailles d'histoire et de philosophie des sciences de l'Ens. Il a maintenu une double activité, scientifique (biologie moléculaire et cellulaire, biologie du développement), historique et philosophique.

*Le dogme central* indique que lorsque «*l'information : l'information séquentielle*» est passée en protéines, elle ne peut pas sortir encore. De façon plus détaillée, le transfert d'informations à partir d'un acide nucléique à un acide nucléique, ou à partir d'un acide nucléique à la protéine peut être possible, mais le transfert de protéine à protéine, ou de protéine à un acide nucléique est impossible. L'information signifie ici la détermination précise de la séquence « information séquentielle », soit des bases dans l'acide nucléique ou des acides aminés dans la protéine **[22]**.

Le dogme central est très souvent confondu avec la voie standard de circulation de l'information de l'ADN à l'ARN à la protéine. Pour répondre à des malentendus sur le dogme, Crick en 1970 a expliqué trois catégories de transferts d'informations séquentielles: les transferts généraux (ceux qui se produisent couramment), les transferts spéciaux (peut se produire dans des situations particulières), et les transferts inconnus. Le dogme central est sur les transferts inconnus : protéine à protéine, protéine à ADN et protéine à ARN et il postule que ces transferts ne se produisent jamais **[22]**.

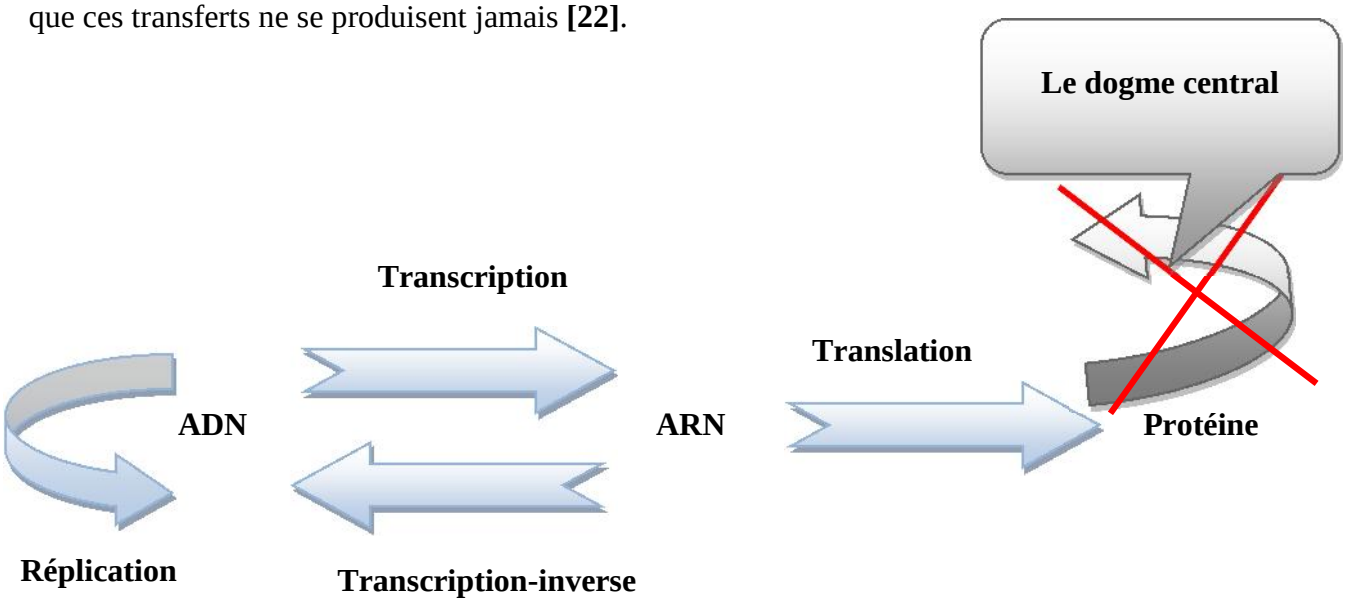



**Figure I.1 : Le flux de base d'informations séquentielle. Le dogme central de la biologie moléculaire : « une fois l'information (séquentielle) a passé en protéines, il ne peut pas sortir encore ».**

### I.3.2. Les acides nucléiques :

Les êtres vivants possèdent au sein de leurs cellules un **programme génétique** (donnant les caractéristiques de leur espèce et leurs caractéristiques individuelles), chez les **eucaryotes**, dans le noyau. Ce programme se transmet de génération en générations sous forme de **chromosomes**, supports des caractères héréditaires.

Compacté dans les chromosomes, notre ADN est comparable à un texte long de 3 milliards de lettres et écrit uniquement avec les lettres A, T, G et C. C'est ce qui constitue **notre identité biologique** et le **patrimoine** génétique que nous transmettons à nos enfants. Dans ce texte sont cachés les gènes nécessaires au développement, au bon fonctionnement et à la reproduction de nos cellules. Chez l'humain, il y a dans la cellule 23 paires de chromosomes où chaque paire de chromosome possède un long et a son propre fonctionnement.

### *I.3.2.1. L'acide désoxyribonucléique (ADN) :*

L'ADN ou l'acide désoxyribonucléique c'est l'acide nucléique qui encode l'information génétique. Il est constitué de deux chaînes de nucléotides (l'unité de base de l'ADN) mono phosphates liés chacun par une liaison ester entre son carbone 3' (alcool secondaire) et le carbone 5' (alcool primaire) du nucléotide suivant**.** Ces deux chaînes de nucléotides sont unies entre elles par des liaisons hydrogènes pour former un hybride en forme de double hélice (c'est le modèle de **Watson et Crick** en 1953)**.** Le nucléotide comporte les substances suivantes **[51]**:

- Un sucre (désoxyribose),
- du phosphate (H3PO4),
- et une des 4 bases azotées : A, T, G, C.

L'ordre dans lequel se succèdent les nucléotides sur l'un des brins de l'ADN (l'autre est complémentaire) constitue une séquence de nucléotides spécifique à chacun d'entre nous « ***l'information génétique est constituée par l'ordre des nucléotides*** ».

Les brins de l'ADN ont les deux caractéristiques suivantes **[51]**:

- *Antiparallèles* : l'un est constitué d'un enchaînement commençant à gauche et se poursuivant vers la droite, l'autre commençant à droite et se poursuivant vers la gauche.
- *Complémentaires* : chaque adénine (A) d'un brin est liée par deux liaisons hydrogène avec une thymine (T) de l'autre brin, et chaque guanine (G) d'un brin est liée par trois liaisons hydrogène avec une cytosine (C) de l'autre brin.

Les dernières découvertes indiquaient que la majorité d'ADN humain est transcrit à des différents **ARN-transcrits** ; mais seulement l'ARN messager (ARNm) parmi ces ARN-transcrits qui est traduit à la suite à une protéine (ADN codant) **[22]**.

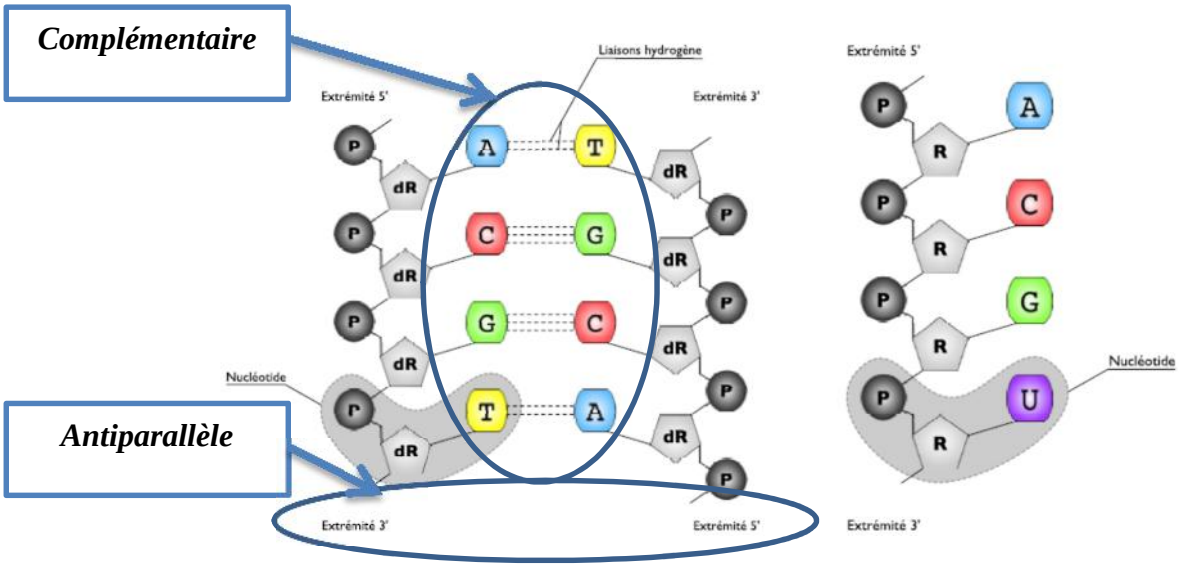
### *I.3.2.2. L'acide ribonucléique (ARN) :*

L'ARN ou l'acide ribonucléique est une macromolécule similaire à l'ADN, constituée d'un enchaînement de nucléotides sur un seul brin. L'ARN a de nombreuses similarités avec l'ADN, avec cependant quelques différences importantes **[6]**:

- L'ARN contient un ribose à la place du désoxyribose de l'ADN, ce qui rend l'ARN chimiquement plus instable et la thymine de l'ADN y est remplacée par l'uracile dans l'ARN, qui possède les mêmes propriétés d'appariement de base avec l'adénine.
- L'ARN est trouvé dans les cellules sous forme de simple brin, tandis que l'ADN est présent sous forme de deux brins complémentaires, formant une double hélice.
- Enfin les molécules d'ARN trouvées dans les cellules sont plus courtes (de quelques dizaines à quelques milliers de nucléotides) que l'ADN du génome (de quelques millions à quelques milliards de nucléotides).

L'ARN dans la cellule peut être codant ou non codant **[6]**:

- **ARN codant** : comme l'ARN messager qui traduits par la suite en protéine,
- **ARN non codant** : comme, les ARN ribosomiques et les ARN de transfert. Contrairement aux ARN messagers, ces ARN sont des molécules fonctionnelles non traduites en protéine**.**




**(a)** **(b)**

**Figure I.2 : La structure des acides nucléiques : (a) la structure d'ADN, (b) la structure d'ARN [6].**

#### *I.3.2.3. Les évènements génétiques :*

Au cours de l'évolution, la transmission du message génétique de cellule à cellule, d'individu à individu, d'espèce à espèce se fait avec des modifications ponctuelles de la structure primaire de l'ADN. Ces modifications sont de trois sortes :

- **La substitution :** est le remplacement d'un nucléotide par un autre dans la structure primaire d'un acide nucléique. Les substitutions de purine à purine ou de pyrimidine à pyrimidine appelés les transitions, sont les plus fréquentes.
- **La mutation** *:* est une modification de la séquence d'ADN (gène), Les mutations expliquent l'existence d'allèles[2] différents pour un même gène. Les mutations peuvent conduire à plusieurs maladies comme le **cancer**.
- **Insertion/Délétion** : délétion, c'est à dire suppression d'un ou de plusieurs nucléotides, insertion, c'est à dire addition d'un ou de plusieurs nucléotides.

### I.3.3. Le gène:

Le gène est un segment d'ADN constitué d'une séquence particulière de nucléotides, correspondant à un ou plusieurs caractères héréditaires codant pour des protéines. Dans la séquence du gène nous distinguons **[51]**:

- Une séquence spécifique nommée le **promoteur** : sur laquelle se fixe l'ARN polymérase, elle est nécessaire pour que la transcription débute. Le site promoteur diffère quelque peu selon les gènes et les organismes. Ce site comporte deux sous séquences spécifiques. Chez les bactéries, il est constitué d'une séquence canonique de **TATAAT**, marque le début de transcription et se situe à une dizaine de bases en amont du gène. Chez les eucaryotes et les archées, cette séquence est trouvée sous la forme **TATAAAA** (boîte **TATA**), située une vingtaine de bases en amont du gène. Pour tous les organismes, il existe la séquence **CAAT** (boîte **CAAT**) située 70 à 80 nucléotides en amont du gène qui sert à la régulation de la vitesse de transcription du gène**,**
- un site d'initiation de la transcription : sur laquelle la transcription du gène commencée.
- une suite d'exons et d'introns : un exon est une partie de la séquence d'un gène transcrit et conservée dans la structure de l'ARNm pour être traduite, par contre un intron est une partie de la séquence d'un gène transcrite mais coupée de la structure de l'ARNm pour ne pas être traduite.

[2] On appelle allèles les différentes versions d'un même gène. Un allèle se différencie d'un autre par une ou plusieurs différences de la séquence de nucléotides. Tous les allèles d'un gène occupent le même locus (emplacement) sur un même chromosome.

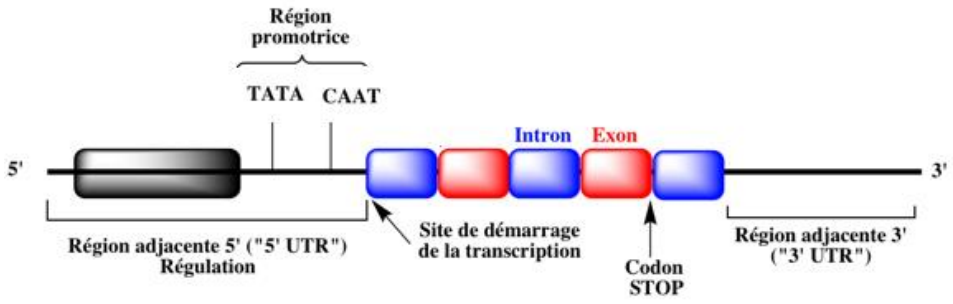



**Figure I.3 : La structure d'un gène [6].**

### *I.3.3.1. L'expression des gènes et le niveau d'expression génique:*

L'expression d'un gène, expression génique ou expression génétique est le processus qui convertit les informations codées dans un gène à des produits fonctionnels dans la cellule. Dans le cadre des projets génomiques impliquant des données d'expression génique, nous nous intéressons au niveau d'expression du gène qui se rapporte au nombre de copies d'ARN transcrits créées par la transcription d'un gène à un moment donné. Les gènes qui sont exprimés comprennent les gènes codant pour des protéines, ainsi que les gènes codant pour des produits ARN fonctionnels. Le plus souvent, l'analyse de l'expression des gènes se focalise sur le niveau d'expression des gènes codant pour des protéines **[22].**

### *I.3.3.2. La régulation de l'expression :*

La régulation de toute voie métabolique se fait principalement à la première réaction pour éviter des synthèses intermédiaires inutiles. Pour l'expression d'un gène, la régulation se fait à quatre niveaux **[51]**:

1. Lors de la transcription (Facteurs de transcription, ARN Polymérase, promoteur …etc.).
2. Lors de la maturation du transcrit (épissage, épissage alternatif, ..etc).
3. Lors de la traduction (Ribosome, ARNt…etc.).
4. Lors de l'activation de la protéine mature[3].

L'objectif de la régulation d'expression des gènes est de garantir le passage de l'information génétique incluse dans une séquence d'ADN à un produit fonctionnel (ARN ou protéine).

## I.3.4. Les protéines:

La découverte des protéines est due au chimiste Gerrit Mulder dans le XIXme siècle mais ce n'est qu'entre les années 1939 et 1941, grâce aux travaux de Linus Pauling, que la structure des protéines commença à être réellement élucidée. Les protéines sont des macromolécules

[3] On appelle protéine mature la forme chimique définitive que la protéine montrera au moment où elle remplira sa fonction dans l'organisme.

impliquées dans la majorité des réactions biologiques. Leurs fonctions incluent la catalyse de processus métaboliques sous la forme d'enzymes, elles jouent un rôle important dans la transmission du signal, mécanismes de la défense et les transports moléculaires, elles sont utilisées aussi comme une matière de construction, par exemple en cheveux. Sont des chaines des petites entités moléculaires, nommées **acides aminés** (aa) qui se composent d'un atome de carbone central noté **Cα**, connecté à un groupe aminé **NH2**, à un groupe carboxyl **COOH** et une chaîne latérale **R** qui est spécifique à un acide aminé particulier. Ces acides aminés sont liés entre eux par des liaisons **peptidiques**. Dans la nature, il existe plus d'une centaine d'acides aminés, cependant, seuls vingt d'entre eux peuvent être intégrés dans les protéines synthétisées **[45].**

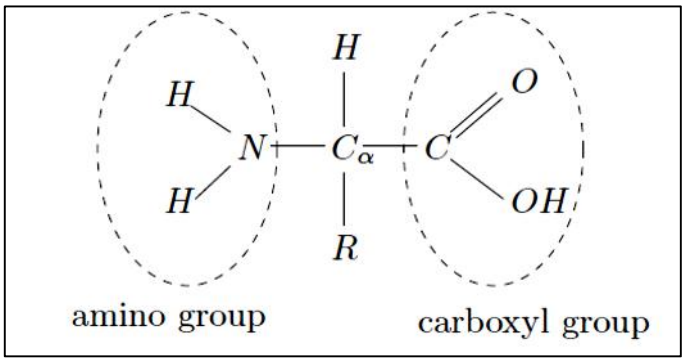



**Figure I.4 : La structure d'un acide aminé [45].**

### *I.3.4.1. La biosynthèse des protéines :*

L'expression des gènes dans les cellules aboutit à la synthèse des ARNs et protéines, dont la structure primaire est déterminée par celle de l'ADN. Cette expression se fait par deux mécanismes principaux : **la transcription** et **la traduction**.

a) La transcription : la synthèse de l'ARNm:

La transcription est le processus qui synthétise un ARN en recopiant la séquence d'un gène. Ce processus est réalisé dans le noyau de la cellule et il se décompose en trois étapes: l'initiation, l'élongation et la terminaison **[22]** :

- Durant la phase d'initiation, l'enzyme d'**ARN polymérase** se fixe sur une région particulière de l'ADN (gène), le site promoteur. La liaison entre l'ADN et l'ARN polymérase permet d'une part d'ouvrir la double hélice et d'autre part de catalyser l'insertion des nucléotides pour former un brin d'ARN.
- L'élongation de la transcription correspond à l'incorporation des nucléotides sur le brin d'ARN. Durant cette phase, l'ARN polymérase progresse de manière séquentielle de

l'extrémité 3'vers l'extrémité 5' du brin d'ADN codant (gène). L'incorporation des nucléotides se faisant par la complémentarité (**G→C, A→U**) entre nucléotides, l'ARN synthétisé est une copie conforme de la région à transcrire.

- La terminaison de la transcription intervient lorsque l'ARN polymérase rencontre un terminateur et le résultat est **l'ARN messager**. Chez les procaryotes, ce terminateur est le plus souvent une région riche en G et en C, suivie d'une série de A sur l'ADN. Chez les eucaryotes, les mécanismes de terminaison de la transcription sont moins connus.

### b) La maturation de l'ARNm :

Chez les eucaryotes, après une étape de transcription l'ARN messager (pré-ARNm) transcrit peut subir des régulations ou bien les transformations post-transcriptionnelles (la maturation). Les transformations sont **[6]**:

- L'addition de la coiffe à l'extrémité 5' : c'est un nucléotide modifié,
- l'addition d'une queue poly A à l'extrémité 3' : c'est une succession de nombreux nucléotides de type Adénosine (A), Les objectifs de la coiffe et la queue poly A sont la protection des ARNm contre les dégradations (garantir la stabilité de l'ARNm lorsque il passe du noyau vers le cytoplasme) et permettent aussi d'invoquer le ribosome pour la traduction.
- L'épissage, c'est le processus dans lequel des fragments de l'ARNm sont excisés et les autres sont conservés. Les fragments excisés sont les introns, les conservés sont des exons.

Un même transcrit primaire peut donner lieu à différents transcrits matures de longueurs différentes issus **d'épissages alternatifs** qui consiste à la délétion (excision) des introns ou des exons qui ne seront pas nécessaires au codage de la protéine.

### c) La traduction ou la translation :

La traduction est un processus qui, comme son nom l'indique, traduit l'information portée par un ARN messager en une protéine, ce processus est réalisé dans le cytoplasme de la cellule en trois étapes :

- L'initiation : le **ribosome**[4] lit l'ARN jusqu'à trouver un groupe précis de trois bases : **AUG (codon start)**. C'est l'initiation

[4] Le ribosome est un complexe composé d'ARN et de protéines ribosomiques, associé à une membrane ou libre dans le cytoplasme. Commun à toutes les cellules procaryotes et eucaryotes, il traduit des triplets de nucléotides (un codon) portés par les ARNm en acides aminés.

- L'élongation : le ribosome lit l'ARN par groupes de trois bases codon, chaque codon est associé à un acide aminé qui lui correspond. Les acides aminés sont liés par des associations nommées liaison peptidique. La correspondance entre les codons les acides aminés se fait par l'intermédiaire **d'un code génétique**[5].
- La terminaison est déterminée par la présence d'un codon **STOP.**

Dans la traduction nous trouvons la notion des cadres ouverts de lectures (en anglais Open Reading Frames « ORF ») est une séquence de codons dans l'ADN qui commence par un **codon start**, se termine par un **codon stop** et ne possède aucun autre codon stop à l'intérieur **[83].** Ils sont utilisés pour détecter des régions codantes potentielles.

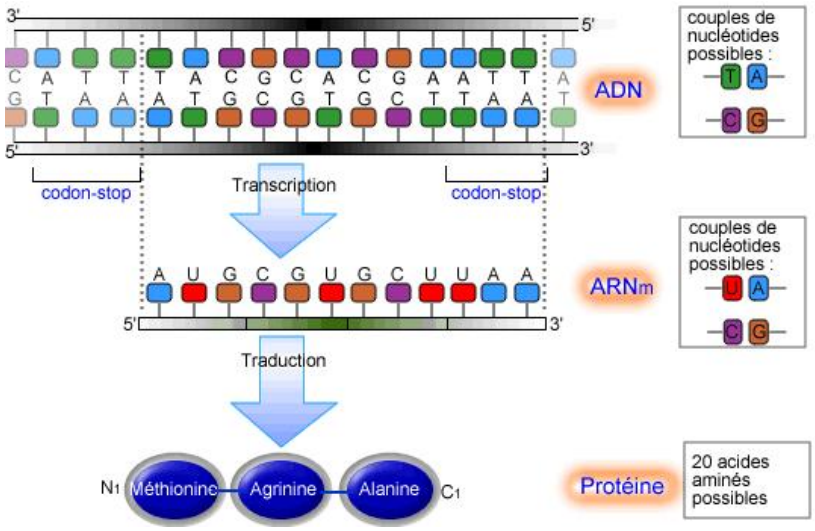



**Figure I.5 : Les étapes de la biosynthèse des protéines [61].**

### I.3.5. Le génome :

Le terme génome peut être défini soit comme l'ensemble complet des informations génétiques ou en tant que l'ensemble du matériel génétique contenu dans la cellule d'un organisme. Lorsqu'elle est appliquée sur le génome humain, cette définition inclut à la fois le génome nucléaire et le génome mitochondrial[6]. Néanmoins, dans le domaine de l'analyse d'expression génique, nous utilisons souvent le terme génome pour désigner le génome nucléaire uniquement, c'est à dire compris que la séquence d'ADN complète d'un ensemble de chromosomes de l'organisme. **[22].**

[5] Comme un dictionnaire qui nous aide à traduire les mots d'une langue étrangère, le code génétique permet dans chaque cellule de passer du langage de l'ADN à celui des protéines, à chaque séquence de trois bases consécutives portées par l'ARN messager, correspond un acide aminé donné et un seul

[6] Les mitochondries sont responsables, au sein de la cellule, de la respiration et par là, de la production d'énergie, elles contiennent leur propre ADN dit ADN mitochondrial ("ADNm") stocké sous forme de petites boucles. L'ensemble d'ADNm présent dans la mitochondrie représente le génome mitochondrial.

Le premier génome complet qui a été séquencé est celui d'Haemophilus influenzae en 1995 suivie par celui de la levure Saccharomyces cervisiae en 1996. La séquence complète du dernier chromosome dans le génome humain, chromosome 1, le plus grand, contenant à lui seul 8% de ce génome, a été obtenue en mai 2006 **[61]**.

### I.3.6. Le transcriptome :

Le transcriptome est l'ensemble des ARN messager présents dans une cellule à un instant T donnée et dans une condition biologique précise, issu de l'expression d'une partie du génome d'un tissu cellulaire ou d'un type de cellule. Chez l'homme, sur ~ 200 000 ARNm transcrits (différents) seuls 10 000 à 20 000 sont exprimés dans une cellule spécialisée, ces derniers constituent le transcriptome de la cellule. De plus, parmi ces transcrits, 4 000 à 6 000 semblent spécifiques de ce type cellulaire. La caractérisation et la quantification du transcriptome dans un tissu donné et à des conditions données permettent d'identifier les gènes actifs, de déterminer les mécanismes de régulation d'expression des gènes et de définir les réseaux d'expression des gènes. Une des techniques utilisées pour mesurer simultanément le niveau d'expression d'un grand nombre de types différents d'ARN messager est celle de la puce à ADN qui est détaillée dans les sections suivantes **[22]**.

### I.3.7. Le protéome:

Une définition simplifiée pourrait indiquer que le protéome est l'ensemble des produits de protéines exprimées par le génome. Cependant, à la différence du génome qui peut être considéré comme une entité stable, le protéome change constamment (principalement due à des interactions protéine-protéine et les modifications des conditions environnementales de la cellule). Ainsi, un protéome peut aussi être interprété comme, toutes les protéines exprimées dans une cellule ou un tissu à un instant donné **[22]**.

## I.4. Quelques techniques de la biologie moléculaire :

Depuis les années cinquante, les biologistes spécialisés en biologie moléculaire ont appris à caractériser, isoler et manipuler les composants moléculaires des cellules et des organismes. Ces éléments comprennent ADN, le référentiel de l'information génétique ; ARN, un proche parent de l'ADN; et les protéines, le type structurel et enzymatique majeur de la molécule dans les cellules.

### I.4.1. Le clonage moléculaire :

Le clonage peut être défini comme une multiplication in vitro[7] d'un organisme, d'une cellule souche ou d'une molécule, en grand nombre d'exemplaires identiques appelés des **clones**. Une des techniques les plus élémentaires en biologie moléculaire c'est le clonage d'ADN dont l'objectif est de cloner les gènes d'intérêt codant pour des protéines spécifiques. Le principe du clonage d'ADN est simple. Il consiste à insérer un segment d'ADN d'intérêt dans une petite molécule capable de se répliquer de façon autonome dans une cellule hôte[8]. Ce type de molécule d'ADN est appelé un vecteur de clonage. Un vecteur de clonage courant est le plasmide bactérien[9]. L'ADN du plasmide est coupé par une enzyme de restriction et lié par l'enzyme d'ADN ligase au fragment d'ADN d'intérêt coupé par la même enzyme de restriction ou une enzyme qui donne des extrémités compatibles, nous obtenons un nouveau vecteur nommé le vecteur recombinant (ADN recombinant[10]). Ensuite, le vecteur est transfecté dans une cellule hôte afin de les cloner en même temps avec la croissance de la cellule **[101]**.

### I.4.2. La réaction en chaîne polymérase (PCR) :

La réaction en chaîne polymérase (en anglais Polymerase Chain Reaction ou PCR) est une technique d'amplification génique in vitro, imaginée par Kary Mullis en 1984 (Prix Nobel dès 1993), elle permet de copier en grand nombre (avec un facteur de multiplication de l'ordre du milliard en quelques heures) une séquence d'acide nucléique (ADN ou ARN) connue, à partir d'une faible quantité. La PCR permet d'amplifier des fragments d'ADN et pour les fragments d'ARN, une étape supplémentaire est nécessaire : **la reverse transcription**. Cette étape est réalisée à l'aide d'une enzyme (**Reverse Transcriptase**) qui copie le mono brin d'ARN en ADNc, ensuite la PCR proprement dite peut commencer ; on parle de **RT-PCR**. Le principe de la PCR est résumé en trois étapes principales **[101]**:

1. **La dénaturation** de l'ADN à amplifier par chauffage (95 °C) pour séparer les deux brins qui le composent.

[7] Consiste à qualifier un processus biologique observé/étudié en éprouvette ou en laboratoire c'est-à-dire en dehors de l'organisme, dans des conditions artificielles, par opposition à in vivo qui consiste à qualifier le processus dans un organisme vivant.

[8] Est une cellule habituellement transfectée par des vecteurs contenant des fragments d'ADN et qui est permet d'amplifier ce vecteur en même temps que leur croissance.

[9] Un plasmide est une molécule d'ADN circulaire naturelle ou modifiée artificiellement qui possède obligatoirement une origine de réplication afin qu'il puisse se répliquer de manière autonome dans la cellule

[10] L'ADN recombinant, une molécule de l'ADN contenant des séquences nucléotidiques provenant d'au moins deux sources différentes.

2. **L'hybridation** deux amorces aux extrémités de la séquence recherchée, les amorces sont de petits brins d'ADN d'environ 20 bases, capables de s'hybrider de façon spécifique, grâce à la complémentarité des bases, sur le brin d'ADN ou sur son brin complémentaire. Les amorces sont choisies de façon à encadrer la séquence d'ADN à amplifier.
3. **L'élongation** grâce à l'action d'une ADN polymérase, elle progresse l'amorce de manière séquentielle de l'extrémité 5' vers l'extrémité 3' du brin d'ADN à amplifier.

Ce cycle est répété un grand nombre des fois pour obtenir une multiplication exponentielle de la séquence d'ADN cible.

### I.4.3. L'hybridation moléculaire sur les biopuces :

L'utilisation des techniques d'hybridation moléculaire a commencé dans les années 1970, mais c'est en 1996 que ce type d'approche a pris toute son ampleur par la miniaturisation de dispositifs expérimentaux. L'avènement des biopuces comme une technique d'hybridation a été le premier pas vers une biologie quantitative à grande échelle. Les microarrays ou les biopuces sont des technologies à haut débit de la biologie moléculaire, sont des objets micro structurés, fabriqués en utilisant des techniques héritées de la micro-électronique (les puces). Le dispositif principal de cette technologie est les puces, qui sont des supports de verre (la plupart des cas), de silicium ou de plastique, la mensuration de ces plaques étant de **25x75 mm** sur lesquels nous trouvons des spots (ou dépôts) étant des « trous » réalisés à l'aide d'aiguilles creuses extrêmement petites. Ces trous sont espacés de **80** à **300** $\mu m$ et la distance entre les spots est **de 250** $\mu m$ .

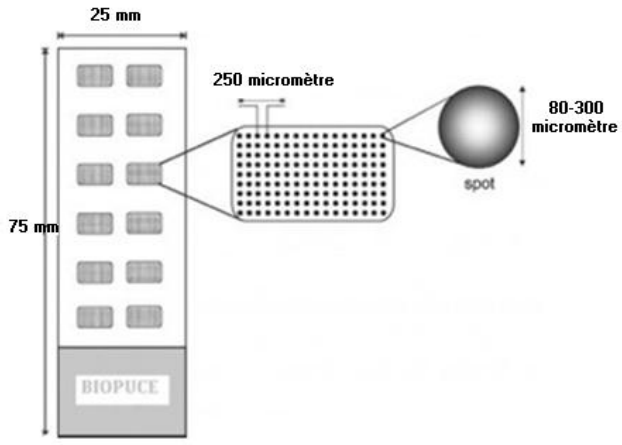



**Figure I.6 : Structure d'une biopuce [114].**

Les biopuce les plus connues sont les puces à ADN et les puces à protéines. Pour les puces à ADN, dans les spots nous trouvons un ADNc simple brin ou sonde (séquence connue) qui est contacté à des transcrits cibles (ARNm transcrit à ADNc) issus d'une cellule cible (séquence

inconnue). L'hybridation est basée sur la complémentarité (A-T et C-G), les cibles reconnaissent, parmi les sondes de la puce, celles qui sont leur complémentaire, et s'y apparient. Les puces à protéines sont utilisées pour observer directement et à grande échelle la présence des protéines, leur degré d'activité et pour analyser les interactions de certaines protéines avec d'autres composants. Le principe de ces puces est très proche de celui des puces à ADN, il s'agit d'apparier des protéines sondes (déposées dans les spots) avec les protéines cibles issues d'échantillons biologiques. L'hybridation est basée sur les interactions : un récepteur fixe et son ligand (amarrage moléculaire), un anticorps et son antigène, … Les interactions protéiques qui auront lieu entre ces anticorps et ces antigènes vont permettre de diagnostiquer la maladie. La principale difficulté de cette technologie réside dans la production d'anticorps (sondes) spécifiques de chacune des protéines cellulaires **[78].**

## I.5. Conclusion :

Dans ce chapitre, nous avons résumé l'essentiel de la biologie moléculaire. Dans un premier temps nous avons expliqué cette discipline scientifique et leur croisement avec les disciplines de la biochimie et la génétique. Ensuite, nous avons présenté quelques notions et terminologies biologiques afin de comprendre tous ce qui suit dans notre recherche. Nous avons montré dans cette section le dogme central de la biologie moléculaire qui précise bien la relation entre les trois macromolécules cellulaire (ADN, ARN et protéine) et qui postule que le passage d'une protéine à un acide nucléique ou d'une protéine à une protéine est impossible. Par la suite, nous avons détaillé le processus biologique qui permet de synthétiser les protéines au sein de la cellule, qui est composé de deux étapes la transcription qui est réalisé dans le noyau de la cellule et la translation ou bien la traduction qui est réalisée dans le cytoplasme de la cellule. Finalement, nous avons élucidé à quelques techniques connues de la biologie moléculaire, avant d'expliquer ces techniques nous avons élaboré quelques outils biologiques qui sont utilisés par ces techniques afin d'étudier les macromolécules cellulaire. Ces techniques représentent les sources d'une grande quantité de données biologiques. L'analyse et l'interprétation de ces données nécessite l'intervention d'une nouvelle discipline nommée la bioinformatique. Dans le chapitre suivant, nous présenterons cette discipline et ses apports à la biologie moléculaire.

## II.1. Introduction :

Les liens entre la biologie et particulièrement la biologie moléculaire et l'informatique commencent très tôt. Depuis les années 1990, cette coopération s'est renforcée à la fois pour aider les biologistes dans leur travail quotidien, mais aussi pour transmettre une partie des découvertes fondamentales en produits informatiques concrets réutilisables. A tel point, qu'il y a un peu plus d'une dizaine d'années, une nouvelle discipline a été créée nommée « *la bioinformatique* », qui est définie généralement par une application de l'informatique dans la biologie. La bioinformatique moderne est née de la convergence de deux aspects de recherche en biologie moléculaire: le stockage des séquences moléculaires sur ordinateurs sous la forme de bases données et l'application d'algorithmes mathématiques et outils informatiques pour analyser et interpréter ces données. Le stockage fait engendre la notion de banques de données biologiques dans lesquels les séquences nucléiques et protéiques sont stockées d'une manière bien organisée où le chercheur peut les consulter et manipuler facilement. L'analyse et l'interprétation des données stockées fait apparaitre des multiplicités des domaines de recherches en bioinformatique et avec l'évolution des technologies à haut débit qui augmente la quantité de données biologiques *omiques*, les défis en bioinformatique sont remarquable. Dans ce chapitre, nous présenterons un état de l'art en matière de bioinformatique, notre chapitre commencera par l'historique et les définitions, pour terminer par les grands défis du domaine.

## II.2. Historique :

Le terme *bioinformatique* n'est apparu dans la littérature scientifique qu'au début des années 1990, cependant ce domaine de recherche existait bien avant l'essor de la génomique **(section II.6)** et des dizaines de laboratoires dans le monde travaillent depuis longtemps en *biomathématiques*[1] ou *biométrie*[2]. La première utilisation réelle du terme bioinformatique a été en 1993 où il apparaît 3 fois dans les articles du domaine puis 9 et 10 fois en 1994-95 pour ensuite augmenter de façon exponentielle. La synthèse des étapes d'évolution de la bioinformatique montre que leurs premières étapes coïncident avec celles de la biologie moléculaire **[81]**. Le tableau suivant retrace l'émergence du domaine, et montre à quel point il est lié à la biologie moléculaire et leur évolution.

[1] Ensemble des méthodes et techniques mathématiques qui permettent d'étudier et de modéliser les phénomènes et processus biologique.

[2] Analyse mathématique des caractéristiques biologiques d'une personne, destinée à déterminer son identité de manière irréfutable. Elle repose sur le principe de la reconnaissance de caractéristiques physiques.

| Année | Les évènements |
|---|---|
| **1953** | Structure en double hélice de l'ADN (Watson-Crick). |
| **1956** | • Séquence en acides aminés de la première protéine: insuline<br>• Fortran ((FORmula TRANslation) : Premier langage informatique de haut niveau. |
| **1958** | Première structure 3D de protéine (myoglobine, Kendrew) |
| **1955-1965** | Premiers langages informatiques, premier ordinateur commercial |
| **1961** | • Nirenberg et Mattaei déchiffrent le code génétique.<br>• Sidney Brenner, François Jacob, Matthew Meselson identifiés l'ARN messager. |
| **1965** | • Jacques Monod et François Jacob découvrent les mécanismes de la régulation génétique impliqués dans le dogme central de la biologie moléculaire, énoncé initialement par Crick.<br>• Ordinateur IBM/360. |
| **1967** | "Construction of Phylogenetic Trees" (Fitch et Margoliash). |
| **1969** | Conception du système d'exploitation Unix |
| **1970** | Premier programme pour la comparaison de séquences protéiques Alignement optimal entre deux séquences (Needleman&Wunsh) |
| **1971** | • PDB - Protein Data Bank (structures 3D macromolécules)<br>• Premier microprocesseur INTEL |
| **1972** | Clonage de fragments d'un plasmide bactérien dans le génome d'un virus pour les reproduire à volonté : c'est l'ADN recombiné. (Berg, Jackson et Symons) |
| **1974** | Programme de prédiction de structures secondaires des protéines (Chou et Fasman) : "Prediction of Protein Conformation". |
| **1977** | Mise au point des techniques de séquençage de l'ADN. |
| **1978** | Matrice de substitution[3] (**PAM**) (Dayhoff et. al.) après l'alignement d'environ 1300 séquences appartenant à 71 familles de protéines. |
| **1980** | Création de la banque EMBL : banque européenne généraliste de séquences nucléiques. |

3 Point Accepted Mutation : Ce type de matrice donne la probabilité que, suite à une mutation par substitution au cours de l'évolution, n'importe quel acide aminé remplace n'importe quel autre acide aminé sans que la fonction de la protéine ne soit altérée.

| | |
|---|---|
| **1981** | • Alignement local de séquences (algorithme de Smith et Waterman)<br>• Naissance du 1er animal transgénique (une souris) (Franck H.Ruddle et John W. Gordon)<br>• PC IBM micro-ordinateur personnel |
| **1982** | • Création de la banque Genbank : banque américaine généraliste de séquences nucléiques.<br>• Localisation du gène responsable de la myopathie de Duchenne. |
| **1984** | • Développement de la réaction de polymérisation en chaîne (PCR) par Mullis.<br>• Création de la banque NBRF : banque américaine généraliste de séquences protéiques.<br>• Logiciel d'analyse de séquence (UW GCG) Devereux et. al. |
| **1985** | • CABIOS (première revue de bioinformatique)<br>• Programme Fasta (Pearson- Lipman) : recherche rapide d'alignements locaux dans une banque. |
| **1986** | • Création de la banque DDBJ : banque japonaise généraliste de séquences nucléiques.<br>• Création de la banque SwissProt : banque généraliste de séquences protéiques créée à l'Université de Genève.<br>• Clonage du gène responsable de la forme de myopathie de Duchenne de Boulogne (dystrophie musculaire). |
| **1987** | • 1ère carte génétique du génome humain.<br>• Apparition de la technologie des puces à ADN.<br>• Genbank, EMBL et DDBJ s'échangent leur contenu et adoptent un système de conventions communes (The DDBJ/EMBL/Genbank feature Table Definition) |
| **1988** | • Création du projet HUGO (Human Genome Organization) pour coordonner les efforts de cartographie et de séquençage entrepris dans le monde et éviter les doublons. |
| **1989** | L'apparition de l'Internet |

| | |
|---|---|
| **1990** | • Programme BLAST (Altschul et al.) : recherche rapide d'alignements locaux dans une banque.<br>• Premier essai de thérapie génique.<br>• Les NIH (National Institutes of Health) et le DOE présentent au Congrès le projet HGP (Human Genome Project) qui vise à séquencer entièrement le génome humain et à identifier l'ensemble des gènes humains. |
| **1991** | • Prédiction de la structure tertiaire de protéines Bowie et. al.<br>• Programme Grail (Mural et al.) pour la localisation de gènes. |
| **1992** | • L'utilisation du terme bioinformatique comme une nouvelle discipline.<br>• Séquençage complet du chromosome III de la levure Saccharomyces cerevisiae.<br>• 13 gènes de maladies génétiques ont été identifiés par clonage positionnel. |
| **1993** | • GeneMark - Programme de Prédiction gènes génomes bactériens Borodovsky et. al.<br>• Etzold et Argos créent SRS, logiciel d'interrogation multi banques accessible sur le web<br>• Mosaic (NCSA), premier navigateur INTERNET. 200 sites web disponibles. |
| **1996** | • Séquençage du 1er génome eucaryote, Saccharomyces cerevisiae par Dujon. |
| **1997** | • Clonage de la brebis Dolly.<br>• PFam - Banque de domaines protéiques par Sonnhammer et. al.<br>• GENSCAN - Prédiction de génomes eucaryotes par Burge et. al. |
| **1998** | • Séquençage du 1er organisme pluricellulaire (premier animal), Caenorhabditiselegans. |
| **1999** | • Publication de la séquence complète des chromosomes 21 et 22 humain |
| **2000** | • 14 mars 2000. Tony Blair et Clinton publient une déclaration selon laquelle les données brutes sur le génome humain doivent être mises gratuitement à la disposition des scientifiques.<br>• 26 juin 2000. A la Maison Blanche, le Dr Craig Venter (projet privé : |

| | |
|---|---|
| | Celera Genomics), Bill Clinton et le Dr Francis Collins (projet public : HumanGenom Project) et Tony Blair (par satellite) annoncent la fin du séquençage "brut" du génome humain.<br>• Publication du "working draft" (brouillon) de la première carte complète du génome humain.<br>• Séquençage du premier génome de plante, Arabidopsisthaliana. |
| **2001** | • Publication des travaux de séquençage du génome humain presque complet.<br>• 13 octobre 2001. Premier clone d'embryon humain |
| **2002** | • Projet protéome humain (HPP)<br>• 2 avril 2002. Clonage du lapin par des chercheurs français de l'INRA |
| **2003** | Séquençage de plusieurs organismes eucaryotes. |
| **2005** | Séquençage à haut débit[4] (high-throughput sequencing en anglais HTS) |
| **2007** | L'apparition de la première génération des appareils de séquençage à haut débit :<br>• le pyroséquençage,<br>• le séquençage avec des terminateurs réversibles<br>• et le séquençage par ligation. |
| **Janvier 2012** | Plus de 3040 génomes eucaryotes et procaryotes séquencés et des milliers en projet (Genomes OnLine)<br>**Plus de 393.milliards de nucléotides** |
| **Nos jours** | • Plus 3000 article publiés concernant la bioinformatique.<br>• Plus de 1400 bases de données biologiques. |

**Table II.1 : les étapes d'évolution de la bioinformatique**

La bioinformatique continue d'être dans une période de croissance très rapide, parce que les besoins d'une matière de stockage de l'ïnformation, la recherche et l'analyse en biologie moléculaire et la génomique ont considérablement augmentés au cours de la dernière décennie. Les types des données recueillies par les biologistes aujourd'hui vont

[4] Un ensemble de méthodes produisant des millions de séquences en un *run* et à faibles coût.

considérablement modifier les types d'informations et de technologies qui vont mettre à la disposition des chercheurs demain **[94].**

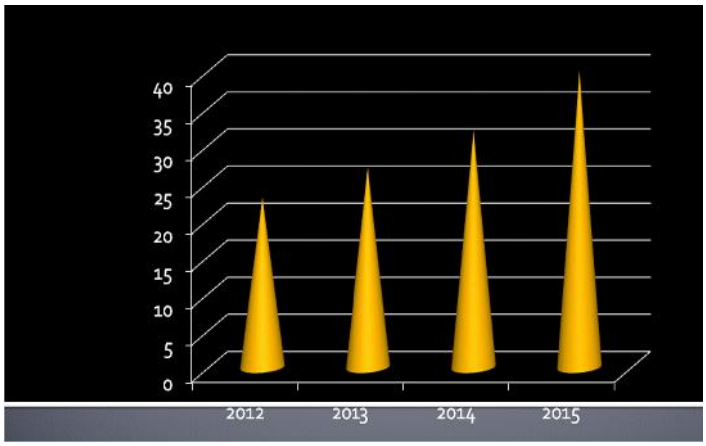



**Figure II.1 : Le taux de croissance mondiale de la bioinformatique. Taux atteint 22% de 2012 à 2015 (source : Realtime Market Research)**

## II.3. Définition :

La bioinformatique est un domaine multidisciplinaire qui utilise des méthodes informatiques, mathématiques, statistiques, combinatoires… pour résoudre un problème biologique. Pour la plupart des membres de la communauté scientifique, la bioinformatique est l'étude de la façon dont la bioinformation est représentée et analysée dans les systèmes biologiques, en particulier les informations obtenues au niveau moléculaire **[94].**

D'après Claverie et al. **[20]** « *la bioinformatique est la discipline de l'analyse de l'information biologique, en majorité sous la forme de séquences génétiques et de structures de protéines. C'est le décryptage de la bioinformation* (*Computational Biology" en anglais). La bioinformatique est donc une branche théorique de la Biologie. Son but, comme tout volet théorique d'une discipline, est d'effectuer la synthèse des données disponibles (à l'aide de modèles et de théories), d'énoncer des hypothèses généralisatrices (ex. : comment les protéines se replient ou comment les espèces évoluent), et de formuler des prédictions (ex. : localiser ou prédire la fonction d'un gène)* ».

D'après Andrade et Sander **[3]** « *Bioinformatics is a science of recent creation that uses biological data, completed by computational methods, to derive new biological knowledge* ». Cette définition, plus moderne et large, sous-entend que la bioinformatique ne se limite

évidemment pas à l'analyse des séquences mais aussi des données concernant les marqueurs moléculaires (biomarqueurs), les données phénotypiques, etc. La bioinformatique est une approche *in silico*[5] de la biologie traditionnelle qui vient compléter les approches classiques *in situ* (dans le milieu naturel), *in vivo* (dans l'organisme vivant) et *in vitro* (en éprouvette).

Dans la pratique, la définition utilisée par la plupart des gens est plus étroite; la bioinformatique pour eux est synonyme de «*biologie computationnelle* ou *computational biology* en anglais », mais il y a une différence entre les deux termes **[69]**:

- *Computational biology* : est l'étude des données biologiques en utilisant des techniques de calcul. L'objectif est d'apprendre de nouvelle biologie, des connaissances sur les systèmes vivre. Donc il s'agit de la science.
- *Bioinformatic* : discipline plus pragmatique, elle vise à créer et appliquer des outils (algorithmes, modèles statistiques, bases de données) dont l'objectif est d'interpréter, classer et comprendre les données biologiques. Il s'agit de l'ingénierie**.**

Selon les définitions données à la bioinformatique, nous pouvons résumer les activités de cette discipline en trois points principaux **[25]**:

- Acquisition et organisation des données biologiques.
- Conception de logiciels pour l'analyse, la comparaison et la modélisation des données.
- Analyse des résultats produits par les logiciels.

## II.4. Les sources de données biologiques (bioinformation) :

Les sources de la bioinformation sont multiple, selon Sean,D. et al. **[94]** Il existe trois sources fondamentales qui sont en train de révolutionner notre compréhension de la biologie humaine et qui créent des défis importants pour la bioinformatique **[94]**:

**1.** *Le projet du Génome Humain et l'étude du génome*: le type le plus dominant de la bioinformation est *la séquence* qui a été activée par le Projet du Génome Humain, c'est un projet international visant à déterminer la séquence complète de l'ADN humain codé dans chacun des 23 chromosomes c'est à dire le génome humain, premièrement le projet a été publié en 2001 et la version finale a été annoncé en 2003 coïncide avec le 50e anniversaire de la découverte d'une structure en double hélice de l'ADN par Watson et Crick. La séquence continue d'être révisée et raffinée et des efforts sont en cours pour séquencer les génomes de nombreux individus différents.

[5] Désigne une recherche ou un essai effectué au moyen de calculs complexes informatisés ou de modèles informatiques.

2. *L'étude du protéome* : nommée aussi la protéomique (**section II.6**) avec cette discipline les chercheurs peuvent découvrir les états des protéines dans l'organisme. Ces états de protéines représentent des nouvelles informations biologiques qui peuvent être utilisées pour identifier les marqueurs d'une maladie humaine.
3. *Les technologies à haut débit* : qui sont utilisées pour recueillir des données sur des milliers ou des millions de molécules simultanément. Avec ces technologies nous pouvons suivre la production et la dégradation de molécules afin d'extraire des nouvelles informations (ex. expression des gènes) sur ces molécules et les utiliser par exemple pour diagnostiquer les maladies.

## II.5. Le stockage de la bioinformation : les Banques de données biologiques :

Le stockage, l'organisation et la diffusion de la bioinformation est l'un des aspects importants dans la bioinformatique c'est-à-dire toutes les informations connues sont mises à disposition des chercheurs du monde entier le plus rapidement possible et elles peuvent être récupérées et utilisées par d'autres chercheurs dans l'avenir. Il s'agit non seulement des séquences nucléiques ou protéiques brutes (successions de bases azotées ou acides aminés) mais également de toutes les annotations[6] des séquences et autres informations connexes. A cette raison, l'utilisation des bases de données d'intérêt biologique a été introduite. Nous distinguons deux types de bases de données **[69]**:

- Celles qui correspondent à une collecte des données la plus exhaustive possible et qui offrent finalement un ensemble plutôt hétérogène d'informations.
- Celles qui correspondent à des données plus homogènes établies autour d'une thématique et qui offrent une valeur ajoutée à partir d'une technique particulière ou d'un intérêt suscité.

En biologie, il est fréquent d'appeler les premières « banques de données » et les secondes « bases de données », mais cette distinction n'est pas universelle en dehors du domaine biologique. Aussi, pour éviter toute confusion sémantique nous parlerons ici de banques de données ou bases de données généralistes (pour les premières) et spécialisées (pour les secondes).

***N.B :*** **La séquence** est l'élément central autour duquel les banques de données se sont constituées.

[6] Des informations relatives à la séquence.

### II.5.1. Les banques de données généralistes :

Ce type de banques vues comme des bibliothèques de fiches descriptives de séquences nucléiques ou protéiques, quel que soit l'organisme dont elles sont issues, et quelle que soit leur nature (ADN, ADNc, ARN, protéine) avec des commentaires structurés issues d'expertises biologiques ou d'analyses bioinformatiques (annotation) **[16]**.

*Avantage* **[16]***:*

- Enorme richesse de séquences en un seul ensemble.
- Grande diversité d'organismes.
- Nombreuses informations qui accompagnent les séquences (annotations, expertise, bibliographie).
- Présence de lien vers d'autres bases de données spécialisées**.**

*Inconvénient* **[16]***:*

- Manque de vérification des données soumises.
- Retard dans l'insertion de nouvelles séquences.

#### a) Les banques nucléiques :

Les données stockées dans ce type de banques sont des données issues de séquençage d'ADN et ARN. Trois banques nucléiques connues, elles partagent des informations et donc contiennent des ensembles presque identiques de séquences**.** Ces trois banques s'échangent systématiquement leur contenu depuis 1987 et ont adopté un système de conventions communes : « The *DDBJ/EMBL/GenBank* Feature Table Definition » **[82] :**

- *La banque EMBL:* créée en 1980 et financée par l'EMBO (**E**uropean **M**oleculary **B**iology **O**rganization), développée au sein du Laboratoire Européen de Biologie Moléculaire situé à Heidelberg (Allemagne), elle est maintenant diffusée par l'EBI : http://www.ebi.ac.uk/embl/. En 24 février 2014, la banque contient 369.5 millions séquences.
- *La banque GenBank* **(Genetic Sequence Databank)**: créée en 1982 par la société IntelliGenetics et diffusée maintenant par le NCBI (**N**ational **C**enter for **B**iotechnology **I**nformation) : http://www.ncbi.nlm.nih.gov/. En février 2014 la banque contient 171123749 séquences. GenBank contient une sous-banque de protéines, traduction des séquences nucléiques, appelée **GenPept**.
- *La banque DDBJ* **(DNA Databank of Japan)***:* créée en 1986 et diffusée par le NIG (National Institute of Genetics, Japon).

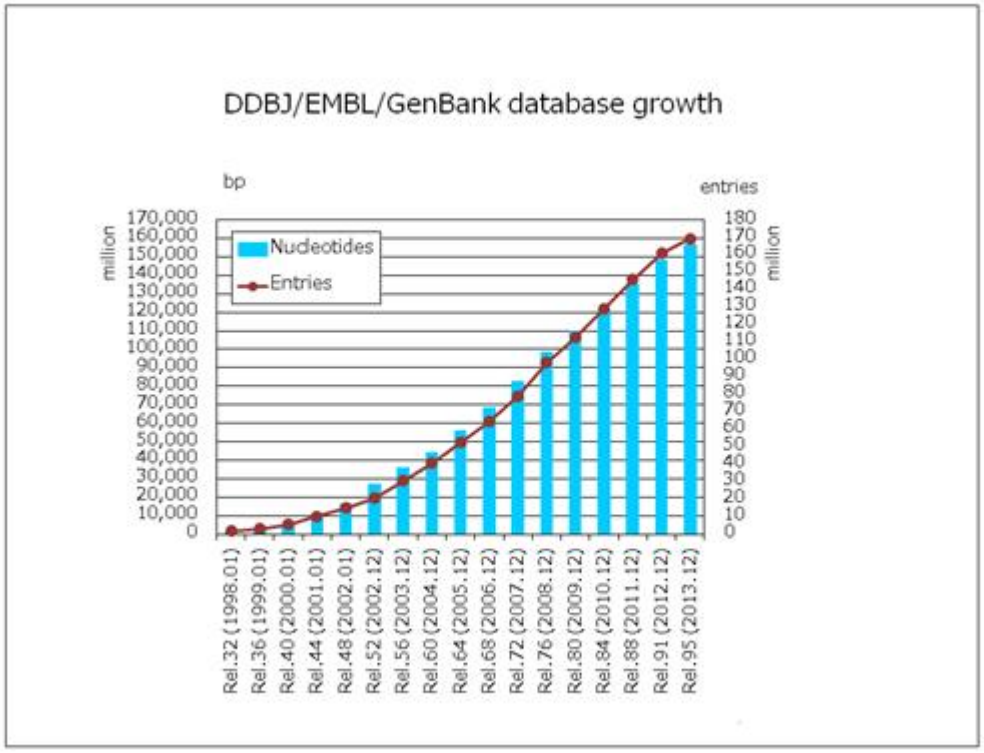



**Figure II.2: La croissance des trois banques : DDBJ/EMBL/GenBank de janvier 1998 à décembre 2013 (source : http://www.ddbj.nig.ac.jp/breakdown_stats/dbgrowth-e.html)**

b) Les banques protéiques :

Les données stockées dans ces bases sont issus d'une traduction de séquences d'ADN ou par le séquençage de protéines (rare car long et coûteux) **[69]** :

- *La banque SwissProt* **:** est une banque protéique crée en 1986 à l'Université de Genève et maintenue depuis 1987 dans le cadre d'une collaboration, entre cette université (via ExPASy, Expert Protein Analysis System) et l'EBI. Celle-ci regroupe aussi des séquences annotées de la banque PIR-NBRF ainsi que des séquences codantes traduites de l'EMBL. En février 2014 la banque contient 542503 séquences compressant 192888369 acides aminés.

- *La banque TrEMBL*: distribuée par l'EBI. Contient la traduction des parties codantes (CDS[7]) des séquences nucléiques stockées dans EMBL à l'exception de celles déjà présentes dans SwissProt**.**

### II.5.2. Les banques spécialisées ou thématique :

Les banques généralistes présentent des avantages (exhaustivité) et des limites (imprécisions, redondance, …). Les banques de données spécialisées sont créées pour des

[7]Coding DNA Sequences

besoins spécifiques (bien précis et pas de redondance), ce sont des banques qui synthétisent l'information pour un organisme particulier ou pour un domaine particulier **[16]**.

*Avantage :*

- Facilité pour mettre à jour les données.
- Vérifier leur intégrité.
- Offrir une interface adaptée.

*Inconvénient :*

- Ne cible pas toujours ce que le nous voulons (c'est-à-dire toutes les banques possibles n'existent pas).

a) Les banques nucléiques spécialisées :

Elles sont spécialisées dans les informations suivantes : ADNc, ARN, structure secondaire d'ARN, signaux et éléments de régulation, alignements, famille de gènes, sondes, amorces.

b) Les banques protéiques spécialisées :

Elles sont spécialisées dans les informations suivantes : motifs, alignement, classification structurale, familles de protéines, interactions, enzymes, modifications protéiques post-traductionnelles, pathologies, bases protéiques sur l'interaction et la thermodynamique des protéines.

c) Les banques structurelles :

Sont des banques spécialisées pour les structures 2D et 3D des protéines. Plusieurs banques connues dans ce contexte **[69]** :

- *La banque PDB* (**Protein Data Bank**): créée en 1971, c'est la banque de référence des structures protéiques obtenues expérimentalement par cristallographie rayon X, spectroscopie RMN et cryo-microscopie électronique (technique la plus récemment utilisée). Les coordonnées des atomes formant la structure d'une protéine, le détail de la séquence, les conditions de cristallisation sont les principales informations disponibles pour chaque structure de la banque. C'est à partir de cette banque que sont détectés les homologues structuraux.
- *La banque SCOP* (**Structural Classification of Proteins**) : banque de données regroupant les protéines de la PDB présentant une relation de similarité structurale et d'évolution.

### II.5.3. Interrogation des banques de données :

Toutes les banques de données possèdent leurs systèmes (outils ou logiciels) d'interrogation où la recherche porte sur les informations relatives à la séquence (non sur la séquence elle-même). Chaque système utilise une syntaxe particulière pour les requêtes d'interrogation (étiquettes, connecteurs logiques, caractères de substitution…etc.) **[16]**. Les systèmes plus utilisés sont :

- *Le système SRS* (**Sequence Retrieval Software**) :

Cet outil a été créé en 1993 par Etzold et argos. C'est un outil facile à utiliser, il permet des recherches simples et croisées (sur plusieurs bases en même temps jusqu'à 90). Nous pouvons y accéder grâce à l'adresse suivante : http://srs.ebi.ac.uk/

- *ENTREZ :*

Développé par NCBI. Ne permet d'interroger que les bases de données du NCBI. Nous pouvons y accéder grâce à l'adresse suivante http://www.ncbi.nlm.nih.gov/sites/gquery.

- *Acnuc :*

Développé au sein du PBIL (Pôle Bioinformatique Lyonnais). Ressemble à un système de SGBD mais ne permet d'interroger qu'une seule banque à la fois. Peut interroger les banques GenBank, EMBL, SwissProt, PIR, TrEMBL.

## II.6. Les champs liés à la bioinformatique:

La bioinformatique est une discipline « **hybride** » liée à plusieurs champs actifs en même temps, en plus de la biologie, les mathématiques, les statistiques elle est liée à plusieurs autres champs.

a) La génomique

Théoriquement, la génomique a commencée avec l'élucidation de la structure en double hélice de l'ADN par Watson et Crick en 1953. Le chemin vers la génomique a été projeté par le développement de certaines technologies telles que le clonage de l'ADN, le séquençage de l'ADN et l'amplification de l'ADN. La génomique comprend la détermination de la séquence de l'ADN entière dans les chromosomes d'un organisme et comprendre la nature de ces séquences. Les séquences d'ADN entières de plus de 500 organismes ont été déterminées. Il a commencé avec la séquence d'un virus bactérien et a progressé vers le génome de l'être humain **[79].**

Au sens large la génomique inclue le séquençage et l'analyse de génome et de l'ensemble de leurs gènes, et au sens exact, la génomique est l'étude de l'ensemble des gènes des

organismes vivants, de leur disposition sur les chromosomes, de leur séquence et de leur fonction. L'objectif est de réaliser l'inventaire des gènes qui s'expriment dans un type cellulaire donné, à un instant donné et dans un environnement donné. La génomique est de deux branches **[79]**:

- *Génomique structurale :* qui se charge du séquençage du génome entier. Elle permet l'annotation des génomes et l'identification des séquences informatives (les gènes avec ou sans introns codant des protéines ou des ARN fonctionnels, les séquences répétées, les éléments transposables, ...).
- *Génomique fonctionnelle :* qui vise à déterminer la fonction et l'expression des gènes séquencés grâce à leurs produits d'expression (ex : nombre de copies d'ARNm produites) dans différentes conditions (avant et après stimulation) ou selon l'origine du tissu (tissu sain, tissu pathologique, tissu d'origine anatomique variable...).

b) La génomique comparative :

Vise à étudier des similarités et des différences dans l'organisation et la structure des génomes. L'objectif de la génomique comparative est de tirer des conclusions particulières à propos de la biologie des espèces et tirer des conclusions plus générales à propos de leur évolution. Par exemple **[22]**:

- *Comparaison de génome de deux ou plusieurs espèces* : permet de déceler des différences et similitudes d'organisation structurale ; permet d'approcher la fonction des gènes ; permet d'agencer les espèces les unes par rapport aux autres et définir leur filiation.
- *Comparaison de génomes d'individus d'une même espèce :* aide à mieux comprendre l'interaction entre les gènes d'un individu et la réponse de son corps aux médicaments ; permet de découvrir des gènes de prédisposition à de nombreuses maladies (cancer, maladies cardiovasculaires, diabète, ...).

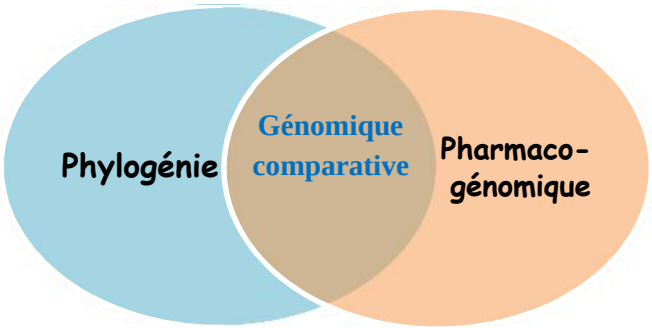



**Figure II.3 : Positionnement de la génomique comparative**

c) La transcriptomique :

La transcriptomique est l'étude de l'ensemble des ARN messagers produits lors du processus de transcription d'un génome. Elle repose sur la quantification systématique de ces ARNm, ce qui permet d'avoir une indication relative du taux de transcription de différents gènes dans des conditions données.

d) La protéomique :

Comment les chercheurs peuvent-ils savoir ce que font les cellules d'un tissu à une certaine heure de la journée ? En étudiant les protéines présentes dans ce tissu à ce moment-là. C'est le domaine de la protéomique. Cette approche permet non seulement d'identifier des protéines mais aussi de comparer leur production après la prise d'un médicament (la protéomique comparative), par exemple, ou en fonction d'un régime alimentaire ou encore au cours du cycle jour-nuit. Au sens large la protéomique est l'étude du protéome, dans le but de déterminer l'activité, la fonction et les interactions des protéines, et cela dans diverses conditions. La protéomique apporte des réponses auxquelles la transcriptomique ne peut répondre **[61]**:

- Compléments d'informations sur les modalités d'expression des gènes pour les organismes dont le génome n'a pas encore été séquencé ou pour lesquels les programmes de prédiction de séquences codantes sont moins fiables.
- Estimation quantitative des concentrations des protéines synthétisées.
- Obtention de données sur la fonction des protéines et les interactions entre protéines ou entre protéines et autres molécules biologiques**.**

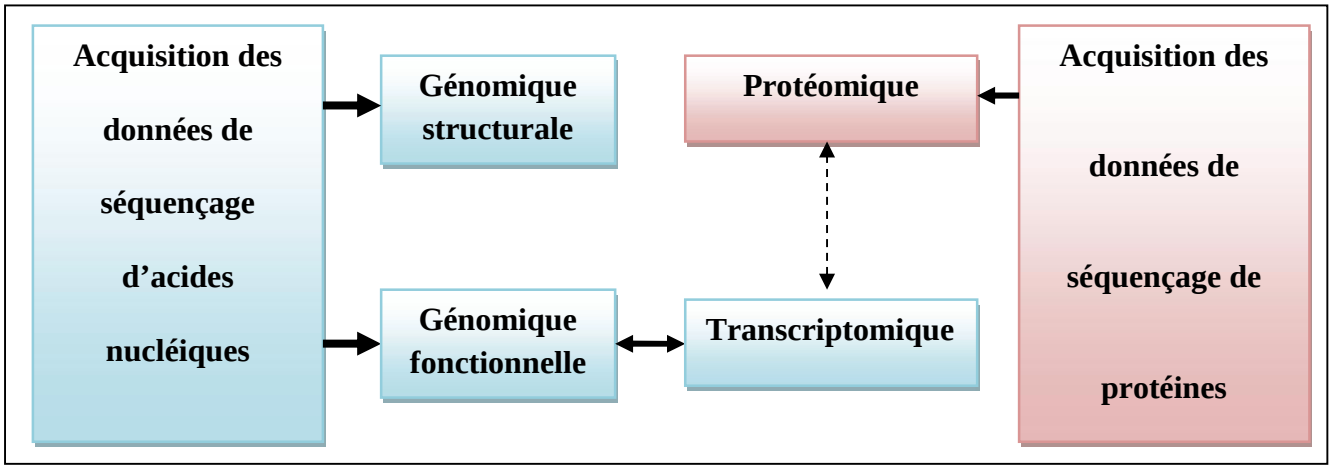



**Figure II.4 : La relation entre la génomique, la transcriptomique et la protéomique**

e) La métabolomique :

La métabolomique c'est la discipline qui s'intéresse à étudier l'ensemble des métabolites (petites molécules : sucres, acides aminés, acides gras, etc.) présents dans une cellule, un tissu, un organe ou un organisme à un temps donné et dans des conditions donnée.

f) La pharmacogénomique:

La pharmacogénomique est l'application d'approches génomiques et des technologies à l'identification de cibles thérapeutiques, a pour objectif l'étude des effets des médicaments qui pris par le patient sur leur génome.

g) La pharmacogénétique :

Pour réduire les risques de mauvaise réception d'un médicament (toxicité et d'effets secondaires), la pharmacogénétique permet aux médecins de prescrire des traitements en fonction de la carte génétique du patient. Cette science permet d'étudier l'influence du patrimoine génétique de l'individu sur le sort des médicaments.

h) Chemoinformatique :

Est un nouveau domaine, il concerne le développement, la création, l'organisation, le stockage, la diffusion, l'analyse, la visualisation et l'utilisation de l'information chimique. C'est une discipline scientifique représente l'interface entre la chimie et l'informatique.

i) La médecine personnalisée :

La médecine personnalisée est une approche récente et innovante fondée sur la connaissance des bases moléculaires des maladies. Elle consiste à choisir le traitement le plus adapté en fonction du profil biologique du patient et en fonction des caractéristiques moléculaires de sa maladie. Elle est liée directement au concept de biomarqueurs, grâce aux biomarqueurs nous pouvons prédire l'efficacité ou la toxicité d'un traitement avant qu'il ne soit prescrit.

## II.7. Les domaines de recherches en bioinformatique:

Lors de ces dernières décennies, le volume connu de données biologique a grandi de manière exponentielle. Ce qui introduit plusieurs domaines de recherche en bioinformatique. Dans ce qui suit nous avons cité quelques-uns

### II.7.1. L'annotation du génome :

Annotation précise des séquences du génome est la première étape importante dans l'analyse du génome. En général, l'annotation consiste à fixer d'une manière exhaustive des

informations aux séquences biologiques. En effet, l'annotation d'un génome est l'attribution à chaque gène sa fonction au sein d'une cellule. Dont l'objectif est d'obtenir une liste de gènes associés à leur fonction pour un organisme donné de cette manière nous pouvons comprendre le fonctionnement d'un organisme dans sa globalité. Dans ce contexte, la bioinformatique vise à annoter tous les génomes séquencés soit le génome humain ou les génomes des autres individus **[25]**.

### II.7.2. La prédiction de structure des protéines :

La séquence d'acides aminés d'une protéine (dite structure primaire) peut être facilement déterminée à partir de la séquence de gènes codantes pour celle-ci. La connaissance de la structure d'une protéine est essentielle pour comprendre la fonction de la protéine (c'est la relation **séquence-> structure-> fonction**) **[4]**. Par manque de meilleurs termes, l'information structurale est généralement classée comme structure secondaire, tertiaire et quaternaire. La prédiction de la structure des protéines est l'une des plus importantes solutions pour la conception de médicaments et la conception de nouvelles enzymes. Une solution générale à ces prédictions reste un problème ouvert pour les bioinformaticiens **[65]**.

### II.7.3. L'analyse et la comparaison des séquences :

Des séquences de nucléotides proches ont très probablement des rôles proches. Par exemple, si la séquence d'un gène **A** ressemblé fortement à la séquence d'un gène B, alors séquence de la protéine codée par **A** sera très proche de la séquence de la protéine codée par **B**. Les deux protéines adopteront donc une structure tridimensionnelle proche et auront donc une fonction semblable. Des conclusions fonctionnelles peuvent donc être tirées suite à la comparaison de séquences nucléotidiques ou même protéiques. La comparaison est un outil de base en bioinformatique. On parle plus communément de l'alignement de séquences qui met en correspondance les parties communes à deux ou à plus de deux séquences. La **Figure II.5** présente un alignement entre deux séquences d'ADN. Des algorithmes spécifiques ont été conçus pour réaliser un alignement. Une problématique différente est celle de retrouver, lorsqu'une nouvelle séquence est identifiée, si elle « ressemble à quelque chose » qui serait déjà connu dans les banques de séquences. L'outil bioinformatique BLAST (Basic Local Alignment Search Tool) a été créé pour cela, il utilise des critères empiriques et mathématiques pour identifier des fragments semblables. Il est extrêmement utilisé en pratique pour tirer des conclusions sur une séquence par rapport aux connaissances déjà présentes dans les banques **[65]**.


**Figure II.5 :** Alignement entre les deux séquences d'ADN :**ACTGTCTACTGAATGT** et **ACGTAGGAACGT**. Les parties rouges identifient les zones communes **[69]**

### II.7.4. Interaction protéine-protéine :

Les interactions protéine-protéine sont impliquées dans la plupart des processus cellulaires exemple anticorps-antigène, hormone-récepteur. La connaissance du réseau d'interaction complet d'un organisme donné facilite la compréhension des processus biologiques tels que les voies de signalisation, les voies métaboliques ou les mécanismes de transcription. Par exemple, les maladies de Creutzfeld-Jacob et d'Alzeihmer sont dues à des interactions aberrantes entre protéines. Cela permet aussi de prédire les fonctions de protéines non annotées à partir des fonctions des protéines avec lesquelles elles interagissent. La question posée c'est qu'il est possible de prédire la structure d'un complexe (interactome) obtenu par l'interaction protéine-protéine à partir des structures des partenaires isolés **[55]**. Une variété de méthodes bioinformatique ont été développées pour faire face au problème d'interaction protéine-protéine, mais il semble qu'il y ait encore beaucoup de travail à faire dans ce domaine.

### II.7.5. L'analyse de données d'expression génique :

Les puces à ADN c'est une technique à haut débit qui permet d'obtenir l'expression des milliers des gènes simultanément. L'analyse de ces données permet d'extraire les gènes exprimés différentiellement qui sont classés comme des biomarqueurs et utilisés pour diagnostiquer les maladies. Una variété des techniques informatiques ont été utilisées pour analyser ces données, les plus connues sont les techniques d'apprentissage automatique : le clustering et la classification. Dans ce contexte les bioinformaticiens continuent de développer des approches spécialisés et des algorithmes pour gérer ces données **[65]**.

### II.7.6. Analyse de l'expression des protéines :

L'expression des protéines est l'un des meilleurs indices de l'activité des gènes actuelle puisque les protéines sont généralement catalyseurs finaux de l'activité des cellules. Les puces à protéine peuvent fournir un aperçu des protéines présentes dans un échantillon biologique. La bioinformatique est très impliquée dans l'interprétation des données (extraction des

connaissances à partir des données) de puces à protéine afin de découvrir des biomarqueurs liés à certaines maladies **[38]**.

### II.7.7. Modélisation des réseaux de régulation :

Les réseaux de régulation sont des réseaux décrivant les interactions entre gènes dans une cellule et qui jouent un rôle fondamental dans le contrôle du fonctionnement et du développement des organismes vivants. La modélisation de ces réseaux permet d'étudier la manière avec laquelle l'expression des gènes favorise ou inhibe l'expression d'autres gènes. C'est un domaine en plein essor en bioinformatique, où elle vise à développer des modèles informatique généralement des modèles de simulation avec deux objectifs : exprimé bien les interactions entre les gènes et la complexité doit être minimale.

### II.7.8. La conception / la découverte de médicaments (Drug Design):

Les progrès récents dans la conception de médicament conduits à créer de nouvelles technologies puissantes qui sont une composante importante dans la bioinformatique. L'une des principales difficultés dans l'utilisation de ces technologies est leur exigence d'une expertise interdisciplinaire dans les sciences physiques, les sciences de la vie, et /ou en informatique. L'identification des gènes anormaux qui causent les maladies après une expérience d'expression génique est au cœur du processus de découverte du médicament. Le but de la bioinformatique consiste à faciliter l'emploi de ces nouvelles technologies puissantes (ex. les puces à ADN) pour identifier les gènes anormaux dans le processus de découverte. Alors, la bioinformatique rend plus simple le tri de ces molécules et réduit donc fortement le temps de recherche et de mise au point des médicaments **[7]**.

### II.7.9. La phylogénie

Toutes les espèces sur terre subissent un changement lent de leurs traits héréditaires au cours d'évolution. Le but de la phylogénie est de comprendre les relations de parenté et de retracer l'historique évolutif d'un gène, d'une famille de gènes, d'une espèce ou de différentes espèces. Les arbres phylogénétiques sont une très bonne manière de schématiser et d'appréhender ces relations rapidement **[7]**. Le problème posé ici est comment construire ces arbres afin d'exprimer bien les relations de parenté. Il s'agit d'un problème très actif en bioinformatique où elle vise à utiliser l'ensemble de la bioinformation disponible sur les espèces, y compris les résultats d'alignements de séquences.

Dans le tableau suivant nous résumons quelques taches de la bioinformatique selon le type de la bioinformation disponible

| Type de données | Tâches bioinformatiques |
|---|---|
| **Séquences nucléiques** | Identification des introns et des exons<br>Prédiction de gènes |
| **Séquences protéiques** | Algorithmes de comparaison de séquences<br>Alignement multiple<br>Découverte de régions conservées<br>Recherche de motifs |
| **Structures macromoléculaires** | Prédiction de structures secondaires<br>Prédiction de structures tertiaires<br>Identification d'interactions moléculaires |
| **Génomes** | Analyse de la structure du génome, Phylogénie<br>Analyse de liaison génétique |
| **Expression des gènes** | Corrélation de l'expression des gènes<br>Recherche de gènes différentiellement exprimés |
| **Voies de signalisation métaboliques** | Simulation de réseaux |

**Table II.2 : quelques tâches de la bioinformatique**

## II.8. Les défis de la bioinformatique : Bioinformatique Prochain

Dans la section précédente nous avons synthétisé quelques domaines de recherche en bioinformatique, à partir de cette synthèse nous pouvons extraire quelques défis du domaine dans le futur **[28]** :

- *Le stockage intelligent de la bioinformation* : avec l'évolution très rapide dans la quantité des données biologiques obtenues par les technologies à haut débit, le défi majeur de la bioinformatique consiste à stocker ces données d'une manière efficace (intelligente) où l'utilisateur peut l'accéder et utiliser facilement.
- *Augmenter l'intégration des données et des méthodes* : l'avenir de la bioinformatique dépend à la capacité d'intégrer des méthodes informatiques, de simulation et de modélisation pour extraire des nouvelles informations ou pour prédire exactement ce qui se passe dans une cellule en temps réel. L'intégration d'une grande variété de sources de

données comme la génomique, la transcriptomique, la protéomique avec une grande variété des méthodes, va nous permettre par exemple d'utiliser les symptômes de la maladie pour prédire les mutations génétiques et *vice versa*.

- *La génomique comparative* : un autre défi en bioinformatique est la génomique comparative à grande échelle c'est-à-dire le développement des outils pratiques pour comparer les génomes entiers d'organismes afin d'augmenter le rythme de découverte en bioinformatique.
- *La modélisation et la visualisation des réseaux complexes de systèmes cellulaires* : qui peuvent être utilisées dans l'avenir pour prédire comment le système cellulaire réagit à un stress prévue ou non.
- *La numérisation des données phénotypique :* la bioinformatique consiste à convertir l'information biologique complexe à un modèle compréhensible pour l'ordinateur. Le problème de la numérisation des données phénotypiques d'une manière lisible par les ordinateurs offre des possibilités intéressantes pour les futurs bioinformaticiens.
- *Traiter et interpréter des nouveaux aspects biologique :* jusqu'à présent, la bioinformatique a été appliquée dans presque tous les domaines d'études biologiques, à partir de génotype[8] et jusqu'à le phénotype[9]. Les domaines les plus récents sont l'interactome, qui intègre des ensembles d'interactions protéine-protéine et localizome, qui décrit les localisations intracellulaires de protéines. Dans l'avenir, le but ultime de la bioinformatique sera l'intégration des bases de données biologiques et de ressources génomiques afin de développer une représentation informatique de cellules et d'organismes vivants par lequel n'importe quel aspect de la biologie peut être examiné et traité dans la bioinformatique.

[8] Ensemble des constituants génétiques d'un organisme, qu'ils soient exprimés ou non.

[9] Ensemble des caractères observables d'un individu. Le phénotype correspond à la réalisation du génotype (expression des gènes) mais aussi des effets du milieu, de l'environnement.

## II.9. Conclusion :

Ce chapitre est consacré à la bioinformatique, nous avons présenté un état de l'art de cette discipline. D'abord, nous avons exposé les origines de la bioinformatique avec une démonstration de l'apport de la biologie moléculaire dans cette discipline. La bioinformatique connues plusieurs définitions, dans la deuxième section nous avons présenté quelques-unes. Ensuite, dans les sections qui ont suivi nous avons détaillé des sources de la bioinformation et la manière de la stocker, ce que nous appelons les banques de données biologiques. La bioinformatique est en interaction avec plusieurs autres disciplines telles que la biologie moléculaire, l'informatique et les mathématiques, nous avons cité dans la sixième section les champs liés à ce domaine multidisciplinaires. Finalement, nous avons exposé quelques domaines de recherche qui sont en plein essor en bioinformatique pour finaliser par les grands défis du domaine qui représente le chemin futur. Au sein de ce très large éventail de recherche, nous nous sommes intéressés par le domaine de la découverte des biomarqueurs pour diagnostiquer les cancers. C'est le problème que nous détaillerons dans le chapitre suivant et que nous traiterons dans notre travail.

## III. 1. Introduction :

La bioinformatique se réfère à la conception, l'implémentation et l'application des technologies informatiques, des méthodes et des outils pour rendre les données «*omiques*» (c'est-à-dire, les données génomiques, transcriptomiques, protéomiques et métabolomiques) significatif. La découverte des nouvelles connaissances à partir de ces données est l'un des objectifs principaux de la bioinformatique. Dans la recherche biomédicale, ces connaissances connues sous le terme « *biomarqueurs* », les biomarqueurs jouent un rôle très important dans ce domaine à des niveaux différents nous citons le diagnostic et le pronostic des maladies comme le cancer. L'identification ou bien la découverte des biomarqueurs d'une certaine maladie nécessite l'utilisation des techniques à haut débit, avec ces techniques nous pouvons manipuler les données biologique existante sur la maladie (les échantillons à étudier, les conditions environnementales…etc.) afin d'extraire des données quantitatives (données omiques) qui sont utilisées à la suite pour découvrir la meilleure connaissance qui représente bien le biomarqueur. La technique la plus utilisée pour découvrir les biomarqueurs du cancer est la technique des puces à ADN. Cette technique a permis de mesurer simultanément sur une seule puce l'expression de centaines, voire de dizaines de milliers de gènes transcrits. La question posée, parmi ces milliers des gènes quels sont les gènes qui représentent le biomarqueur, nous pouvons dire le bon biomarqueur ? Le rôle d'un bioinformaticiens ici est de développer une approche analytique qui permet d'identifier ces biomarqueurs à partir d'une haute dimensionnalité de données. Dans ce contexte, nous présentons ce chapitre afin de clarifier le concept biomarqueur ainsi le bon biomarqueur et d'expliquer comment la technique des puces à ADN est utilisée pour le découvrir.

## III.2. Les biomarqueurs : Définition et classification

Les essais cliniques basés sur les biomarqueurs ont été appliquées depuis plus de cinquante ans, mais leurs applications potentielles pour la détection et la classification des maladies, la stratification des patients et la découverte de médicaments ont augmentés depuis le début du XXIe siècle **[35]**.

### III.2.1. Définition :

Le terme « biomarqueur » a émergé au cours des dernières années pour devenir de plus en plus commun de nos jours. La première question à se poser est donc la signification de ce terme générique.

*« Le biomarqueur défini comme une caractéristique qui est objectivement mesurée et évaluée comme un indicateur d'un processus biologique normale, pathologique ou réponse à une intervention thérapeutique »* **[12]**.

D'après cette définition, les biomarqueurs peuvent être divisés en trois types principaux **[35]**:

- *Type « 0 » :* biomarqueurs utilisés pour estimer l'émergence et le développent d'une maladie.
- *Type « 1 »* : biomarqueurs qui permettent de prédire les réponses aux interventions thérapeutiques.
- *Type « 2» :* biomarqueurs qui, en principe, pourraient être utilisés comme critères d'évaluation dans le cadre d'essais cliniques.

Selon leur application à la détection des maladies, trois classes principales de biomarqueurs peuvent être spécifiées: les biomarqueurs de criblage, de diagnostic et de pronostic **[35] :**

- *Les biomarqueurs de criblage*, sont utilisés pour prédire l'apparition éventuelle de la maladie chez les patients asymptomatiques.
- *Les biomarqueurs de diagnostic,* sont utilisés pour faire des prédictions sur des patients suspectés d'avoir la maladie.
- *Les biomarqueurs de pronostic,* sont utilisés pour prédire le résultat d'un patient souffrant d'une maladie.

Les biomarqueurs peuvent également être considérés comme des indicateurs de changements fonctionnels et structurels dans les organes et les cellules. Ces modifications peuvent être associées soit à des facteurs de causalité ou des conséquences d'événements normaux et pathologiques **[35]**.

En outre, les biomarqueurs peuvent être considérés comme des cibles thérapeutiques potentielles, par exemple lorsque leur rôle causal dans la maladie est démontré **[35].**

Les biomarqueurs peuvent être des gènes, des protéines, des peptides (morceaux de protéines) ou des métabolites dont leur niveau est en train de changer en cas de maladies et ils peuvent ensuite être utilisés pour déterminer le stade de la maladie chez un patient **[95].**

| La maladie | Les biomarqueurs |
|---|---|
| **Diabète de type 1 et 2** | Glucose, Fructosamine, Hémoglobine A1c, Evaluation rétinale, Mesures néphropatiques, Evaluation neuropathie périphérique |
| **Hypertension** | • Angiotensine-I et -II, Rénine plasmatique, Aldostérone, Activité ACE<br>• Pression du sang, Mesure de la fréquence cardiaque |
| **Asthme, maladie pulmonaire obstructive chronique, arthrite rhumatoïde** | • Cytokines, Leukotriènes, Chimiokine<br>• Tests des fonctions pulmonaires |
| **Cancer de la prostate** | PSA (prostate specific antigen) |
| **Cancer du pancréas, du côlon, du rectum, de l'estomac.** | gène **CA 19.9** |
| **Cancer digestifs, ovariens, thyroïdiens.** | gène **CA 72,4** |
| **Cancer du sein** | gène **CA 15,3** |
| **Maladie du foie** | ALT, gamma GT |

**Table III.1 : Exemple des maladies et leurs biomarqueurs**

Le terme biomarqueurs n'est pas un nouveau concept même s'il n'était pas utilisé sous cette dénomination et d'une manière fréquente dans le passé. Pour autant, en médecine clinique la mesure et le suivi de la température corporelle pour une infection ou la détermination de la glycémie pour le diabète de type 2[1] sont des exemples simples qui sont utilisés depuis longtemps et qui répondent à la définition des biomarqueurs.

[1] Représente près de 90% des cas de diabète dans le monde. Egalement appelé « diabète non insulino-dépendant», « diabète gras » ou « diabète de l'adulte », ce type de maladie résulte généralement d'une surcharge pondérale et d'un manque d'exercice physique, et touche plus particulièrement les adultes, bien que les enfants soient de plus en plus atteints.

Par la diversité des types de biomarqueurs possibles comme nous avons déjà cité, le National Institute of Health (USA) a également propose une classification des biomarqueurs comme suit **[12]**:

| Dénomination | Définition |
|---|---|
| Biomarqueur | Caractéristique biologique mesurée de façon objective et évaluée comme un indicateur soit d'un processus biologique normal ou pathologique. |
| Biomarqueur de type 0 | Marqueur biologique de la progression de la maladie relie à un paramètre clinique connu. |
| Biomarqueur de type I | Marqueur biologique qui reflète les effets d'une thérapeutique selon son mécanisme d'action. |
| Biomarqueur de type II | Marqueur biologique considère comme un critère de substitution : une modification de ce biomarqueur est associée à un bénéfice clinique ou à un risque. |
| Critère de substitution [ou ≪ Surrogate endpoint ≫] | Catégorie de marqueurs destines à se substituer à un critère d'évaluation clinique devant permettre de déterminer le bénéfice clinique ou le risque à partir de données épidémiologiques, thérapeutiques ou physiopathologiques. |
| Critère d'évaluation clinique [ou ≪ Clinical endpoint ≫] | Caractéristique ou variable qui reflète l'état du patient. |
| Marqueur pronostic | Marqueur permettant de différencier des catégories de patients à différents risques pour une évolution déterminée, indépendamment du choix du traitement administre (ou du choix de ne pas administrer de traitement). |
| Marqueur prédictif | Marqueur permettant de prévoir les éventuels bénéfices (efficacité) et risques (toxicité) d'un traitement selon le statut du marqueur. |

**Table III.2: Définition des biomarqueurs selon le National Institute of Health [12]**

Selon la technique d'acquisition des données biologiques (*données moléculaires*) utilisée, les techniques de génomique, de transcriptomique, de protéomique ou de métabolomique. Nous pouvons classer les biomarqueurs comme suit :

- *Les biomarqueurs génomiques* fondés sur l'ADN, sont habituellement détectés par des techniques de séquençage ou hybridation génomique sur biopuce…etc.
- *Les biomarqueurs transcriptomiques* fondés sur l'ARN, sont détectés par des techniques de puce à ADN (visant à évaluer le niveau d'expression des gènes)…etc.
- *Les biomarqueurs protéomiques* fondés sur les protéines, sont mesurés par des techniques des puces à anticorps, des puces à protéines…etc.
- *Les biomarqueurs métabolomiques* pour les métabolites, sont détectés principalement par des techniques de RMN (résonance magnétique nucléaire).

### III.2.2. Les critères pour un bon biomarqueur :

Un biomarqueur de recherche doit répondre à un certain nombre de critères qui sont de deux ordres : biologiques et statistiques **[108]** :

***a) Plausibilité biologique*** :

- Etre mis en évidence d'un point de vue épidémiologique comme un élément marqueur de la maladie.
- Etre un élément causal des mécanismes pathologiques.
- Les modifications de sa mesure doivent être en lien avec l'évolution de la maladie.
- Il doit être mesurable et évaluable objectivement.
- Il doit conduire à la création de données sur la sécurité de son utilisation (la reproductibilité)

***b) Critères statistiques :***

- Les données statistiques ayant permis sa validation doivent être robustes.
- Variation corrélée avec l'évolution de la maladie.
- Sa variation ne doit pas être «masquée» par d'éventuels effets secondaires.

Il y a des autres critères pour garantir le succès dans les essais cliniques :

- Il doit pouvoir se substituer à un critère d'évaluation clinique.
- Il doit être prédictif de l'effet du traitement que cela soit en termes de bénéfices ou d'effets secondaires.
- Son effet ou rôle en tant que paramètre de substitution doit être comparable à celui qu'il aurait en présence d'un médicament de la même classe pharmacologique.

## III.2.3.Les biomarqueurs dans la cancérologie :

### *III.2.3.1. Description du cancer :*

Le cancer définis comme une perte de contrôle «accidentelle» de la régulation des cellules qui aboutit à leur prolifération anarchique. Il existe de nombreux types de cancer, mais quelque soit le type ils commencent tous par la croissance hors contrôle de cellules anormales **[62].**

L'origine du mot cancer est créditée au médecin grec Hippocrate, qui est considéré comme le «père de la médecine ». Hippocrate utilisé les termes *carcinos* et le *carcinome* pour décrire la non formation et la formation d'ulcère des tumeurs. En grec, ces mots font référence à un *crabe*, probablement appliquée à la maladie parce que la forme des propagations d'un cancer similaire la forme d'un crabe.

Le médecin romain Celsus, plus tard traduit le terme grec au terme *cancer*, le mot latin pour le crabe. Galen, un autre médecin romain, utilisé le mot *oncos* (Grec pour le gonflement) pour décrire les tumeurs. Bien que l'analogie de crabe d'Hippocrate et de Celsus est encore utilisée pour décrire les tumeurs malignes, le terme de Galen est maintenant utilisé comme une partie du nom pour les spécialistes du cancer « oncologues » **[62].** A partir de cette description nous pouvons noter les terminologies suivantes :

*a)* *La tumeur* : pathologique résultant de la multiplication excessive des cellules. Elle résulte d'un déséquilibre entre la mort des cellules et leur renouvellement et elle échappe aux systèmes de régulation contrôlant la division des cellules. La tumeur peut être maligne ou bénigne.

- Tumeur bénigne: tumeur sans gravité, d'évolution favorable.
- Tumeur maligne: tumeur grave, entraînant des symptômes anormaux.

*b)* *Le cancer* : tumeur maligne formée par la multiplication désordonnée des cellules d'un tissu ou d'un organe.

*c)* *La cancérologie :* équivalente à l'oncologie, c'est la science qui s'intéresse à étudier les cancers.

Durant les années 1970, les scientifiques ont découvert trois familles de gènes liés au cancer: les oncogènes, les suppresseurs de tumeurs et les gènes réparateurs **[62] :**

- *Oncogènes*: ces gènes provoquent la croissance de cellules hors de contrôle. Ils sont formés par des changements ou mutations de certains gènes normaux de la cellule appelés *proto-oncogènes*. Proto-oncogènes sont les gènes qui contrôlent la croissance et la division cellulaire. Actuellement, plus de cent oncogènes sont identifiés. Les plus connus sont les gènes *Ha-ras*, *myc*, ou *abl*.

- *Les gènes suppresseurs de tumeurs*: ce sont des gènes normaux qui ralentissent la division cellulaire(les freins). Lorsque les gènes suppresseurs de tumeurs ne fonctionnent pas correctement, les cellules peuvent se développer hors de contrôle ce qui conduit au cancer.
- *Les gènes de réparation* : qui sont capables de détecter et de réparer les lésions de l'ADN qui ont modifié les oncogènes ou les gènes suppresseurs de tumeur. Ces gènes de réparation sont également inactivés dans les cellules cancéreuses.

*« Le cancer se produit lorsque les oncogènes sont activés à un moment inapproprié, ou les gènes suppresseurs de tumeur et les gènes de réparation sont inactivés quand ils devraient entrer en action. Il en résulte une croissance excessive qui prend la forme de tumeurs ».*

### *III.2.3.2. Types du cancer :*

Il existe plus de cents différents types de cancer. La maladie peut toucher presque n'importe quel organe du corps, de la peau au côlon. On distingue quatre principaux groupes de cancers :

***a)*** ***Les carcinomes :*** sont les tumeurs qui prennent naissance dans le revêtement extérieur ou intérieur des organes internes (appelé tissu épithélial) et sur la surface extérieure du corps;

***b)*** ***Les leucémies :*** sont les cancers des éléments constituants du sang;

***c)*** **Les lymphomes :** sont les tumeurs qui se forment dans le système lymphatique;

***d)*** **Les sarcomes :** sont les tumeurs qui prennent naissance dans le tissu conjonctif, comme les muscles, les os et le cartilage.

### *III.2.3.3. Evolution du cancer :*

Le cancer c'est une maladie qui progresse d'une manière très rapide, cette progression est passée principalement par quatre stades :

***a)*** ***Stade1 :*** au début, des cellules normales se divisent plus rapidement qu'elles ne devraient et le nombre total de cellules augmente. Nous parlons alors d'une *hyperplasie*.

***b)*** ***Stade2*** **:** appelé *dysplasie*, les nouvelles cellules cancéreuses deviennent déformées. Elles constituent alors un amas croissant de cellules, appelé *tumeur primitive*.

***c)*** ***Stade3*** **:** la tumeur commence à pousser et à écraser les cellules voisines. À mesure qu'elle grossit, elle se creuse un chemin et envahit les cellules voisines, ce processus s'appelle *invasion*.

***d)*** ***Stade4 (final)*** : lorsque les cellules cancéreuses atteignent un vaisseau sanguin ou un ganglion, elles peuvent emprunter la circulation sanguine ou le liquide lymphatique pour

se rendre à d'autres parties du corps où elles recommencent à se diviser. Ce processus s'appelle *métastase*, ce qui signifie que le cancer s'est propagé à d'autres régions du corps.

#### *III.2.3.4. Les biomarqueurs en cancérologie :*

Lorsque nous parlons de biomarqueurs cancéreux, nous se réfère généralement aux protéines, gènes et autres molécules qui affectent la croissance, la multiplication et la lyse des cellules cancéreuses. Les biomarqueurs du cancer sont donc des outils précieux pour la détection de la maladie, le diagnostic, le pronostic et le choix du traitement **[2].**

La mise en place des biomarqueurs nécessite une compréhension globale des mécanismes moléculaires et processus cellulaire qui profondes l'initiation du cancer, en particulier en mettant l'accent sur la façon dont de petits changements dans seulement quelques gènes ou des protéines régulatrices peuvent perturber une variété de fonctions cellulaires et lancer le cancer.

Un défi majeur dans le diagnostic du cancer est d'abord de découvrir les biomarqueurs qui représentent bien cette maladie et d'établir la relation exacte entre ces biomarqueurs et la pathologie clinique, ainsi que, pour être en mesure de détecter des tumeurs non invasive à un stade précoce.

Les Biomarqueurs du cancer sont des traits quantifiables qui aident les oncologues lors de la première interaction avec les patients suspects à **[2]**:

- identifier les personnes à risque (Est-ce qu'il y a un risque pour développer un cancer ?),
- diagnostiquer à un stade précoce (y a-t-il un cancer ? Quel est le type de ce cancer ? Quel est le stade (à quel point il est évolué) de ce cancer ?),
- choisir la meilleure modalité de traitement (est ce que ce cancer répond à ce médicament ou non? Si non, essayer des autres traitements),
- surveiller la réponse au traitement (comment ce cancer réagit- il à ce traitement ?).

## III.3. La découverte des biomarqueurs :

La découverte de biomarqueur a un rôle important dans le diagnostic précoce de la maladie et le pronostic des résultats du traitement. Le succès du processus de découverte de biomarqueurs dépend de nombreux facteurs, y compris le phénotypage précis des échantillons biologiques, la pertinence des méthodes d'analyse qui produisent les données, l'exhaustivité des techniques de prétraitement qui extraient l'information et affiner les données analytiques brutes. Un autre facteur important est la qualité de la méthode d'analyse qui détecte le nombre

limité de composés (biomarqueur) qui ont le potentiel de discriminer les classes d'échantillons **[95].**

### III.3.1. Le Framework de découverte de biomarqueurs :

La découverte de biomarqueurs repose généralement sur l'idée que les espèces moléculaires (c'est à dire gènes, protéines, etc.) qui affichent les plus grands changements dans les phénotypes peuvent être signalées en tant que biomarqueurs potentiels.

L'approche traditionnelle utilisée pour la découverte de biomarqueurs consiste à analyser un seul gène ou protéine et l'identification de ses valeurs «anormales», basé sur des hypothèses biaisées vers des processus ou des voies biologiques spécifiques. En général, il existe trois méthodes traditionnelles d'identification des valeurs anormales de biomarqueur **[22]**:

1. Identification basée sur *des seuils de référence*, dans cette méthode, la distribution des valeurs de biomarqueur dans un groupe de référence est estimée et les valeurs anormales sont définies en utilisant des valeurs extrêmes en fonction de seuil centiles. Par exemple, la valeur de la concentration de protéine au-dessus de la valeur du 99% peut être considérée comme anormale et une indication d'une maladie.
2. Identification basée sur *des seuils de discrimination*, le seuil de discrimination peut être défini après la comparaison de la distribution des valeurs de biomarqueurs entre les groupes de patients (par exemple, groupe de contrôle contre groupe de maladie) en fonction de leurs différences ou les chevauchements. Par exemple, une valeur de concentration de protéines supérieure à 100 pg /mL peut être associé à une complication ou d'une maladie clinique spécifique. Un seuil de discrimination aurait pour but de maximiser la capacité de distinguer entre ces groupes.
3. L'approche basée sur *les seuils de risque*, cette approche vise à détecter des valeurs de biomarqueurs qui seraient associés à une augmentation du risque au-delà d'un point critique sur le suivi. Par exemple, une valeur de pression sanguine systolique au-dessous de 115 mmHg peut être définie comme «souhaitable», comme une valeur au-dessus de cette limite est liée à une augmentation du risque de maladie vasculaire.

Actuellement, la découverte de biomarqueurs peut être considérée comme un processus itératif incrémental comporte plusieurs étapes comme il est montré dans la **Figure III.1.**

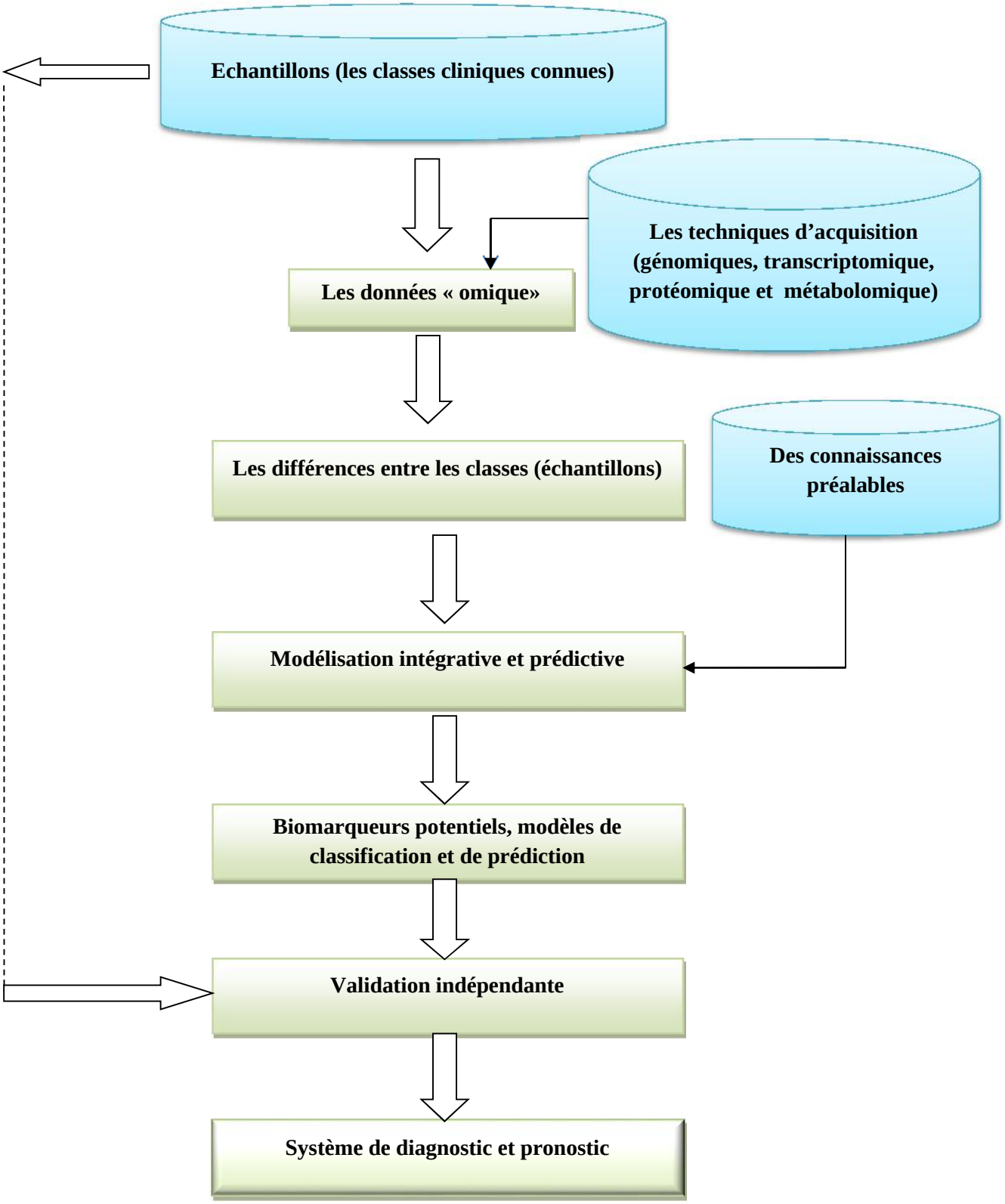



**Figure III.1: Le Framework de découverte de biomarqueurs [22].**

## III.3.2. Les puces à ADN pour la découverte de biomarqueurs et le diagnostic du cancer :

### III.3.2.1. Les puces à ADN :

C'est une nouvelle technologie qui permet d'évaluer le niveau relatif d'accumulation des transcrits de gènes dans une cellule, un tissu, un organe, un organisme ou encore un mélange complexe, à un moment donné et dans un état donné par rapport à un échantillon de référence. Elle regroupe un ensemble très divers de méthodes et de technologies comme les technologies d'acquisition des images, les méthodes d'analyse des images, les techniques d'amplification de la biologie moléculaire…etc.

#### *a) Historique:*

Les premières puces à ADN sont apparues en 1993, mais leur concept date de 1987. La technologie des puces à ADN est basée sur le principe d'hybridation développé par Southern en 1974. Ce principe stipule que deux fragments d'acides nucléiques complémentaires peuvent s'associer et se dissocier de façon réversible sous l'action de la chaleur et de la concentration saline du milieu **[78]**.

Historiquement les *macroarrays*, les *microarrays* et *les véritables puces à ADN* correspondent à trois méthodes différentes d'analyse **[78] :**

- *Les macroarrays* : utilisaient des clones d'ADN complémentaire (ADNc) disposés sur des membranes de nylon (avec un espacement de l'ordre du millimètre) en association avec des cibles radioactives.
- *Les microarrays* : plus miniaturisés, comportaient quelques milliers de gènes représentés par des produits PCR déposés tous les 200 à 400 microns sur une lame de verre et des cibles marquées par fluorescence.
- *Véritables puces à ADN* : associaient à chacun des gènes d'un organisme un ensemble d'oligonucléotides[2] synthétisés *in situ*[3]. La première de ces puces à ADN s'appelait la «Gene ChipTM HIV PRT». Commercialisée en 1998 par Affymetrix, elle avait été conçue pour l'analyse des mutations du virus HIV.

Aujourd'hui ces trois distinctions n'ont plus vraiment lieu d'être, d'autant plus que ces techniques sont utilisées de façon croisée, comme le montre l'exemple de puces à ADN utilisant des produits PCR et des cibles radioactives. Les terminologies *puce à ADN* et

[2] Petit segment d'ADN (quelques dizaines de nucléotides) simple brin.
[3] Signifie sur place ; elle est utilisée pour désigner une opération ou un phénomène observé à l'endroit où il se déroule (sur place).

*microarray* sont donc employées de façon indifférente. Les termes de *biopuce* ou *microréseau* sont également employés dans la littérature française **[78].**

Les puces à ADN sont utilisées dans plusieurs domaines tels que la microbiologie, la recherche et action des médicaments, environnement et agriculture, mais dans notre travail nous intéressons à leur application dans la découverte des biomarqueurs pour le diagnostic des maladies comme le cancer.

### *b) Principe :*

L'idée conceptuelle de la puce à ADN est très simple. Il s'agit de greffer sur une petite puce des fragments synthétiques d'ADN appelés *sondes* espacés de quelques micromètres et représentatifs de chacun des gènes étudiés. Ce micro dispositif (la puce) est ensuite mis au contact des acides nucléiques à analyser, au cours de l'étape d'hybridation. Ces acides nucléiques, appelés *cibles*, correspondent aux ARNm ou aux ADNc qui ont été préalablement couplés à un marqueur fluorescent ou radioactif[4]. Ce contact entre cibles et sondes conduit à la formation d'hybrides qualifiés par leurs coordonnées, et quantifiés grâce à la lecture des signaux radioactifs ou fluorescents **[71]**.

**i) Le support :**

Le support est une lame de microscope sur lesquels sont fixées les sondes, se présente sous la forme d'une surface (matrice) avec une taille de *25x75 mm*, plate ou poreuse (percées de puits) et composées de matériaux tels que le verre, les polymères, le silicium ou l'or et le platine. L'élément principal de la puce à ADN est l'unité d'hybridation « le spot » sur lequel sera greffée une sonde d'ADN synthétique. Les spots, présents en un grand nombre d'exemplaires et sont répartis régulièrement sur toutes la surface de la puce.

**ii) La préparation des sondes :**

Les sondes greffées ou synthétisées peuvent être de différentes tailles et types en fonction de la problématique biologique, des contraintes expérimentales et des moyens disponibles. La conception des puces à ADN utilisant les sondes de types ADNc, nécessite une étape d'obtention des sondes grâce à une amplification PCR à partir d'ADN génomique (génome connu). Le contrôle de ces produits d'amplification est important (taille de la séquence est de 100 à 500 bases) pour obtenir la bonne sonde. Les ADN amplifiés sont ensuite *dénaturé* pour obtenir des ADNc (le dépôt en simple) brin pour permettre par la suite une hybridation avec

[4] Consiste à utiliser les isotopes radioactifs comme l'iode 125 et le tritium pour marquer l'élément à étudier.

les cibles. L'ADNc complémentaire ne contient que les parties codantes (exons) et il est stable. Les sondes préparées sont déposées par un robot ou bien le *spotter* sur la puce, la couche de polymère qui recouvre la puce permettant la fixation des sondes par simples liaisons électrostatistiques, cette opération est réalisée à l'aide d'aiguilles creuses. Le diamètre des spots peut varier de 80 à 300 $\mu m$, et la distance entre deux dépôts consécutifs est de l'ordre de 250 $\mu m$ dans les deux directions **[71]**.

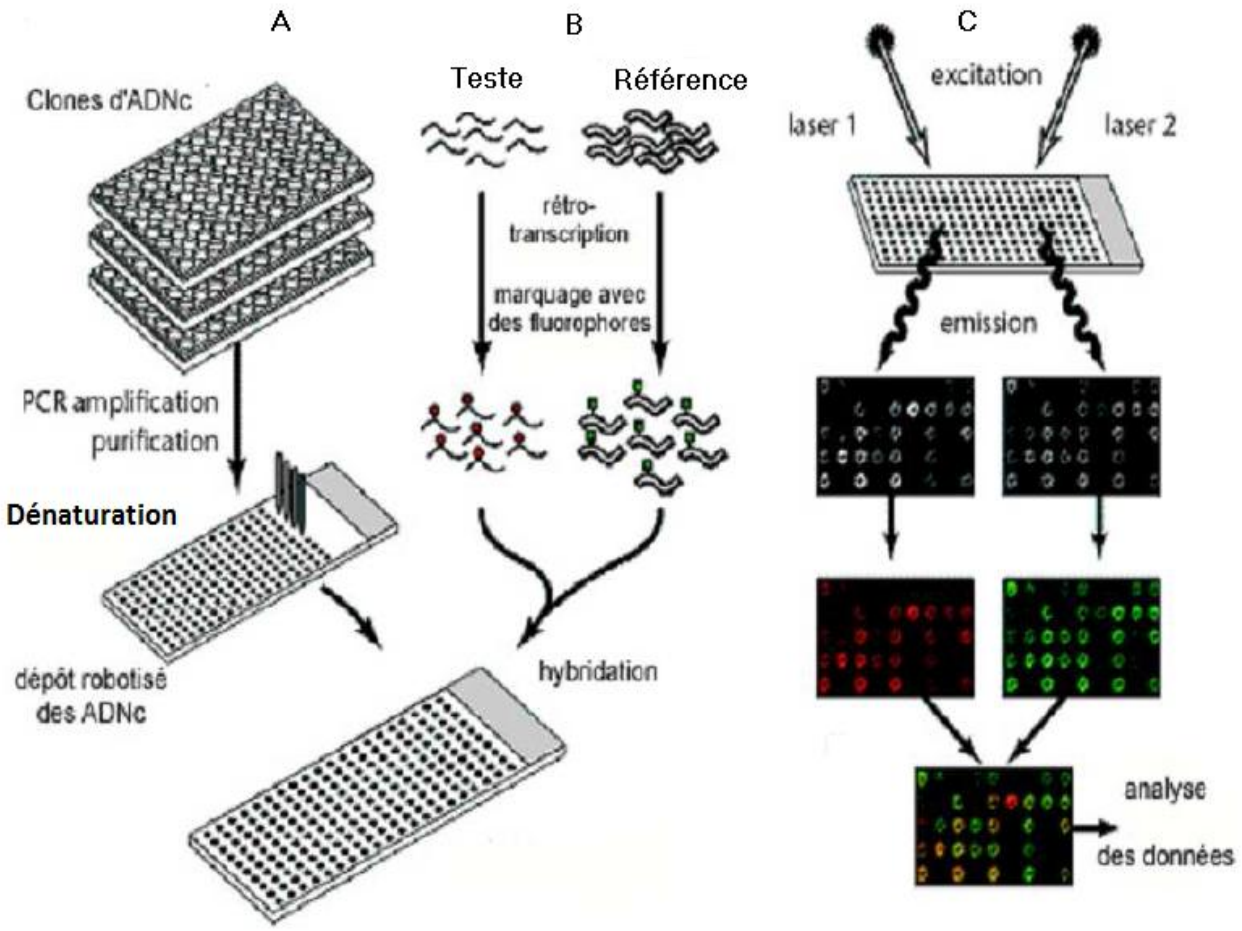



**Figure III.2 : Principe de la technologie des puces à ADN. (A)** Les séquences des sondes sont déterminées de façon à optimiser leur spécificité et leur sensibilité. Les sondes synthétisées sont déposées par un robot sur la surface de la lame selon un plan défini. **(B)** Les ARNm sont extraits des échantillons biologiques à comparer, rétro-transcrits et marqués avec deux fluorochromes différents puis mélangés avant hybridation. **(C)** La lecture des lames est réalisée par un scanner (microscope à fluorescence) couplé à un photomultiplicateur (PMT). L'image est alors analysée de façon à quantifier le signal. Les données sont ensuite normalisées, analysées et interprétées **[96]**.

Pour les puces à ADN qui utilise les sondes de type oligonucléotides, la synthèse de ce type des sondes est chimique c'est-à-dire ne nécessite pas des techniques de laboratoires (PCR). Les oligonucléotides sont synthétisés soit directement in situ (comme les Affymetrix : www.affymetrix.com) c'est-à-dire sur place sur la puce grâce à des technologies spécialisées ou avant le dépôt sur le support par l'aiguille creuse (la taille moyenne est de 25 à70 bases) **[22].**

Le résultat de cette étape est une puce pour lequel **[22]**:

- Chaque position sur la puce est connue.
- Chaque spot contient des fragments spécifiques d'un seul gène.
- L'ensemble des gènes du génome étudié doivent être représentés si possible sur la puce.
- Ces fragments fixés (sondes) sur la puce ne sont pas marqués.

**iii)La préparation des cibles et l'hybridation :**

Les cibles sont des échantillons à étudier. La préparation des cibles consiste à extraire les transcrits présents (ARNm) dans les cellules à partir de ces échantillons. Ensuite, les ARNm extraits sont rétro transcrits (la transcription inverse par RT-PCR) en ADNc. Selon la technologie de puce utilisée, les ADNc sont identifiées par un marquage radioactif ou fluorescent. Bien que moins sensibles que les marquages radioactifs, certains systèmes de marquages fluorescents présentent l'avantage de pouvoir identifier plusieurs cibles sur la même puce. Par exemple, un échantillon cancéreux peut être marqué par une cyanine rouge (Cy5) et un échantillon sain peut être marqué par une cyanine verte (Cy3) **[71]**.

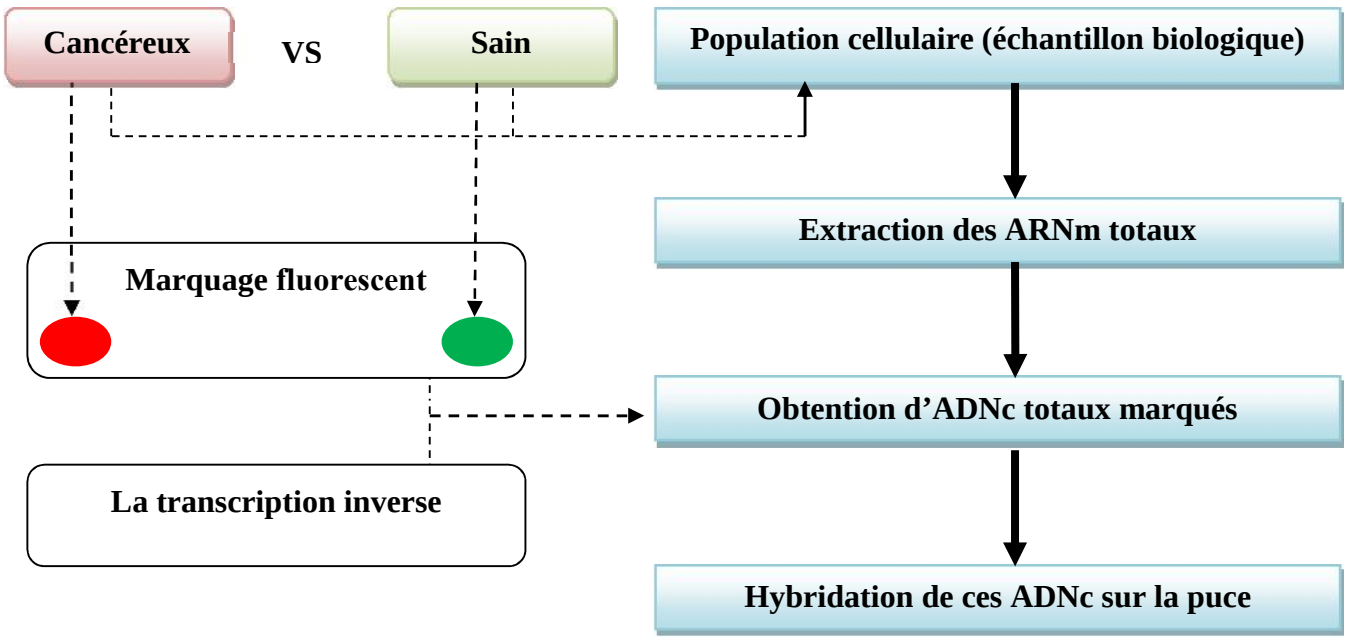



**Figure III.3 : Les étapes de la préparation des cibles à partir des échantillons cancéreux et sein**

Les ADNc marqués sont ensuite déposés sur la puce sur laquelle sont fixées des dizaines de milliers de sondes, recouvrant ainsi tous les gènes présents dans une cellule. Les ADNc vont alors s'hybrider sur ces sondes. Après plusieurs étapes de lavages, seules les séquences spécifiques fortement liées resteront hybridées. Les ADNc marqués à la fluorescence fixée à une sonde vont alors générer un signal dont l'intensité dépendra de la quantité d'ADNc, de la force de l'hybridation déterminée par le nombre de paires de bases, les conditions d'hybridation (comme la température) ainsi que les lavages. Le rapport ou bien le ratio des intensités obtenues pour chaque fluorochrome offre une comparaison directe des variations d'expression entre les deux échantillons **[96]**.

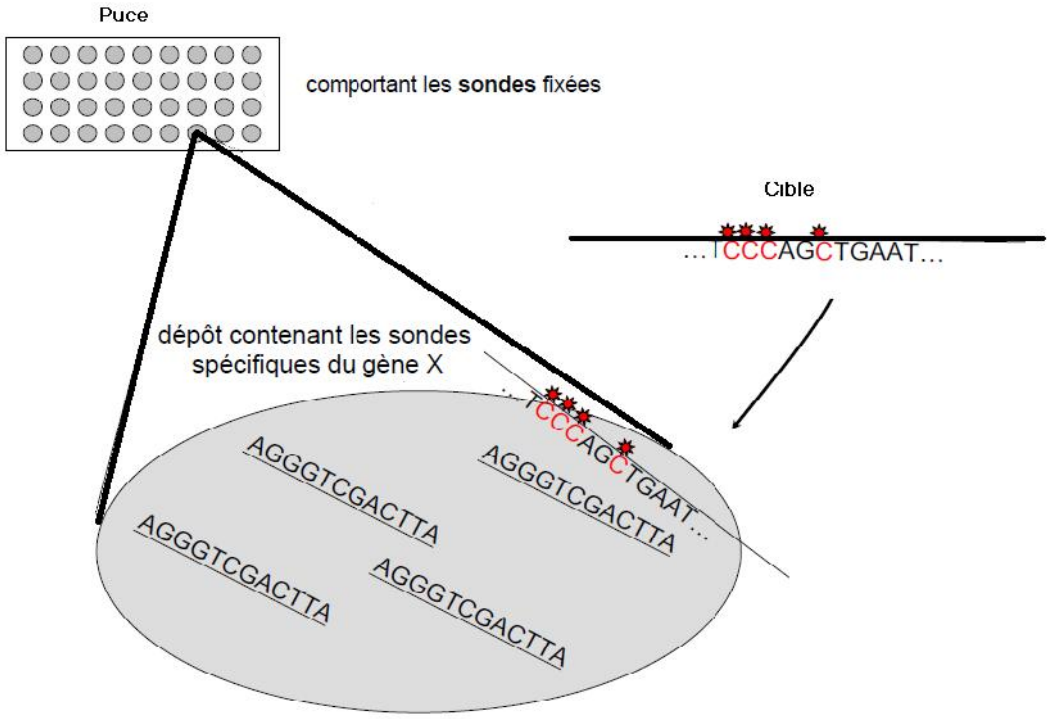



**Figure III.4 : Le principe d'hybridation dans les spots de la puce [71].**

**iv) La lecture des résultats d'hybridation :**

La lecture des résultats se fait grâce à un scanner. Dans le cas des technologies à fluorescence, son principe est celui d'un microscope confocal couplé à un ou plusieurs lasers. Chaque laser excite spécifiquement un fluorochromes. L'émission est amplifiée par un photomultiplicateur et transformée en signal digital c'est-à-dire en image. Chaque pixel de l'image scannée représente une mesure de fluorescence. Pour les puces à ADN deux couleurs, deux images en niveau de gris sont générées (une pour chaque fluorochromes). Ces images sont converties en fausses couleurs (allant généralement du vert au rouge) et superposées. Un ensemble des étapes d'analyse d'images permet d'extraire des informations qualitatives

(diamètre, niveau de saturation) et semi quantitatives (intensité du signal et du bruit de fond) pour chaque complexe sonde-cible (*spot*) dans chacun des fluorochromes. Des méthodes et outils bioinformatiques sont ensuite nécessaires pour analyser et extraire des connaissances.

**v) L'analyse de l'image :**

C'est une étape importante, elle a un impact considérable sur l'interprétation biologique des données. Dans le cas d'un double marquage, le but de cette analyse est de quantifier le niveau d'expression des gènes. Cette mesure basée sur rapport d'intensité entre les deux niveaux des fluorochromes détectés. L'analyse de l'image d'une expérience de puce à ADN comporte trois sous étapes :

- *La localisation des spots :*

Consiste à déterminer les coordonnées de chaque spot de la puce à l'aide d'une grille théorique définie lors du plan de dépôt des sondes qui doit être juste que possible. Pour localiser un spot sur une image, c'est-à-dire faire correspondre un modèle idéal de puce avec une image acquise, un nombre important de paramètres doit être estimé (espace entre les spots, espaces entre blocs d'une puce ….).

- *La segmentation des spots :*

Consiste à classer les pixels de l'image à deux classes *fond* et *signal*. Ceci sous-tend une analyse du signal au niveau de chaque spot et un découpage de l'image en différentes régions, chacune ayant des propriétés propres. A cette étape, la taille variable, la forme complexe et les irrégularités des spots compliquent la tâche de segmentation.

- *La quantification :*

Après l'identification des pixels *signal*. La quantification consiste à calculer l'intensité du signal (par des formules mathématiques et physiques). Ces intensités corresponds les niveaux d'expression des gènes disponibles sous la forme d'un seul nombre et qui représente l'abondance de gène transcrit dans un échantillon biologique « point de départ » **[22]**.

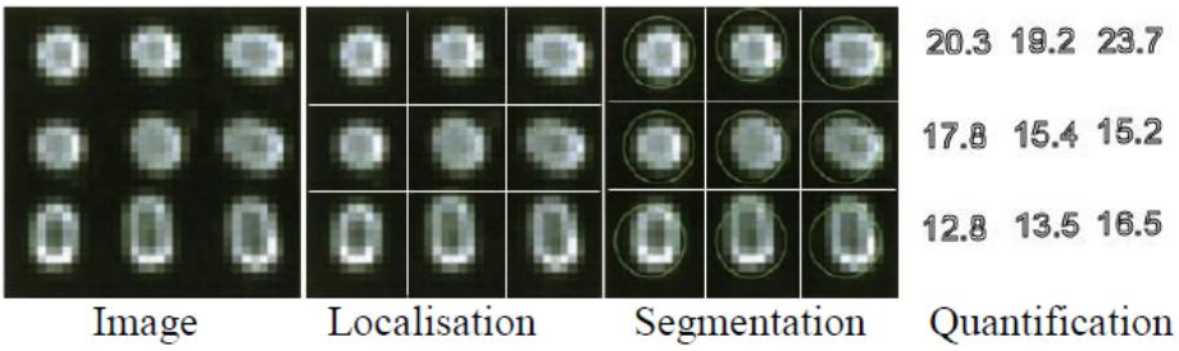



**Figure III.5 : Les étapes d'analyse d'une image d'une expérience de puce à ADN [60].**

Après la quantification, une étape de *prétraitement* est nécessaire pour éliminer certaines anomalies, cette étape englobe le *filtrage* et la *normalisation*. Le filtrage vise à éliminer les gènes qui ont une faible variance (non spécifiques) et la normalisation qui repose sur l'hypothèse fondamentale que la plupart des gènes ont le même niveau d'expression dans les échantillons, ce qui signifie que peu de gènes sont différentiellement exprimés. L'objectif principal de la normalisation est corrigé les différences systématiques entre les mesures sur la même puce qui ne représentent pas de véritables variations biologiques **[93].**

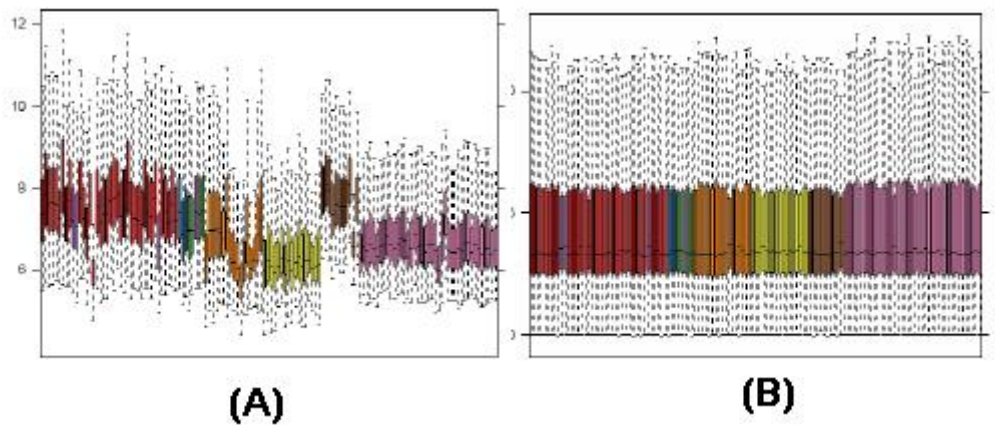



**Figure III.6 : La normalisation des données d'expression génique,** (A) : les données avant la normalisation, (B) les données après la normalisation **[96]**

Après l'étape de prétraitement (analyse bas niveau), nous obtenons une matrice numérique, les données de cette matrice nommées les données *d'expression des gènes* ou *d'expression génique* qui est utilisée à la suite par les bioinformaticiens pour établir une liste de gènes dont l'expression est significativement différente entre les échantillons, puis à les classer par famille et par niveau d'expression afin d'établir des hypothèses physiopathologiques et de répondre à des problématiques biologiques importantes comme la recherche de biomarqueurs **[96]**.

### III.3.2.2. L'analyse des données d'expression génique pour la découverte des biomarqueurs:

Les puces à ADN sont maintenant utilisées de plus en plus pour l'analyse à haut débit de données biologiques afin d'identifier et d'extraire des nouvelle connaissances comme les biomarqueurs. Les biomarqueurs dans une expérience de puce à ADN sont *les gènes exprimés différentiellement*. En termes statistiques, nous disons qu'un gène est exprimé de façon différentielle si la distribution des valeurs d'expression dans les échantillons sains

diffère de la distribution dans les échantillons de cancer dans une expérience de puce qui compare échantillons provenant de deux classes d'individus, sain et cancéreux. Nous disons qu'un gène rend un bon biomarqueur si un test permettant de distinguer les échantillons de cancer à partir d'échantillons sains basée sur l'expression de ce gène a une *sensibilité* et une *spécificité* élevées **[64] [60].**

Alors la découverte de biomarqueur dans le contexte d'analyse des données d'expression génique est l'identification d'un sous-ensemble optimal de gènes (exprimés différentiellement) qui se différencie de manière significative entre les classes et peut être utilisé pour la prédiction précise de l'appartenance d'un échantillon à une classe. Bien qu'il puisse arriver que biomarqueur compose d'une seule variable (un sel gènes), le plus souvent un bon biomarqueur signifie la recherche d'un sous ensemble de gènes, qui peut séparer les classes **[22]**. Les données d'une expérience de puce à ADN sont réorganisées dans des matrices $X_{m\,n}$ où les $m$ lignes sont les échantillons à étudier et les *n* colonnes sont les gènes exprimés, chaque $X_{ij}$ correspond le niveau d'expression de *l'ième* gène dans le *jème* échantillon. En associant à cette matrice un vecteur Y qui représente les classes des échantillons par exemple saine et cancéreux. Alors l'analyse pour découvrir les biomarqueur consiste à appliquer des algorithmes et des heuristiques sur cette matrice afin d'extraire le petit sous ensemble de gènes qui représente le bon biomarqueur.

En bioinformatique, l'analyse de ces données fait appel aux techniques d'apprentissage automatique soit pour une analyse supervisée ou non supervisée.

- Pour *l'analyse non supervisé* elle s'agit le clustering des données : regroupé les gènes dans des clusters différents avec une similarité maximale intra les clusters et minimale inter les clusters. Ces clusters peuvent utiliser à la suite dans une analyse supervisée.
- Pour *l'analyse supervisée*, c'est l'analyse appropriée pour la découverte de biomarqueurs, elle permet d'obtenir les biomarqueurs avec le modèle de classification approprié.

Dans la figure (**Figure III. 7. (A))** nous intéressons au chemin **A-B-D** qui est détaillé dans la partie **(B)** de la figure qui illustre la séquence des étapes de découverte de biomarqueurs. *La réduction de la dimensionnalité* vise à éliminer les gènes non informatifs seulement les gènes informatifs restants et qui corresponds les gènes exprimés différentiellement, *la construction du modèle de classification* en fonction de la méthode de sélection de caractéristique, les deux premières étapes peuvent être séparés ou couplés ; ensuite, *une étape de validation* est nécessaire afin d'évaluer les performances de ce modèle, il est préférable de faire cette validation sur un ensemble de test indépendant, la mise en œuvre du modèle de classification.

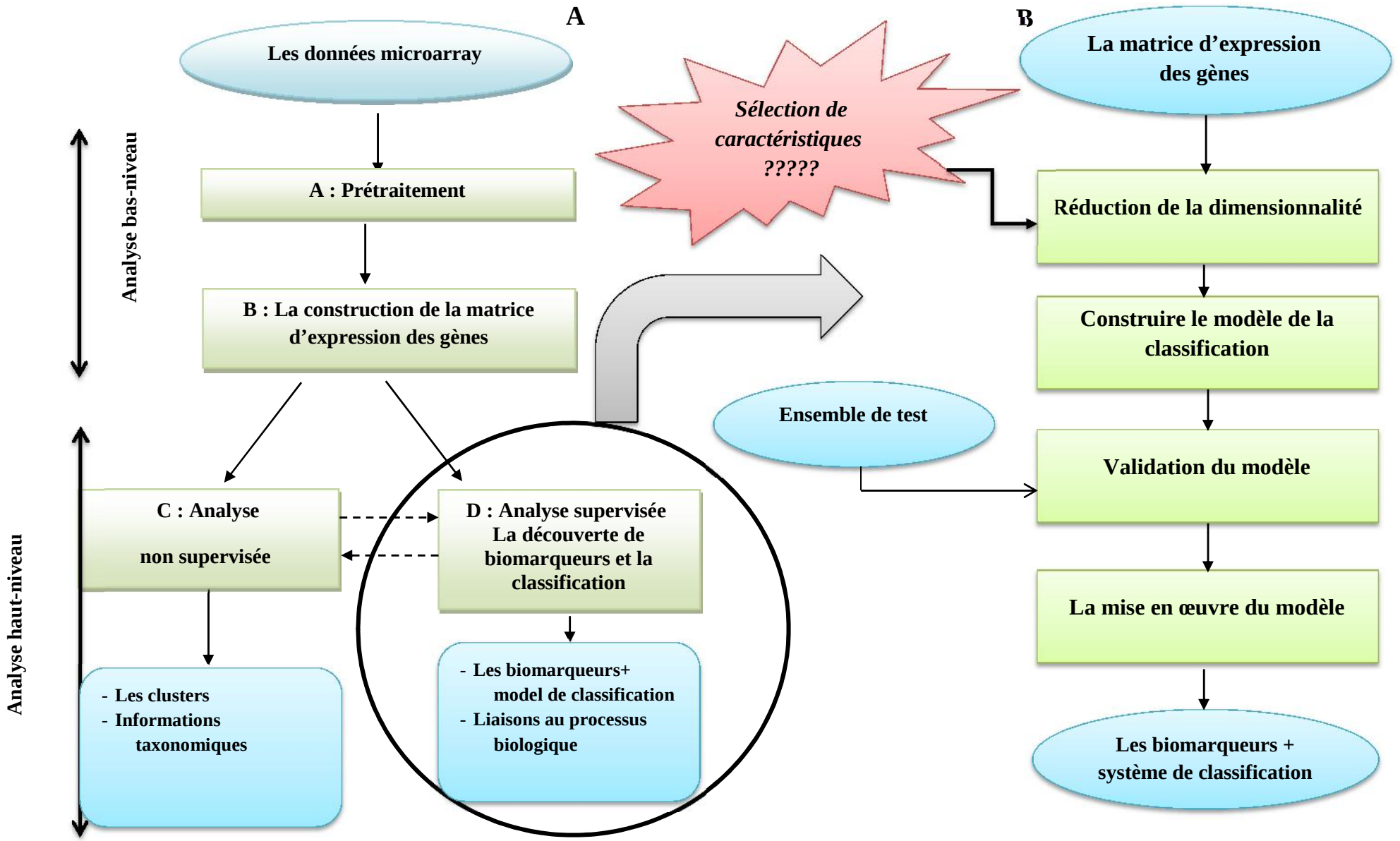



**Figure III. 7 : L'analyse des données des puce à ADN, (A) Les étapes d'analyse de données d'expression génique, (B) les étapes de découverte de biomarqueurs**

A la fin, nous pouvons dire que les principaux objectifs de ces étapes sont :

- Identifier les petits sous-ensembles de gènes qui peuvent être utilisés pour la classification efficace des nouveaux échantillons,
- fournir des classificateurs ou bien des modèles de classification rapides et économiques qui peuvent être facilement mises en œuvre dans la pratique clinique,
- relier les biomarqueurs avec des processus biologiques sous-jacents **[22].**

Une étape de *validation* d'un biomarqueur et le modèle de classification approprié sur *un ensemble de test* indépendant est suffisante pour son déploiement. La validation se fait même si nous ne comprenons pas la biologie sous-jacente de la maladie. Cette opinion est plus facile à accepter pour certains types de biomarqueurs, par exemple, ceux de diagnostic. Cependant, pour d'autres types, tels que les biomarqueurs prédictifs, il est nécessaire de comprendre la relation entre le biomarqueur et les mécanismes de la maladie **[8]**.

## III.4. Conclusion :

Ce chapitre est consacré à la découverte des biomarqueurs qui est à la base du diagnostic des maladies. Dans la première section de ce chapitre, nous avons expliqué le concept de biomarqueur en donnant leurs différents types selon plusieurs classifications avec une exposition des critères de validité qui permet de dire qu'un biomarqueur est bon ou non. Ainsi, nous avons montré le rôle de biomarqueur en cancérologie qui définit comme l'unité de base avec lequel nous pouvons diagnostiquer les cancers. Ensuite, dans la deuxième section nous avons décrit le Framework pour découvrir les biomarqueurs où nous avons exposé d'abord la méthode traditionnelle utilisée pour cette tâche. L'évolution des technologies à haut débit fournie un grand apport dans la tâche d'identification des biomarqueurs. Parmi ces technologies nous trouvons les puces à ADN où nous avons détaillée leurs principes et étapes de construction dans la deuxième section. Par la suite, nous avons parlé sur l'analyse des données d'expression génique qui peut être supervisée ou non supervisée et nous avons déduit que l'analyse supervisée est l'appropriée pour découvrir les biomarqueurs. En bioinformatique et d'un point de vue apprentissage automatique la découverte de biomarqueurs pour le diagnostic d'une maladie (ex. le cancer) équivalent à un problème de réduction de dimensionnalité avant une tâche de classification, le processus ou l'algorithme heuristique pour réaliser cette réduction et qui conduit à le sous ensemble optimal de gènes informatifs appelé *la sélection de caractéristiques*. Dans le chapitre suivant nous représenterons cette technique avec leurs principales approches et comment elle est appliquée pour découvrir les biomarqueurs.

## IV. 1. Introduction :

L'analyse des données d'expression génique pour découvrir les biomarqueurs c'est une analyse supervisée qui consiste à réduire la dimensionnalité des données avant une tâche d'apprentissage (la classification). Avec cette réduction nous pouvons faire un *filtrage* de gènes c'est-à-dire toutes les gènes redondantes, non pertinentes et bruyante sont éliminées, seulement les gènes les plus informatives restants, d'un point de vue de biomarqueur sont les gènes exprimés différentiellement. L'étape de la classification qui suit la réduction vise à construire un modèle qui permet de classer et évaluer les gènes si elles discriminent bien entre les classes échantillons ou non, si oui cet ensemble de gènes représente le bon biomarqueur et le classificateur permet de classer efficacement les nouveaux échantillons. Les méthodes de réduction de dimensionnalité sont deux types, la sélection de caractéristique et la transformation de données. Dans notre travail nous utiliserons la première, car la deuxième technique conduit à une modification des données initiales qui peut conduire à des résultats complètement différents et insuffisants en raison de la sensibilité de ce type de données. Dans ce chapitre, nous présenterons les techniques utilisées pour la découverte de biomarqueur à savoir la sélection de caractéristique et les techniques d'apprentissage automatique. Nous considérons ici le gène comme une caractéristique, les étiquettes des classes sont les étiquettes des échantillons et les exemples sont les échantillons biologiques.

## IV.2. La sélection de caractéristiques :

Au cours de la dernière décennie, la motivation pour l'application des techniques de sélection de caractéristiques en bioinformatique est passée d'un exemple illustratif de devenir un prérequis indispensable à la construction du modèle pour la classification des maladies. En particulier, la nature de haute dimension de nombreuses tâches de modélisation en bioinformatique, allant de l'analyse de séquence sur l'analyse des micro réseaux à l'analyses des données d'expression génique, la littérature a donnée naissance à une multitude de techniques de sélection de caractéristiques qui sont présentées dans le domaine.

### IV.2.1. La motivation :

Un problème fondamental de l'apprentissage automatique est d'approximer la relation fonctionnelle $f()$ entre une entrée $X = \{x_1, x_2, \dots, x_n\}$ et la sortie $Y$, basé sur une mémoire de points de données (des exemples), $\{x_i, y_i\}, i = 1 \dots m$ où les $x_i$ sont des vecteurs de nombres réels $x_i = \{x_{i1}, x_{i2}, \dots, x_{in}\}$ et les $y_i$ sont des nombres réels. Parfois, la sortie Y n'est pas

déterminée par tout l'ensemble des entités en entrée X. Au contraire, il est décidé seulement par un sous-ensemble d'entre eux $\{x_{(1)}, x_{(2)}, \dots, x_{(m')}\}$ où $m' < m$ . Avec la disponibilité de données et du temps, il est bon d'utiliser toutes les caractéristiques d'entrée, y compris les caractéristiques non pertinentes, pour approximer la fonction sous-jacente entre l'entrée et la sortie. Mais dans la pratique, il y a deux problèmes qui peuvent être évoqués par les caractéristiques non pertinentes impliquées dans le processus d'apprentissage :

- Les caractéristiques non pertinentes en entrée vont induire des coûts plus élevés de calcul.
- Les caractéristiques non pertinentes en entrée peuvent conduire à un *surapprentissage*.

Par exemple, dans le domaine du diagnostic médical, notre but est de déduire la relation entre les symptômes et leur diagnostic correspondant. Si par erreur on inclut le numéro d'identification du patient comme une caractéristique d'entrée, un processus d'apprentissage automatique peut arriver à la conclusion que la maladie est déterminée par le numéro d'identification.

A cette raison une diminution de l'espace d'entré par une technique de sélection de caractéristiques avant un processus d'apprentissage est très importante pour éviter ces problèmes. La sélection de caractéristique permet d'obtenir seulement les données les plus informatives liées au problème étudié (le nombre de caractéristiques obtenu inférieur au nombre des échantillons).

### IV.2.2. Définition :

La sélection de caractéristique (en anglais Feature Selection) est un sujet important dans le data mining[1] et l'apprentissage automatique. Pour la classification, l'objectif de la sélection de caractéristiques est de sélectionner un sous-ensemble de caractéristiques pertinentes pour construire des modèles de classification et de prédiction efficaces où la pertinence dépend toujours des objectifs et critères du problème à résoudre **[47]**. Par la suppression de caractéristiques inutiles et redondantes, la sélection de caractéristiques peut améliorer la performance des modèles de prédiction par la réduction de l'effet de la malédiction de dimensionnalité, l'amélioration des performances de généralisation, accélérer le processus

[1] Désigne l'ensemble des techniques et méthodes dans les domaines des statistiques, des mathématiques et de l'informatique qui permettent de sortir d'un grand volume de données, des connaissances précises sur des éléments inconnus auparavant. Cette technique permet d'analyser et d'interpréter des données volumineuses, contenues dans une ou plusieurs bases de données afin de dégager des tendances.

d'apprentissage et améliorer le modèle d'interprétation **[56]**. Nous pouvons définir le problème de la sélection de caractéristiques comme suit :

- soit **G** = $\{g_1, g_2, \dots g_n\}$ un ensemble de caractéristiques de taille *n* où *n* représente le nombre total de caractéristiques étudiées.
- Soit **F** une fonction qui permet d'évaluer un sous-ensemble de caractéristiques.
- Nous supposons que la plus petite valeur de **F** soit obtenue pour le meilleur sous-ensemble de caractéristiques. Donc, l'objectif de la sélection est de trouver un sous-ensemble $G'(G' \subseteq G)$ de taille $n'(n' \leq n)$ tel que **: [46]**

$$F(G') = \min F(Z) / Z \dot{c} G$$

Où $|\mathbf{Z}| = \mathbf{n}'$. Le sous ensemble sélectionnée de taille *n'* nommé dans le domaine biomédical la *signature biologique* où *n'* soit donné par l'utilisateur ou déterminé à la fin de la sélection immédiatement.

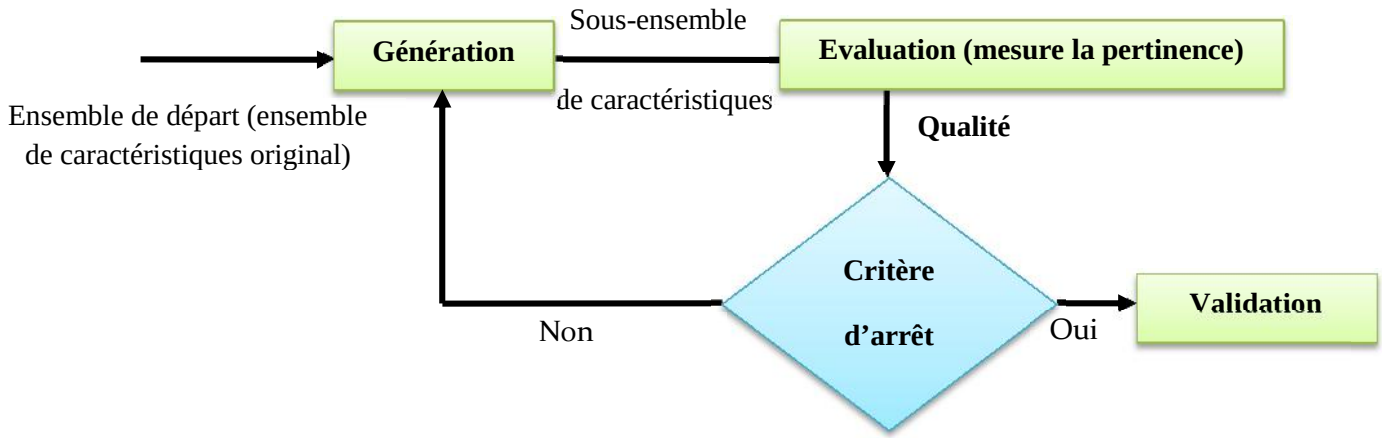



**Figure IV.1 : Schéma générale d'un algorithme de sélection de caractéristiques [68].**

La figure (**Figure IV.1)** montre qu'un algorithme de sélection de caractéristiques est itératif et il est composé de quatre éléments essentiels **[46]**:

1. **Génération** : vise à sélectionner le sous ensemble de caractéristiques candidates.
2. **Evaluation** : calculer la valeur de pertinence du sous-ensemble sélectionné, l'évaluation peut être indépendante à la génération (l'approche filter), dépendante complètement à la sélection (l'approche wrapper) ou semi-dépendante à la sélection (l'approche embedded).
3. **Le critère d'arrêt** : quand arrêter la recherche dans l'espace de caractéristiques? Certains algorithmes terminés son exécution, tandis que d'autres ont besoin d'un critère d'arrêt doit veiller à ce que la solution trouvée est une bonne ou non. Le critère d'arrêt détermine

également la taille de l'ensemble de caractéristiques sélectionné ce qui est une question délicate.

4. **Validation** : consiste à vérifier la validité du sous-ensemble par le test sur un ensemble de test indépendant.

### IV.2.3. Quelques notions liées à la sélection de caractéristiques:

Les caractéristiques sélectionnées doivent satisfaire la propriété de la *pertinence*. Il existe plusieurs définitions pour cette notion, les plus connues sont celles présentées dans **[58] [59].** Selon ces définitions, une caractéristique est classée comme étant : de *forte pertinence, faible pertinence* et *aucune pertinence.*

- *Forte pertinence* : une caractéristique $gi$ est dite très pertinente si son absence entraine une détérioration significative de la performance du système de classification utilisée.
- *Faible pertinence* **:** une caractéristique $gi$ est dite peu pertinente si elle n'est pas très pertinente et s'il existe un sous-ensemble V tel que la performance de $V \cup \{gi\}$ soit significativement meilleure que la performance de V.
- *Aucune pertinence :* les caractéristiques qui ne sont ni « peu pertinentes » ni « très pertinentes » représentent les caractéristiques non pertinentes. Ces caractéristiques seront en général supprimées de l'ensemble de caractéristiques de départ.

Parmi plusieurs caractéristiques, nous pouvons dire qu'il existe des caractéristiques redondantes et bruyantes. Deux d'entre eux pourrait dégrader les performances de la classification. Une classification des différentes caractéristiques se compose de :

- *Caractéristiques pertinentes*: les caractéristiques qui, par eux-mêmes ou dans un sous-ensemble avec d'autres caractéristiques, ont des informations sur la classe.
- *Caractéristiques redondantes*: ceux qui peuvent être éliminées parce qu'il y a une autre caractéristique ou sous-ensemble de caractéristiques qui fournissent déjà les mêmes informations sur la classe.
- *Caractéristiques bruyants:* les caractéristiques qui ne sont pas redondants et ne dispose pas d'informations sur la classe.

### IV.2.4. Les approches de sélection de caractéristiques :

Les techniques de sélection de caractéristiques sont généralement de trois types Filter, wrapper et embedded mais avec l'apparition du problème de l'instabilité une nouvelle approche a été proposée nommée l'approche ensemble.

### IV.2.4.1. L'approche Filter :

L'approche Filter consiste à évaluer le pouvoir discriminant de caractéristiques basées uniquement sur les propriétés intrinsèques des données. En règle générale, ces méthodes évaluent un score de pertinence et un système de seuil est utilisé pour sélectionner les caractéristiques les mieux marquantes **[21].** Dans cette approche l'évaluation de caractéristiques se fait généralement indépendamment d'un classificateur. Alors, elle peut considérer comme une étape de prétraitement (filtrage) avant une phase d'apprentissage pour la classification**.** Saeys et al. **[89]** ont été parlé sur les avantages pratiques de ces approches en déclarant que : «*même si le sous-ensemble de caractéristiques n'est pas optimale, il peut être préférable en raison de leur évolutivité et de calcul statistique*. ». Par conséquent ces méthodes peuvent négliger les caractéristiques de pruneau qui, semblent sans importance mais qui peuvent expliquer les phénomènes étudiés, lorsqu'ils sont pris en considération avec les autres **[76].** Les approches Filter peuvent être **[103]**:

- *Univariées* où chaque élément est étudié indépendamment des autres caractéristiques ou,
- *multivariées* où chaque élément est étudié comme partie d'un groupe de caractéristiques multiples.

D'une manière générale, deux stratégies de filtrage peuvent être identifiées **[76]** :

- La première stratégie, considère le problème de sélection comme un problème de classement, nous choisissons les caractéristiques ayant le meilleur score tandis que le reste est rejeté connues comme des *méthodes de classement (ranking methods)*;
- la deuxième stratégie où les caractéristiques sont sélectionnées par l'optimisation d'une fonction de coût particulier qui est souvent défini comme un compromis entre le caractère informatif maximale et une redondance minimale à l'intérieur du sous-groupe sélectionné de caractéristiques sont les *méthodes de recherche spatiale (space search methods*)

#### *a) Quelques fonctions de score :*

Soit $\boldsymbol{X} = \{\boldsymbol{x_k} | \boldsymbol{x_k} = (\boldsymbol{x_{k1}}, \boldsymbol{x_{k2}}, \dots, \boldsymbol{x_{kn}}), \boldsymbol{k} = \boldsymbol{1}, \dots, \boldsymbol{m}\}$ un ensemble de m exemples d'apprentissage dans un espace de représentation comportant *n* caractéristiques. Soit $\boldsymbol{Y} = \{\boldsymbol{y_k}, \boldsymbol{k} = \boldsymbol{1}, \dots, \boldsymbol{m}\}$ où $y_k$ représente l'étiquette de la classe de l'exemple$\boldsymbol{x_k}$.

$\boldsymbol{x^i} = (\boldsymbol{x_{1i}}, \boldsymbol{x_{2i}}, \dots, \boldsymbol{x_{mi}})$ représente la *ième* caractéristique, nous pouvons calculer le *score* de cette caractéristique en utilisant un des critères d'évaluation suivants (nous citrons les critères les plus utilisés) **[46]**:

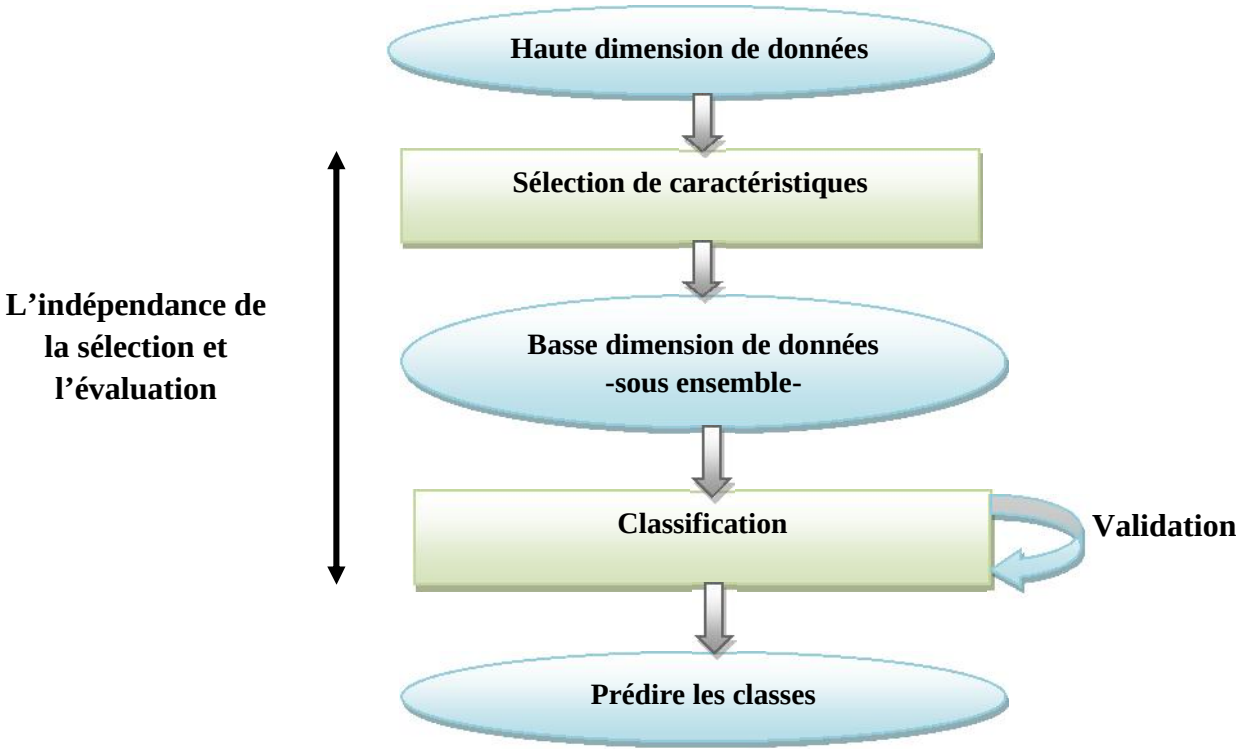



**Figure IV.2: Le principe de base de l'approche Filter**

***i) Le critère de la corrélation :***

En statistique et probabilité, le terme de corrélation est réservé pour désigner la liaison entre deux ou plusieurs variables quantitatives. Une mesure de cette corrélation est obtenue par le calcul du coefficient de corrélation linéaire. Il est compris entre -1 et 1 et il est estimé comme suit (dans le cas d'une classification binaire {0,1}):

$$C(i) = \frac{\sum_{k=1}^{m}(x_{ki} - u_i)(y_k - u_y)}{\sqrt{\sum_{k=1}^{m}(x_{ki} - u_i)^2 \sum_{k=1}^{m}(y_k - u_k)^2}}$$

Où $u_i$ et $u_k$ représentent respectivement les valeurs moyennes de la ième caractéristiques et des étiquettes de l'ensemble d'apprentissage.

***ii) L'information mutuelle :***

L'information mutuelle c'est une notion importante en théorie de l'information utilisée pour mesurer la dépendance (dépendance statistique) entre deux variables **[37] [85]**. Plus cette valeur est élevée plus les variables sont liées, quand elle est nulle, les variables sont indépendantes. Elle est estimé empiriquement par :

$$MI(x^i, y) = \sum_{x\,y} P(x^i, y) \log \frac{P(x^i, y)}{P(x^i)P(y)}$$

Où les probabilités $P(x^i), P(y)$ et $P(x^i, y)$ représentent respectivement les distributions marginales de $x^i$ et y et la probabilité conjointe.

***iii) Le critère de Fisher :***

C'est un critère qui permet de mesurer le degré de séparabilité des classes à l'aide d'une caractéristique donnée. Il est défini par :

$$F(i) = \frac{\sum_{c=1}^{C} n_c (u_c^i - u^i)^2}{\sum_{c=1}^{C} n_c (\sigma_c^i)^2}$$

Où $n_c$ , $u_c^i$ et $\sigma_c^i$ représentent respectivement l'effectif, la moyenne et l'écart type de la ième caractéristique au sein de la classe **c**. $u^i$ est la moyenne globale de la ième caractéristique.

***iv)* SNR (Signal-to-Noise Ratio) :**

Est une mesure utilisée en sciences et technologie qui compare le niveau d'un signal souhaité pour le niveau de bruit de fond. Il est défini comme le rapport de la puissance du signal à la puissance de bruit. Dans le contexte de sélection des caractéristiques est un score qui mesure le pouvoir de discrimination d'une caractéristique entre deux classes.

$$SNR(i) = \frac{(\mu_{Ci1} + \mu_{Ci2})}{(\sigma_{Ci1} + \sigma_{Ci2})} \quad [9]$$

Avec $u_c^i$ et $\sigma_c^i$ représentent respectivement la moyenne et l'écart type de la ième caractéristique au sein de la classe **c**, les classe sont 1 ou 2.

### *b) La méthode Max-relevance, Min-Redundancy (mRMR) :*

*Max-relevance, Min-Redundancy* (mRMR) est une méthode de filtrage qui vise à sélectionner les caractéristiques avec la plus grande pertinence et peu redondante avec la classe cible, elle est proposée par Peng et al. en 2005 **[85]**. En mRMR, la pertinence maximale *mR* et la redondance minimale *MR* de caractéristiques sont calculés en utilisant l'information mutuelle selon les deux formules suivantes :

$$Redondance(i) = \frac{1}{|n|^2} \sum_{i,j \in n} I(i,j) \qquad Pertinence(i) = \frac{1}{|n|^2} \sum_{i \in n} I(i,Y)$$

$|\mathbf{n}|$ : représente la taille de l'ensemble de caractéristiques.

I(i, j) : est l'information mutuelle entre la *i ème* et la *j ème* caractéristique.

I(i, Y) : est l'information mutuelle entre la *i ème* caractéristique et l'ensemble des étiquettes de la classe Y. Le score d'une caractéristique est la combinaison de ces deux facteurs:

$$Score(i) = Pertinence(i) - Redondance(i) \quad ou$$

$$Score(i) = \frac{Pertinence(i)}{Redondance(i)}$$

Plusieurs travaux montrent l'efficacité de cette méthode dans le domaine d'analyse des données d'expression génique **[68] [86] [92] [106] [122].**

### IV.2.4.2. L'approche Wrapper (approche enveloppante):

L'inconvénient principal des approches Filter est l'indépendance au classificateur. Pour résoudre ce problème, Kohavi et John ont introduit en 1997 **[67]** les approches wrapper. Elles sont basée sur la sélection de sous-ensemble de caractéristiques les plus discriminantes en réduisant au minimum l'erreur de prédiction d'un classificateur particulier alors, un appel de l'algorithme de classification est fait plusieurs fois c'est-à-dire à chaque sélection d'une caractéristique, nous calculons le taux de classification pour juger la pertinence d'une caractéristique. Cette approche est principalement critiquée en raison de leurs énormes besoins informatiques (la complexité). Plus que cela, il n'y a aucune garantie que la solution proposée sera optimale si un autre classificateur est utilisé pour la prédiction **[21].** Mais, elles sont généralement considérées comme étant meilleures que celles de filtrage. Elles sont capables de sélectionner des sous-ensembles de caractéristiques de petite taille qui sont performants pour le classificateur utilisé **[47]**.

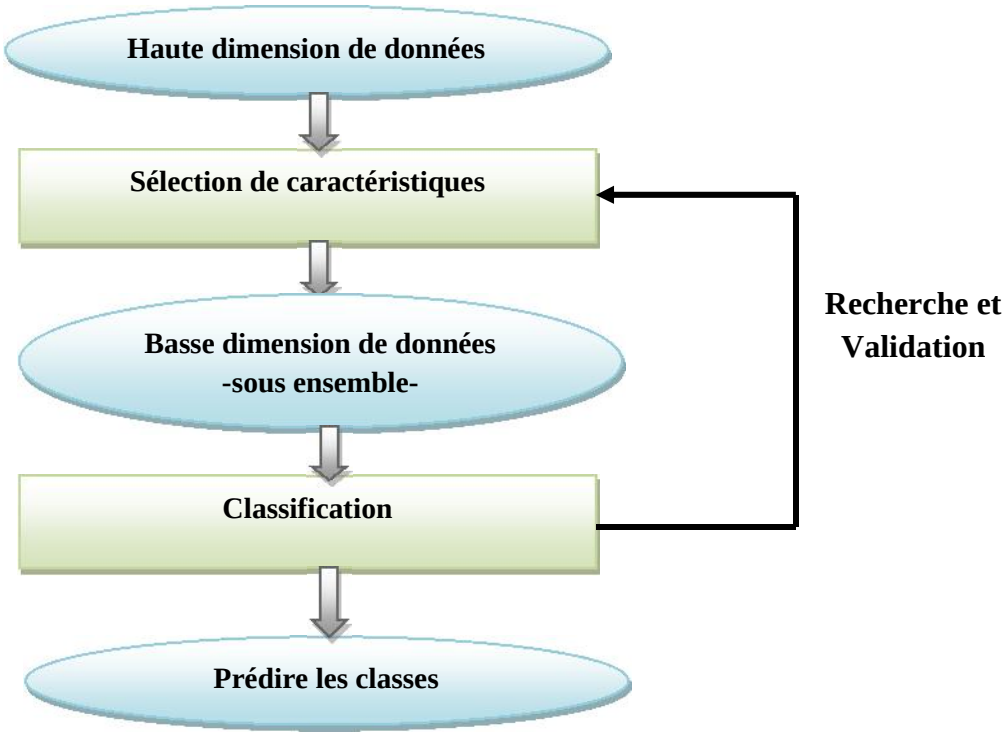



**Figure IV.3: Le principe de base de l'approche Wrapper**

### IV.2.4.3. L'approche Embedded (approche intégrée) :

L'approche Embedded possède aussi l'interaction avec l'algorithme d'apprentissage comme mais elle inclue la sélection de caractéristiques lors du processus d'apprentissage. Dans les approches wrapper, la base d'apprentissage est divisée en deux parties : une base d'apprentissage et une base de validation pour valider le sous-ensemble de caractéristiques sélectionné. Par conséquent, les approches Embedded peuvent se servir de tous les exemples d'apprentissage pour établir le système. Cela deux avantages principaux sont déduit :

- L'amélioration des résultats.
- La rapidité par rapport aux approches Wrapper parce qu'elles évitent que le classificateur recommence à zéro pour chaque sous-ensemble de caractéristiques**.**

Alors nous pouvons considérer les approches Embedded comme des approches *intermédiaires* qui englobent entre la rapidité et la liaison avec l'algorithme de classification. Une des méthodes de type embedded la plus utilisée dans l'analyse des données d'expression génique pour la découverte de biomarqueurs c'est la méthode *SVM-RFE* qui est un hybride entre la technique d'élimination récursive de caractéristique et le classificateur SVM.

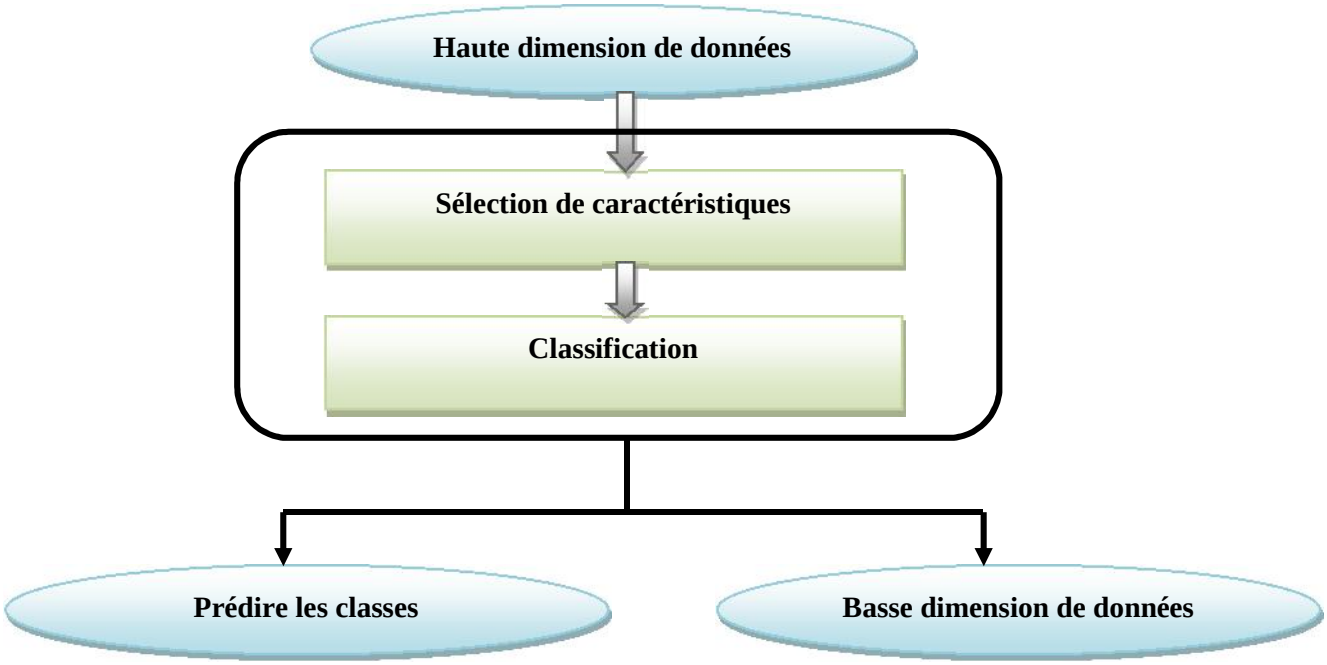



**Figure IV.4: Le principe de base de l'approche Embedded**

#### *a) L'élimination récursive des caractéristiques RFE:*

L'élimination récursive des caractéristiques en anglais Recursive Feature Elemination (RFE) correspond à l'élimination successive des caractéristiques apportant le moins à la qualité de la discrimination pour un classificateur donné. L'apprentissage est tout d'abord

réalisé avec l'ensemble des *n* caractéristiques, la variable la moins discriminante est supprimée, puis l'apprentissage est réalisé sur les n-1 caractéristiques restantes et ce processus est itéré jusqu'à obtenir le nombre de caractéristiques désiré. La complexité de la RFE dépend directement du classificateur choisi et du nombre de caractéristiques éliminées à chaque itération. En termes d'application à l'analyse des données d'expression génique, les SVMs linéaires semblent être une méthode bien adaptée pour être combinée à la RFE **[43].**

### *b) La méthode SVM-RFE :*

La SVM-RFE est une méthode de sélection de type embedded présentée en 2002 par Guyon ,I. et al.[**43**], elle est basée sur l'élimination récursive de caractéristiques et utilise le classificateur SVM pour évaluer et sélectionner le sous-ensemble de caractéristiques optimal non redondants **[112]**. L'équation du meilleur hyperplan séparateur dans le SVM est de la forme : $D(x) = w.x + b = 0$ où $w$ est un vecteur de poids, $b$ est une constante et $x$ correspond aux coordonnées d'un point dans l'espace de caractéristiques. Chaque variable $x_i$ est associée à un poids ($w_i$) qui détermine son pouvoir discriminant à chaque itération. Il est montré dans **[44]** que le coût de suppression de la *ième* caractéristique est de l'ordre de $w_i^2$ . La procédure de sélection est décrémentale et élimine donc progressivement les caractéristiques de faible poids. L'algorithme est comme suit :

**Répéter**

a) entrainer le classificateur SVM (linéaire) $D(x) = w.x + b$

b) calculer les $w_i^2$

c) supprimer la (les) caractéristique(s) correspondant au(x) poids le(s) plus faible(s).

d) Jusqu'à un critère d'arrêt est satisfait

**Figure IV.5 : L'algorithme de base de la méthode SVM-RFE [43].**

Cette méthode est très utilisée dans le domaine d'analyse des données d'expression génique où elle montre leur efficacité pour trouver les meuilleures solutions (biomarqueurs) **[86] [90] [100] [112]**. Au début la méthode SVM-RFE a été proposé pour une classification binaire ensuite une extention à une classifiction multi classes a été proposée **MSVM-RFE [109].**

## IV.2.5. La stabilité d'un algorithme de sélection de caractéristiques:

L'objectif principal d'un algorithme de sélection de caractéristiques est de produire une *signature robuste/stable* qui maximise les performances. La robustesse ou bien la stabilité d'un algorithme de sélection de caractéristiques peut définir comme : « un algorithme de sélection de caractéristiques est stable quand il y a des petits changements dans l'identité des échantillons d'apprentissage, les caractéristiques sélectionnées ne changent pas beaucoup » **[39] [48].**

Ces dernières années, la stabilité est devenue un sujet d'intérêt dans le domaine biologique où les gens du domaine cherchent toujours à valider encore leurs conclusions en appliquant les résultats (ex. biomarqueurs) découverts sur de nouvelles données. Plusieurs raisons qui pourraient conduire à réduire la stabilité d'une méthode de sélection de caractéristiques :

- Concevoir un algorithme pour sélectionner le nombre minimum de caractéristiques avec la plus grande précision de la classification sans prendre en considération la stabilité lors de la conception d'un tel algorithme.
- L'instabilité pourrait également être causée par la variance dans les données.
- *« petit échantillon - haute dimension »* est l'un des plus gros contributeurs à l'instabilité d'un algorithme de sélection.

Dans la littérature, juste un peu de travaux explorent la question de la robustesse dans le domaine d'analyse des données d'expression génique et la pluparts de ces travaux ont montrés la difficulté d'obtenir une signature reproductible avec un petit nombre d'échantillons **[23] [30] [34] [39] [90] [100].**

### IV.2.5. 1. Mesure de la stabilité :

Pour mesurer la stabilité d'un algorithme de sélection de caractéristiques, il faut avoir une procédure de test et une mesure de similarité.

#### *a) Les perturbations pour tester la stabilité :*

La mesure de la stabilité nécessite comme première étape une perturbation de l'ensemble de données, la perturbation consiste à supprimer des instances de façon aléatoire à partir d'un ensemble de données original afin de créer un ou plusieurs ensembles de données réduits. Ensuite une méthode de sélection de caractéristiques appliquée à chacun des différents sous-ensembles de données (tous les ensembles de données réduites et parfois l'ensemble de données original) et de créer une liste une signature pour chacun des sous-ensembles de données. Il existe plusieurs techniques de perturbation de l'ensemble de données, la plus simple consiste à sélectionner d'une manière aléatoire 80% des échantillons de la base initiale

pour construire la base perturbée. Il y a une autre technique qui est basée sur les chevauchements fixes des échantillons, elle est proposée par Wang et al. **[105]**. Cette technique consiste à générer deux ensembles de données de même taille à partir le jeu de données original avec une quantité contrôlée de chevauchements entre eux.

### *b) Mesure la similarité :*

Une mesure de la stabilité nécessite une mesure de la similarité entre les sous-ensembles de caractéristiques sélectionnés après l'application de l'algorithme de sélection sur les bases perturbées**.** Cela dépend évidemment du langage de représentation utilisé par l'algorithme de sélection de caractéristique donnée pour décrire ses préférences caractéristiques. Nous pouvons distinguer trois types de langages de représentation de préférences caractéristiques : Dans le premier type un poids ou un score est attribué à chaque entité indiquant son importance. Le deuxième type de représentation est une simplification de la première où, au lieu de poids, les rangs sont affectés à des caractéristiques. Le troisième type est constitué d'ensembles d'éléments sélectionnés dans lesquelles aucune pondération ou le classement est pris en considération **[103]**. Formellement, soit les exemples d'apprentissage sont décrits par un vecteur de caractéristiques $\mathbf{X} = (\mathbf{x_1}, \mathbf{x_2}, \dots, \mathbf{x_n})$ , puis un algorithme de sélection de caractéristiques produit soit **[23]:**

- Un poids-score : $w = (w_1, w_2, \dots, w_n), w \in W \subseteq R^n$
- Un classement : $r = (r_1, r_2, \dots, r_n), 1 \leq r_i \leq n$
- Ou sous ensemble de caractéristiques : $s = (s_1, s_2, \dots, s_n), s_i \in \{0,1\}$, avec 0 indique l'absence d'une caractéristique et 1 la présence d'une caractéristique.

- Pour mesurer la similarité entre deux poids w,w' produits par un algorithme de sélection de caractéristique donnée, nous utilisons le coefficient de corrélation Pearson

$$\mathbf{S_w(w, w')} = \frac{\sum_i (\mathbf{w_i} - \boldsymbol{\mu_w})(\mathbf{w'_i} - \boldsymbol{\mu'_w})}{\sqrt{\sum_i (\mathbf{w_i} - \boldsymbol{\mu_w})^2 \sum_i ((\mathbf{w'_i} - \boldsymbol{\mu'_w})^2}}$$

  Où $\boldsymbol{\mu}$ correspond la moyenne, $S_w$ prend les valeurs entre [-1,1] ; la valeur 1 signifie que les poids sont parfaitement corrélés, une valeur 0 signifie qu'il n'y a pas de corrélation, tandis qu'une valeur de -1 qu'ils sont anticorrélée.

- Pour mesurer la similarité entre deux classement r,r', nous utilisons le coefficient de Spearman de corrélation des rangs

$$S_R(r,r') = 1 - 6\sum_i \frac{(r_i - r'_i)^2}{n(n^2-1)}$$

Où $r_i$ et $r'_i$ sont les rangs de la caractéristique i dans les classements r et r' respectivement. Ici aussi, la gamme possible de valeurs est dans [-1,1]. Une valeur de 1 signifie que les deux classements sont identiques, une valeur de 0 qu'il n'y a pas de corrélation entre les deux classements, et une valeur de -1 qu'ils ont des ordres exactement inverses.

- Enfin, nous mesurons la similarité entre deux sous-ensembles de caractéristiques sélectionnés à l'aide d'une simple métrique de similarité comme l'indice de Dice, Tanimoto, Jaccard…etc.

### IV.2.5. 2. Méthodes pour améliorer la stabilité :

Plusieurs méthodes sont proposées pour améliorer la stabilité d'un algorithme de sélection de caractéristique et en général sont membres de l'une des deux catégories: les méthodes basées groupe (*Group Feature Selection*) et les méthodes basées ensemble (*Ensemble Feature Selection*) **[103].**

***a) Group Feature Selection :***

Le principe de cette approche consiste à assembler les caractéristiques fortement corrélées dans le même groupe et les caractéristiques qui ne sont pas corrélées dans des groupes séparés ensuite la sélection de caractéristiques est effectuée sur ces groupes résultants**.** L'idée derrière cette méthode est que, les groupes sont constitués d'éléments fortement corrélés alors ils auront la même pertinence et le choix d'une caractéristique par groupe traite le problème de la redondance. Ainsi, la liste finale de caractéristiques sera plus stable car elle est constituée seulement de caractéristiques les plus représentatives de chaque groupe **[103]**.

***b) Ensemble Feature Selection :***

L'idée de cette approche est inspirée de la méthodologie d'ensemble qui est appliquée dans l'apprentissage. Elle consiste à appliquer les algorithmes de sélection de caractéristiques plusieurs fois sur des sous ensemble de données perturbés par la technique d'échantillonnage et les résultats sont combinés dans une seule décision **[28] [53] [90] [100] [103]**. Bien que plusieurs résultats sont combinés, les caractéristiques qui sont souvent les plus performants se déplace vers le haut de la liste, tandis que ceux avec des performances faibles se déplace plus bas. Ainsi, la liste de caractéristiques finale sera plus stable. Il y a trois façons principales pour appliquer cette approche **[103] :**

- *Diversité de données* : consiste à appliquer une méthode de sélection de caractéristiques à des différentes versions échantillonnées du même ensemble de données.
- *La diversité fonctionnelle* : est réalisée par l'application d'un ensemble de différentes techniques de sélection de caractéristiques sur le même ensemble de données.
- *La façon hybride* : utilise deux d'entre eux, l'application de différentes techniques de sélection de caractéristiques à différentes versions échantillonnées.

Après l'application d'une des trois approches pour sélectionner les caractéristiques, la deuxième étape consiste à utiliser l'une des nombreuses fonctions d'agrégation disponibles pour combiner les résultats qui sont générés. Quelques exemples de méthodes d'agrégation comprennent l'agrégation exponentielle, l'agrégation moyenne et médiane, et l'agrégation basée seuil.

## IV.3 : L'apprentissage supervisé : la classification :

La classification est une technique d'apprentissage automatique où dans le jeu de données nous trouvons les étiquettes des exemples. Avant d'expliquer cette technique, il faut parler d'abord sur l'apprentissage automatique.

### IV.3.1. L'apprentissage automatique :

L'apprentissage cette notion englobe toute méthode permettant de construire un modèle de la réalité à partir de données, soit en améliorant un modèle partiel ou moins général, soit en créant complètement un nouveau modèle **[5].** L'apprentissage automatique (**machine learning** en anglais), qui est l'un des sous-domaines de l'intelligence artificielle. Il fait référence au *développement*, à *l'analyse* et à *l'implémentation* des algorithmes qui permettent à une machine d'évoluer grâce à un processus d'apprentissage, et ainsi de remplir des tâches qu'il est difficile ou impossible de remplir par des moyens algorithmiques plus classiques. Dont l'objectif d'extraire et d'exploiter automatiquement l'information présente dans un jeu de données. Selon la nature des données à étudier (si nous avons les sorties désirées ou non), il est possible de classer l'apprentissage automatique en: apprentissage supervisé vs apprentissage non supervisé **[84] :**

*a) Apprentissage supervisé :* dans ce type d'apprentissage, un expert est employé pour étiqueter correctement les exemples et l'apprenant doit alors trouver ou approximer un modèle qui permet d'affecter la bonne étiquette à ces exemples et il est capable de prédire

au mieux les étiquettes de nouveaux exemples. Quand les valeurs des sorties sont discrètes, donc nous avons dans la classification dans le cas contraire c'est *la régression*[2].

*b)* *Apprentissage non supervisé*: se distingue de l'apprentissage supervisé par le fait qu'il n'y a pas de sortie a priori (les étiquettes ou labels) dans le jeu de données, l'algorithme doit découvrir par lui-même ces sorties. Le clustering c'est une technique de l'apprentissage non supervisé**.**

Le *surapprentissage* ou *sur-ajustement* (en anglais *overfitting* ) est un problème très connu dans l'apprentissage automatique. En général, il arrive lorsque la base d'apprentissage comporte des données approximatives ou bruitées. Si nous obligeons le modèle de prédiction à répondre de façon quasi parfaite relativement à ces exemples, donc il devient biaisé par des données erronées et il perd sa souplesse et sa généralisation. Dans la classification, un bon moyen pour éviter le surapprentissage est la sélection de caractéristiques. Donc, le modèle créé à partir d'un nombre limité de caractéristiques sera plus général.

### IV.3.2. La classification :

La classification est la procédure qui permet d'identifier les classes appropriées pour des objets à partir de certains traits descriptifs. Une règle ou une procédure de classification est obtenue par un système d'apprentissage à partir d'un ensemble d'exemple, cette règle devra classifier correctement les exemples et capable de classifier correctement de nouveaux cas descriptif (exemple/échantillon). Dans la classification il faut avoir au moins deux classes prédéfinies.

La classification et la prédiction sont souvent utilisées dans la recherche biomédicale différemment que dans les statistiques. Dans les statistiques, elles sont associées à différents types de variables étudiées, continues pour la prédiction et catégoriques ou bien discrète pour la classification. En recherche biomédicale, ils sont souvent utilisés de manière indifféremment. Par exemple, l'affectation d'un échantillon biologique à une des classes différenciées peut être appelée résultat de prédiction; une estimation de la généralisation d'un système de classification peut être appelée l'exactitude de prédiction du système de classification **[22]**.

[2] Est un ensemble de méthodes statistiques très utilisées pour analyser la relation d'une variable par apport à un ou plusieurs autres variables. La régression d'une variable aléatoire y sur le vecteur de variables aléatoires x désignait la moyenne conditionnelle de y sachant x .

### IV.3.2.1. Les techniques d'évaluation d'un modèle de classification

Le processus d'apprentissage pour un problème de classification se compose classiquement de trois phases :

- Apprendre un modèle sur un jeu de données nommé *l'ensemble d'apprentissage*;
- Evaluation de ce modèle sur un jeu de données extrait du jeu de données initial nommé *l'ensemble de test*;
- Tester le modèle obtenu sur un jeu de données disjoint (nouvelles données).

L'étape de d'évaluation des performances d'un modèle appris nécessite l'utilisation d'un ensemble des données non utilisées pour l'apprentissage afin de ne pas biaiser les évaluations de performances, il s'agit de l'ensemble de test. Typiquement, dans notre contexte, il est nécessaire de faire appel à des méthodes d'évaluation pour répondre à des questions du type : *« tel biomarqueur est-il un indicateur pertinent concernant la probabilité de présence d'une pathologie chez tel groupe de sujets ?* ». La méthode d'évaluation la plus utilisée et particulièrement dans le domaine d'analyse des données d'expression génique nommée *la validation croisée* (ou cross validation en anglais) **[98]**, elle permet d'obtenir une estimation des performances du prédicteur, en exploitant la totalité du jeu de données. Ceci est obtenu en faisant plusieurs tests sur différents ensembles d'apprentissage et de test pour terminer par un calcul de la moyenne des résultats. Si nous obtenons une bonne précision en moyenne, ainsi qu'un écart-type faible, notre méthode de prédiction pourra être considérée comme bonne **[55]** :

#### a) La validation croisée : k-Fold CV:

Consiste à partitionner l'ensemble initial de données en *k* parties (folds) disjointes, d'apprendre sur l'union de *k-1* parties et d'évaluer les performances sur la partie non utilisée. Ce processus est itéré *k* fois, ainsi tous les exemples de l'ensemble initial de données auront été utilisés une fois en test et *k-1* fois en apprentissage. Le choix de la valeur de k dépend de la taille des données à analyser. Les valeurs classiquement utilisée sont 5 et 10. La performance réelle égale à la moyenne des performances estimées.

#### b) La validation croisée : Leave One Out (LOOCV) :

Est une généralisation de la validation-croisée k-fold avec k=n, n étant le nombre d'exemples. Une LOOCV consiste à construire un classificateur à partir de toutes les observations sauf une et le tester sur l'observation restante. Cette technique d'évaluation est préférable lorsque le nombre des échantillons est petit. Tandis que, les données sont trop

volumineuses, le recours à cette méthode n'est plus viable car le coût de calcul devient rapidement prohibitif (apprentissage de *n* modèles).

#### c) La validation croisée : Repeated Random SubSampling :

L'idée de base consiste à découper aléatoirement la population des données, classiquement en 2/3 pour l'ensemble d'apprentissage et 1/3 pour l'ensemble de test (qui est appelé 3-fold dans la littérature). Une fois la prédiction faite et les performances du prédicteur estimées, l'ensemble initial est reconstitué. Nous recommençons ensuite cette procédure plusieurs fois et nous calculons la moyenne des performances estimées.

### IV.3.2.2. Les critères d'évaluation :

Une fois le modèle prédictif (classificateur) construit, une évaluation de la performance est nécessaire. Cette évaluation consiste à mesurer, pour chaque étiquette, le nombre d'exemples correctement associés à cette étiquette, ainsi que le nombre d'exemples qui y sont incorrectement associés afin de mesurer le taux de bonne classification. Pour cela, les quantités suivantes ont été définies **[55]**:

- *VP (Vrais Positifs)* : si une prédiction a été faite positive et l'exemple montrait réellement un résultat positif.
- *FP (Faux Positifs)* : si une prédiction a été faite positive alors que l'exemple montrait réellement un résultat négatif.
- *FN (Faux Négatifs)* : si une prédiction a été faite négative alors que l'exemple montrait réellement un résultat positif (l'inverse de FP).
- *VN (Vrais Négatifs)* : si une prédiction a été faite négative et l'exemple montrait réellement un résultat négatif (l'inverse de VP).

Ces quantités sont regroupées dans une matrice appelé *la matrice de confusion*, illustrée dans le tableau suivant (**Table IV.1**) :

| | | **Réel** | |
|---|---|---|---|
| | | + | - |
| **Prédit** | + | VP (Vrais Positifs) | FP (Faux Positifs) |
| | - | FN (Faux Négatifs) | VN (Vrais Négatifs) |

**Table IV.1 : La matrice de confusion**

A partir de cette matrice, de nombreux critères d'évaluation peuvent être calculés, nous présentons quelques-uns que nous manipulerons dans notre travail **[55]** :

- ***L'accuracy ou l'exactitude:*** **Acc=VP+VN/VP+VN+FP+FN**. Qui permet d'évaluer la performance globale d'un classifier. Cette mesure représente le pourcentage de cas, positifs ou négatifs, correctement identifiés.
- ***La sensibilité :*** **Se=VP/VP+FN**. Définie comme le pourcentage de cas positifs correctement identifiés. Cette mesure est largement utilisée dans le domaine biomédical ;
- ***La spécificité :*** **Sp=VN/FP+VN**. Définie comme le pourcentage de cas négatifs correctement identifiés. Cette mesure est issue du domaine du traitement du signal, son utilisation dans le domaine biomédical est toujours associée à la sensibilité.

#### IV.3.2.3. Quelques techniques de classification :

Dans la littérature, il existe plusieurs techniques de classification. Dans ce qui suit, nous citerons quelques-uns que nous avons utilisés dans notre travail.

##### *a) Le k plus proche voisin :*

Le classificateur de k plus proches voisins (k-Nearest Neighbor ou k-NN en anglais) existé depuis les années cinquante, il est défini comme l'un des algorithmes de classification les plus simples **[102]**. Pour prédire la classe d'un nouvel objet, nous le comparons à ses voisins les plus proches par une mesure de distance **[54]**.

**Paramètre** : le nombre **k** de voisins
**Donnée** : un échantillon de m exemples et leurs classes
La classe d'un exemple *x* est *c(x)*
**Entrée** : un enregistrement *x'*
Déterminer les **k** plus proches exemples de *x'* en calculant les distances
Combiner les classes de ces **k** exemples en une classe **c**
**Sortie** : la classe de *x'* est c (*x'*)=**c**

**Figure IV.6 : L'algorithme de k-plus proche voisins [54].**

L'opérateur de distance le plus souvent utilisé est la distance euclidienne, cependant, en fonction du problème, nous pouvons également utiliser les distances de Hamming, de Mahalanobis…etc. Cette méthode ne nécessite aucun apprentissage pour un nouvel exemple se présentant, il suffit de calculer sa distance à tous les exemples présents dans l'ensemble

d'apprentissage et de garder les k premiers pour prendre une décision. Le choix de la valeur de k est très important et elle influe directement sur les résultats de la classification.

Comme il est facile à utiliser et à comprendre, il n'y a pas de connaissance préalable nécessaire sur les données, l'algorithme k-NN a été utilisé souvent à titre de comparaison avec d'autres techniques d'apprentissage automatique (par exemple SVM) lorsqu'elle est appliquée à des données d'expression génique. Cet algorithme fonctionne assez bien mais pas aussi bon, mais pas loin par rapport aux méthodes plus sophistiquées **[104]**.

### *b) Classificateur bayésien:*

Le Classificateur bayésien est basé sur la théorie de Bayes, il est utilisé pour calculer les probabilités d'appartenance à une classe c'est à dire la probabilité qu'un tuple donné appartient à une classe particulière ou non. Sa précision et sa vitesse sont extraordinaire lorsqu'elle est appliquée à un grand ensemble de données **[49]**.

- Le théorème de Bayes : $\mathbf{P(H|X)} = \frac{\mathbf{P(X|H)P(H)}}{\mathbf{p(X)}}$

Avec : $\mathbf{P(H|X)}$ la probabilité a posteriori de H sachant que X.

X sont des observations connues et **P(H)** la probabilité a priori. $\mathbf{P(X|H)}$ et $\mathbf{P(H)}$ sont généralement estimés à partir des données fournies.

- Dans le cas d'une classification, supposons qu'il existe *m* classes $\mathbf{(C_i/\ i=1,\ldots,\ m)}$. Étant donné un tuple $\mathbf{X} = (\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n})$, le classificateur prédit que *X* appartient à la classe ayant la plus haute probabilité a posteriori, sachant que X.

$$\mathbf{P(C_i|X)} = \frac{\mathbf{p(X|C_i)P(C_i)}}{\mathbf{P(X)}}$$

L'objectif est alors de *maximiser* $\mathbf{p(X|C_i)}$ car **P(x)** est fixé et la probabilité a priori pour les $\mathbf{C_i/\ i=1,\ldots,\ m}$, c'est à dire $\mathbf{P(C_i)}$ ne sont pas connus, ils sont supposés être identiques. Pour prédire la classe d'un objet donnée, nous avons juste besoin de trouver la classe $C_i$ tel que $\mathbf{p(X|C_i)P(C_i)}$ est le maximum avec les valeurs qui sont estimées à partir de l'ensemble d'apprentissage.

Ce classificateur est souvent utilisé grâce à leur simplicité, il manipule directement le théorème de Bayes, mais à cause de leur hypothèse naïve de l'indépendance de caractéristiques, il est très sensible à leur corrélation. L'application des classificateurs bayésiens est très fréquente dans le domaine d'analyse des données de puce à ADN, la plupart des travaux montrés leur efficacité **[104]**.

### *c) Les machines à vecteur de support :*

Les machines à vecteurs de support ou séparateurs à vastes marges (notées SVM pour *Support Vector Machines* en anglais) c'est une nouvelle méthode utilisée pour résoudre les *problèmes de discrimination*. Cette méthode découle directement des travaux de *Vapnik* en théorie de l'apprentissage à partir de 1995. Une SVM est un algorithme d'apprentissage, permettant d'apprendre un séparateur, trouver un séparateur revient à construire une fonction qui prend un vecteur de notre ensemble et peut dire de quelle classe il est. Un SVM linéaire calcule l'équation d'un hyperplan (une droite) qui sépare les classes des données et maximise la marge entre lui-même et les individus les plus proches. L'équation de l'hyperplan est de la forme $D(x) = w.x + b$ où $w$ est un vecteur de poids, $b$ est une constante et $x$ correspond aux coordonnées d'un individu dans l'espace des variables. La marge de sécurité introduite est portée par les observations les plus proches de l'hyperplan : ce sont *les vecteurs de support* (**Figure IV.7)** qui donnent leur nom à la méthode **[50].**

Les SVMs s'appliquent aussi des problèmes *non linéairement séparables*, cela nécessite le passage dans un espace de plus haute dimension de manière à se ramener à une séparation linéaire. Ce passage nécessite une transformation non linéaire φ des vecteurs d'entrées, le calcul de l'hyperplan séparateur optimal nécessite seulement la connaissance du produit scalaire entre deux points images de φ . Il suffit donc de définir une fonction noyau à valeur réelle vérifiant $K(x_1, x_2) = \varphi(x_1).\varphi(x_2)$, un SVM est alors uniquement caractérisé par sa fonction noyau. Les noyaux classiques sont linéaires, polynomiaux ou gaussiens **[50]**.

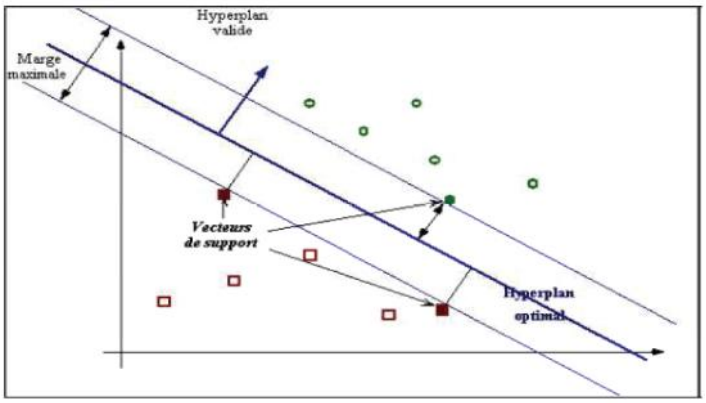



**Figure IV.7 : La représentation schématique d'un SVM [50].**

Les SVMs sont largement utilisés grâce à leurs avantages : elles sont très précises; elles sont moins sensibles au *surapprentissage*; elles permettent de modéliser des données très complexes, non linéaires et elles fournissent une description compacte du modèle appris.

Avec ces avantages, la technique SVM montre leur efficacité dans le domaine d'analyse des données d'expression génique où les utilisateurs de cette méthode ont été obtenus des signatures biologiquement très pertinentes qui ont une haute précision **[104].**

### IV.3.3. Ensemble de classificateurs :

Le principe général des méthodes d'ensemble dans l'apprentissage automatique consiste à construire une collection de classificateurs, pour ensuite agréger l'ensemble de leurs résultats de classification. Le succès de ces méthodes d'ensemble peut résumer en deux points :

- Chaque prédicteur individuel doit être relativement bon.
- Les prédicteurs individuels doivent être différents les uns des autres.

Pour le premier point est un point nécessaire, car l'agrégation des prédicteurs tous mauvais ne pourra vraisemblablement pas donner un bon prédicteur. Pour le deuxième point est également naturel, car l'agrégation des prédicteurs semblable donnera encore un prédicteur semblable et n'améliorera pas les prédictions **[73]**.

#### IV.3.3. 1. Les méthodes de construction d'un ensemble de classificateur :

Les méthodes de construction d'un ensemble de classificateurs sont multiples mais les plus utilisées et les plus connues sont les méthodes de Boosting et le Bagging **[73].**

##### *a) Boosting :*

Cette méthode introduite par Freund and Schapire en 1996, le Boosting est une des méthodes d'ensemble les plus performantes à ce jour, elle est basé sur l'idée de la combinaison des classificateurs simple ou bien faible (obtenus par un apprenant faible comme les arbres de décision, les réseaux de neurones…etc.) en un seul classificateur solide ou bien fort afin d'obtenir une plus grande précision que les classificateurs faibles. La variante AdaBoost (Adaptive Boosting), qui a d'abord été introduit par Freund et Schapire en 1996 **[36],** est un algorithme qui améliore l'algorithme simple de Boosting via un processus itératif. L'idée principale de cet algorithme est associée un poids à chaque exemple dans l'ensemble d'apprentissage. Au départ, le même poids est attribué à tous les exemples, puis, à chaque nouvelle itération, un classificateur est appris à partir des données pondérées et selon l'erreur de prédiction associée à chaque exemple le poids sera modifié.

**Entrée** : ensemble d'apprentissage[illegible] $A = \{a^1, a^2, \ldots, a^N\}$, avec $a^i = (x^i, y^i)$

M : le nombre de classificateurs

**Sortie** : un classificateur H (méta-classificateur)

1. Initialiser les poids $w_i = 1/n$ , $i \in \{1, \ldots, N)$, et met m=1
2. Tant que $m \leq M$ faire
3. Entrainer l'apprenant Faible sur A, utilisant les poids $w_i$ , produisant classificateur $H_m$
4. Calculer l'erreur pondérée de $H_m$, $err_m = \sum_{i=1}^{N} w_i^{(m)} h(-y_i H_m(x_i))$
5. Calculer le poids de classificateur faible, $\alpha_n = \frac{1}{2}\log(\frac{1-err_n}{err_n})$
6. Pour chaque exemple i=1,…,N, mettre à jour les poids $v_i^{(m)} = w_i^{(m)} \exp(-\alpha_n y_i H_m(x_i))$
7. Renormalinser les poids, calculer $S_m = \sum_{j=1}^{n} v_j$ et pour i=1,…,N , $w_i^{(m+1)} = v_i^{(m)}/S_m$
8. m=m+1
9. Finq
10. Le classificateur final : $H(x) = sign(\sum_{j=1}^{m} \alpha_j H_j(x)$

**Figure IV.8 :L'algorithme général d'AdaBoost pour une classification binaire [73].**

### *b) Bagging :*

Le *Bagging* , acronyme pour *bootstrap aggregating* a été présentée par Breiman en 1996 **[14]**. L'idée du Bagging consiste à générer différents ensembles de données (échantillons bootstrap) à partir de l'ensemble d'apprentissage initial d'une manière uniformément au hasard et avec remise à l'aide d'une méthode nommée bootstraping, ce qui permet de construire une collection de prédicteurs variés. L'étape d'agrégation permet alors d'obtenir un prédicteur plus performant **[73]**.

**Entrée** : ensemble d'apprentissage $A = \{a^1, a^2, \ldots, a^N\}$, avec $a^t = (x^t, y^t)$

B : le nombre d'échantillons bootstrap.

**Sortie** : un classificateur H (méta-classificateur)

1. Pour b=1 à B faire
2. Établi, avec la remise, N échantillons à partir A, obtenant le b-ième échantillon bootstrap $A_b^*$.
3. A partir chaque échantillon bootstrap $A_b^*$ , apprendre le classificateur $H_b$
4. Finpour
5. Produire le classificateur final par le vote majoritaire de $H_1, \ldots, H_B$, avec $H(x) = (\sum_{b=1}^{B} H_b(x))$

**Figure IV.9 :L'algorithme de base de Bagging [73].**

### IV.3.3. 2. Les techniques de combinaison (le vote majoritaire):

Plusieurs techniques d'agrégation sont couramment utilisées pour combiner les classifications fournies par les différents classificateurs **[73]**. Mais la technique la plus utilisée c'est la technique de *vote majoritaire*. Le principe de base de cette technique consiste à considérer la sortie de chaque classificateur comme un vote pour une classe. Puis le nombre de votes pour chacune des classes est compté et l'ensemble choisit la classe ayant le vote majoritaire. L'inconvénient majeur de cette technique est lié à l'hypothèse que les différents classificateurs ont une fiabilité similaire, dans ce cas la décision sera difficile. La précision de chaque classificateur joue un rôle très important dans la décision finale et pour prendre en compte la précision individuelle de chaque classificateur, nous pouvons l'associer un poids proportionnel à sa précision. Il y a des autres techniques comme la pondération de performance, les techniques basées modèle additif…etc.

### IV.3.3. 3. Le méta ensemble :

Le méta ensemble est un nouveau concept proposé afin d'améliorer les résultats des méthodes d'ensemble de classificateurs. Il consiste à construire l'ensemble d'ensemble de classificateurs c'est-à-dire utilisé plusieurs méthodes d'ensemble (Bagging, Boosting…etc.) qui fournies des différents résultats de classification ensuite une étape d'agrégation est effectuée sur ces résultats dont l'objectif d'exploiter les avantages de chacune des méthodes **[91]**. Cette idée a été d'abord étudiée par Dettling **[27]** qui a proposé de combiner les algorithmes Bagging et Boosting (appelé *BagBoosting*) pour la classification des données de

puces à ADN. L'hypothèse sous-jacente est que l'ensemble Boosting possède des inconvénients qui sont résolus par le Bagging tandis que, le Bagging possède aussi des inconvénients qui sont traités par le Boosting. Donc, la combinaison de ces deux méthodes d'ensemble peut aboutir à un outil de prédiction qui pourrait atteindre des performances très élevées. La figure suivante illustre ce concept.

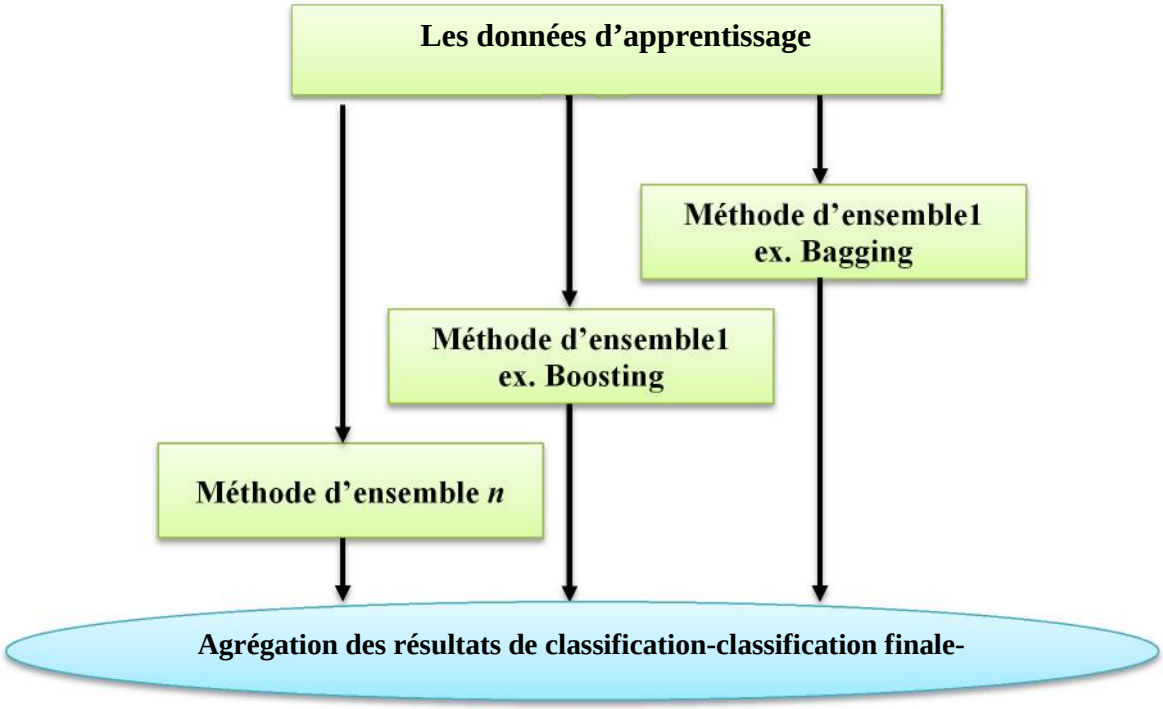



**Figure IV.10 : L'organisation en couches dans le méta ensemble**

## IV.4. Conclusion :

L'analyse de données de puces à ADN pour découvrir les biomarqueurs confronte les chercheurs à un problème de haute dimensionnalité. L'objectif est souvent de trouver un nombre restreint de gènes qui représentent la bonne signature biologique. Dans ce chapitre, nous avons traité les techniques utilisées pour résoudre ce problème qui sont la sélection de caractéristiques et la classification .D'abord, une revue du domaine de la sélection de caractéristiques a été présentée. Après avoir exposé les motivations de cette technique et leur définition, nous avons présenté les approches de sélection filter, wrapper et embedded avec les avantages et les inconvénients de chacune, où les approches filter ont l'avantage de la rapidité mais elles sont critiquées par l'indépendance au classificateur, les approches wrapper demande un temps de calcul très élevé mais la dépendance au classificateur permet d'obtenir des résultats performantes et les approches embedded qui sont définis comme des approches intermédiaire entre la rapidité et la qualité du solution. Puis nous avons expliqué le critère de la stabilité, leur avantage dans le domaine biologique, les causes de l'instabilité et les solutions proposées dans la littérature. Ensuite, dans la dernière section nous avons présenté la deuxième technique utilisée qui est la classification, nous avons exposé quelques techniques de classification et les nouveaux concepts ensemble et méta ensemble de classificateur.

La plupart des travaux montre l'efficacité des méthodes wrapper pour trouver des sous-ensembles de gènes qui ont une petite taille et une haute performance même le temps de calcul est élevé. Ce type de méthodes généralement définis comme des hybridations entre les algorithmes d'optimisation et les algorithmes de classification. Dans le chapitre suivant nous présenterons trois algorithmes d'optimisation qui sont les algorithmes génétiques, l'optimisation par essaim particulaire et la sélection clonale et comment sont utilisées pour une sélection de caractéristiques.

## V.1. Introduction :

Les phénomènes physiques et biologiques ont été à la source de nombreux algorithmes s'en inspirant, nous citons les algorithmes évolutionnaires, les algorithmes de système immunitaire artificiel…etc. L'apparition des algorithmes évolutionnaire qui s'inspirent de la théorie de l'évolution fait un grand pas dans les domaines de la résolution de problèmes complexes. Dans ce type des algorithmes nous trouvons par exemple les algorithmes génétiques, l'optimisation par essaim de particules qui se présente comme une alternative aux algorithmes génétiques et aux colonies de fourmis pour l'optimisation de fonctions non-linéaires. Ainsi, les algorithmes de système immunitaire artificiel qui s'inspire du fonctionnement du système immunitaire humain et qui fournit avec leur collection des algorithmes (la sélection négative, la sélection clonale…etc.) une grande évolution pour les problèmes d'optimisation. Dans ce type des algorithmes, ils ne s'agissent pas de trouver une solution analytique exacte, ou une bonne approximation numérique, mais de trouver des solutions satisfaisant au mieux à différents critères, souvent contradictoires. S'ils ne permettent pas de trouver à coup sûr la solution optimale de l'espace de recherche, du moins nous pouvons constater que les solutions fournies sont généralement meilleures que celles obtenues par des méthodes plus classiques, pour un même temps de calcul. La recherche de la solution se fait d'une manière itérative, elle commence par une population initiale et déroule pendant certain nombre des itérations pour terminer par le retour de la solution optimale si un certain critère d'arrêt est satisfait. L'aspect itératif de ces algorithmes permet d'explorer l'espace de recherche et améliorer la solution à chaque itération afin d'obtenir la solution optimale. L'application de ces algorithmes dans la sélection de caractéristiques est pas récente mais elle connue un essor très rapide et plusieurs adaptations et variantes sont proposée où les résultats obtenus sont de haute qualité c'est-à-dire un petit ensemble de caractéristiques avec une erreur minimale de classification, mais elle est critiquée toujours par leur complexité. Dans ce chapitre nous expliquerons ces algorithmes et comment sont appliqués dans le domaine de la sélection de caractéristiques.

## V.2. Les algorithmes génétiques :

Les algorithmes génétiques (AG) initiés par John Holland en 1960, il a été étudié avec son groupe les systèmes évolutifs et, en 1975, il a été introduit le premier modèle formel des algorithmes génétiques (the Canonical Genetic Algorithm CGA) dans son livre *Adaptation in Natural and Artificial Systems*. Ce modèle servira de base aux recherches ultérieures et sera plus particulièrement repris par Goldberg qui publiera en 1989, un ouvrage de

vulgarisation des algorithmes génétiques qui est ajouté à la théorie des algorithmes génétiques les deux idées suivantes **[97]**:

- Un individu dans la population est lié à un environnement par son code (le codage ou la représentation).
- Une solution est liée à un problème par son indice de qualité (fitness).

### V.2. 1. Le principe de base des algorithmes génétiques :

Les algorithmes génétiques (GA) sont des algorithmes de recherche heuristique adaptative basé sur la conjecture de la sélection naturelle et la génétique. Le concept de base des algorithmes génétiques est conçu pour simuler les processus nécessaire dans un système naturel pour l'évolution, en particulier ceux qui suivent la survie du plus apte, inspiré par le principe de Darwin[1]. Ils représentent une exploitation intelligente d'une recherche aléatoire dans un espace de recherche défini pour résoudre un problème donné **[107]**.

Un AG standard nécessite tout d'abord un codage de l'ensemble des paramètres du problème d'optimisation à résoudre en une chaîne de longueur finie. Ensuite, le principe est simple, il s'agit de simuler l'évolution d'une population d'individus par des changements héritables, en passant d'une génération à la génération suivante jusqu'à un critère d'arrêt. Nous commençons par générer aléatoirement une population initiale d'individus. A chaque génération, des individus sont sélectionnés en utilisant fonction objectif appelée la fonction d'adaptation ou la fonction de fitness. Puis, les opérateurs de croisement et de mutation sont appliqués et une nouvelle population est créée. Ce processus est itéré jusqu'à un critère d'arrêt est satisfait. Le critère d'arrêt le plus couramment utilisé est le nombre maximal de générations que le nous désirons effectuer. La figure (**Figure V.1)** représente le principe de l'AG standard.

1. Codage du problème sous forme d'une chaîne (binaire, réelle…etc.)
2. Génération aléatoire d'une population initiale. Celle-ci contient un pool génétique qui représente un ensemble de solutions possibles.

Répéter

3. Calcul d'une valeur d'adaptation pour chaque individu. Elle sera fonction directe de la proximité des différents individus avec l'objectif, nous parlons ici sur l'évaluation (fitness).

[1] Les animaux les plus adaptés à leur milieu qui survivent. Ce sont donc eux qui auront le plus de chance de se reproduire, et donc de transmettre leurs gènes.

4. Sélection des individus doit se reproduire en fonction de leurs parts respectives dans l'adaptation globale.
5. Croisement et mutation des individus des parents.
6. Vérifier le critère d'arrêt, si il est satisfait nous quittons sinon, sur la base de ce nouveau pool génétique nous repartons à partir du point 3.

**Figure V.1 : Le principe de base de l'AG standard [97].**

### V.2. 1. 1. Les niveaux d'organisation d'un AG :

Les informations dans un algorithme génétique sont organisées en cinq niveaux, la population initiale est constituée d'un ensemble des individus, chaque individu doté d'un génotype constitué d'un ou plusieurs chromosomes. Le chromosome est un ensemble de gènes et chaque gène correspond un paramètre lié au problème d'optimisation à résoudre, il est représenté sur un bit **[97]**.

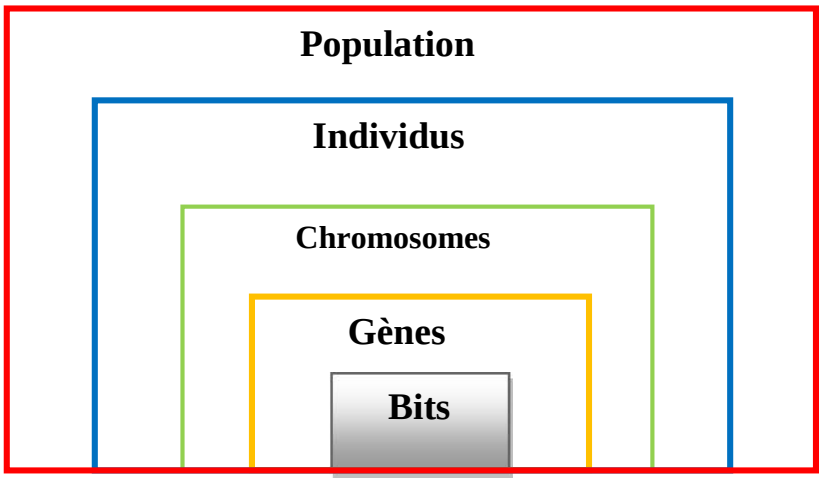



**Figure V.2 : Les niveaux d'organisation de l'information dans les AGs.**

### V.2. 1. 2. Les opérateurs génétiques :

Nous avons déjà vu que les AGs sont des algorithmes évolutionnaire basée sur le principe de la reproduction. La reproduction dans ce cas nécessite l'application des opérateurs génétiques qui sont : *la sélection, le croisement et la mutation*.

#### *a) La sélection :*

La sélection a pour objectif d'identifier les individus qui doivent se reproduire. Cet opérateur ne crée pas de nouveaux individus mais identifie les individus sur la base de leur fonction d'adaptation qui vont servir de parents dans l'étape suivante. Cette opérateur est le

plus important puisqu'il permet aux individus d'une population de survivre, de se reproduire ou de mourir. En règle générale, la probabilité de survie d'un individu sera directement reliée à son efficacité relative au sein de la population. Plusieurs opérateurs de sélection existent parmi lesquels la sélection à la roulette, par tournoi, par élitisme et par rang.

***i) La sélection à la roulette* [40] *:***

C'est la méthode la plus connue et la plus utilisée. Avec ce type de sélection, chaque individu a la probabilité d'être sélectionné proportionnelle à sa performance, donc plus les individus sont adaptés au problème, plus ils ont de chances d'être sélectionnés. Ainsi un individu $x_i$, dans une population de taille N, a la probabilité suivante d'être sélectionné :

$$P_s\,(\,x_i) = \frac{fitness(\,x_i)}{\sum_{j=1}^{N} fitness(\,x_j)}$$

***ii) La sélection par tournoi* [77]*:***

Le principe de cette méthode consiste à effectuer un tirage avec remise de deux individus de la population P, et nous le faisons *combattre*. Celui qui a la fitness la plus élevée l'emporte, nous répétons ce processus *n* fois de manière à obtenir les *n* individus qui serviront de parents.

***iii) La sélection par élitisme :***

Le principe de cette méthode consiste à trier d'une manière décroissante les individus selon leurs performances et les meilleurs (un ou plusieurs) sont copiés dans la nouvelle population ; Puis, nous générons le reste de la population selon l'algorithme de sélection usuel. Cette méthode amélioré considérablement les algorithmes génétiques car elle permet de ne pas perdre les meilleures solutions au cours d'évolution **[46]**.

***iv) La sélection par rang* [24]*:***

Cette méthode consiste à attribuer à chaque individu un classement proportionnel à sa performance mesurée en utilisant la fonction de fitness, l'individu le plus mauvais prendra le rang 1 et le meilleur aura le rang N tel que N est la taille de la population courante, alors la probabilité de sélection d'un individu $x_i$ devient :

$$P_s\,(\,x_i) = \frac{Rang(\,x_i)}{\sum_{j=1}^{N} Rang(\,x_j)}$$

***b) Le croisement :***

C'est l'opérateur qui permet de créer de nouvelles chaînes (enfants) en échangeant de l'information entre deux chaînes (parents). Les enfants obtenus héritent partiellement des caractéristiques des parents sélectionnés dans l'étape de sélection. Le rôle fondamental d'un

croisement est de permettre la recombinaison des informations présentes dans le patrimoine génétique de la population. Le processus de croisement appliqué à chaque paire d'individus avec une certaine probabilité $P_c$. Plus $P_c$ est élevée, plus il y a de nouveaux individus qui apparaissent dans la nouvelle population. En général, le croisement est de deux types : à un point ou à des points multiples. Dans le cas d'un seul point, nous sélectionnons d'une manière aléatoire un point pour chaque couple d'individus et d'effectuer une permutation des séquences entre ces individus. Le croisement à multi points, consiste à sélectionner plusieurs points d'une manière aléatoire et l'échange entre les parents se fait sur les différentes parties des séquences cernées par ces points (**Figure V.3).**

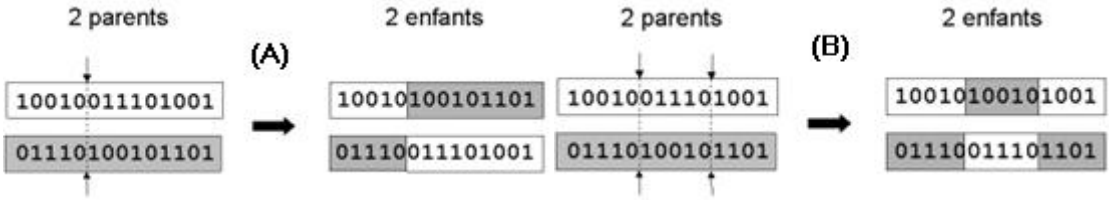



**Figure V.3 : Les mécanismes du croisement, (A) : croisement à un point, (B) croisement à deux points [46].**

Il existe un autre type de croisement nommé *le croisement uniforme*, il consiste à définir un masque c'est-à-dire une chaine de bits {0,1} de même longueur que les individus parents sur lesquels il sera appliqué. Lorsque le bit du masque égal à 0, l'enfant hérite le bit du premier parent, sinon il hérite de celui du second parent. Le second enfant est le complémentaire du premier. Ce croisement peut être considéré comme une généralisation du croisement multi points sans connaissance préalable du point de croisement **[97]**.

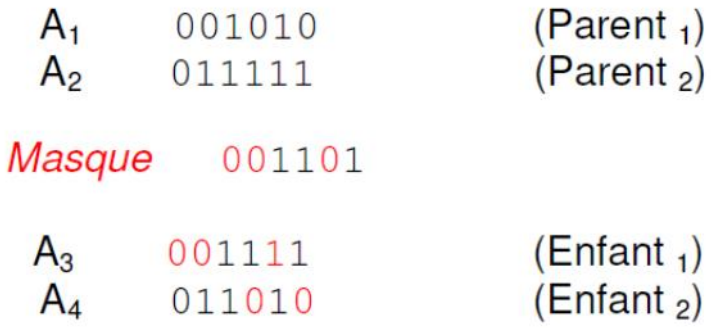



**Figure V.4 : Exemple d'un croisement uniforme [97].**

Nous pouvons noter que le nombre de points de croisements (point, deux points ou multi points) ainsi que la probabilité de croisement $P_c$ permettent d'introduire plus ou moins de diversité dans la recherche.

### *c) La mutation :*

La mutation est exécutée seulement sur un seul individu. Elle représente la modification aléatoire et occasionnelle de faible probabilité de la valeur d'un bit d'individu, pour un codage binaire cela revient à changer un 1 à 0 et vice versa. Cet opérateur introduit de la diversité dans le processus de recherche des solutions et peut aider l'AG à ne pas stagner dans un optimum local c'est-à-dire permet d'explorer l'espace de recherche.

Alors pour appliquer un algorithme génétique, l'utilisateur doit déterminer les paramètres suivants:

- *La taille de la population* c'est-à-dire le nombre des individus liée à la nature du problème à résoudre
- *La stratégie de sélection* associée aussi au problème à résoudre
- *La probabilité du croisement* $P_c$., plus elle est élevée, plus il y a de nouveaux individus qui apparaissent dans la nouvelle population.
- *La probabilité de mutation* $P_m$., généralement très petite pour éviter l'exploration chaotique dans l'espace de recherche.
- *Le critère d'arrêt*, qui est généralement un nombre maximal des itérations.

## V.2. 2. Avantages et inconvénients :

Les algorithmes génétiques possèdent plusieurs avantages qui permettent de les utiliser dans des domaines multiples mais ils ont aussi des inconvénients qui imposent des restrictions dans leur utilisation, nous citons quelques-uns

### *a) Avantage* **[66]**:

- Les AGs est l'un des algorithmes évolutionnaires qui peuvent balayer d'une manière rapide un vaste ensemble de solutions et les mauvaises propositions n'affectent pas la solution finale car ils sont tout simplement éliminés.
- Les AGs permettent de trouver *les bonnes solutions* sur des problèmes très complexes d'une manière simple, ils s'agissent de déterminer entre deux solutions quelle est la meilleure, afin d'opérer leurs sélections. Ils sont appliqués dans les domaines où un grand nombre de paramètres entrent en jeu et nous avons besoin d'obtenir de bonnes solutions en quelques itérations seulement.
- Ils ont la capacité de résoudre tous les problèmes d'optimisation qui peuvent être décrit avec le codage de chromosomes.

- L'algorithme génétique est une méthode qui est très facile à comprendre et il n'exige pas la connaissance des mathématiques.

***b) Inconvénients* [66]*:***

- Un AG est très lent et ne peut pas toujours trouver *la solution optimale* quand le problème possède un grand nombre de paramètres.
- Il n'y a pas de garantie absolue qu'un algorithme génétique peut trouver toujours un optimum global.
- Comme d'autres techniques d'intelligence artificielle, l'algorithme génétique ne peut pas assurer des temps de réponse constants. Cette propriété limite leur utilisation dans les applications *en temps réel*.

## V.3. L'optimisation par l'essaim particulaire :

L'optimisation par essaims particulaires (en anglais Partical Swarm Optimization PSO) est un algorithme évolutionnaire stochastique[2], basée sur la reproduction d'un comportement social. Elle est inventée par Russel Eberhart (ingénieur en électricité) et James Kennedy (socio-psychologue) en 1995 **[63].** Cet algorithme s'inspire à l'origine du monde du vivant grâce à des observations faites lors des simulations informatiques de vols groupés d'oiseaux et de bancs de poissons. Ces simulations ont mis en valeur la capacité des individus d'un groupe en mouvement à conserver une distance optimale entre eux et à suivre un mouvement global par rapport aux mouvements locaux de leur voisinage **[41].** La PSO se base donc sur la collaboration des individus entre eux et elle s'appuie sur le concept *d'auto-organisation*[3]. Grâce à des règles de déplacement très simples (dans l'espace des solutions), les particules peuvent converger progressivement vers l'optimum global.

### V.3.1. Le principe de PSO :

L'optimisation par essaim de particules repose sur une population d'individus appelée essaim, ces individus ou bien particules sont originalement disposés d'une manière aléatoire et homogène, qui se déplacent dans l'espace de recherche et constituent, chacune, une solution potentielle. A chaque itération de l'algorithme, les particules se déplacent dans l'espace en prenant en compte leur meilleure position et la meilleure position de son

[2]Sont les algorithmes qui utilisent itérativement des processus aléatoires. Donc, plusieurs exécutions successives de tels programmes ne produisent pas forcément le même résultat.
[3] Désigne la capacité des éléments d'un système à produire et maintenir une structure à l'**échelle** du système sans que cette structure apparaisse au niveau des composantes et sans qu'elle résulte de l'intervention d'un agent extérieur

voisinage. Dans les faits, nous calculons la nouvelle vitesse à l'instant *t+1* pour la particule *i* par la formule suivante **[32]**:

$$V_i(t+1) = a.V_i(t) + b_1\left(X_{pbest_i} - X_i(t)\right) + b_2\left(X_{vbest_i} - X_i(t)\right) \ldots \textbf{(V.1)}$$

Où chaque particule caractérisé par **[41]**:

1. $X_i(t)$ : sa position dans l'espace de recherche à l'instant *t*
2. $X_i(t)$ : sa vitesse à l'instant *t*
3. $X_{pbest_i}$ : la position de la meilleure solution par laquelle elle est passée
4. $X_{vbest_i}$ : la position de la meilleure solution connue de son voisinage
5. $pbest_i$ : la valeur de fitness de sa meilleure solution
6. $vbest_i$ : la valeur de fitness de la meilleure solution connu du voisinage

Avec **: a :** appelé la masse d'inertie, est une constante.

$b_1 = c1 * r1$ . Où c1 est une constante appelée le poids cognitif(ou personnel ou local) et r1 est une variable aléatoire dans la plage [0,1].

$b_2 = c2 * r2$ . Où c2 c'est une constante nommée le poids social ou globale et r2 est une variable aléatoire dans la plage [0,1].

A partir de la nouvelle vitesse calculée la particule peut calculer leur position à l'instant *t+1* selon la formule suivante :

$$X_i(t+1) = X_i(t) + V_i(t+1) \ldots \textbf{(V.2)}$$

Où $X(t)$ représente la position actuelle de la particule. L'algorithme de base de la PSO est illustré dans le pseudo code suivant (**Figure V.5 )**:

**Initialisation**

*1 Particule*= 1solution potentielle du problème d'optimisation

*nb :* le nombre des particules

Pour chaque particule

- Fixer sa position d'une manière aléatoire dans l'espace de recherche $X_i$.
- Fixer sa vitesse d'une manière aléatoire $V_i$.
- Définir *F* la fonction de fitness
- Définir le voisinage (la taille du voisinage)

**Répéter**

*Pour i=1 à nb faire*

Calculer sa fitness F($x_i$)

**Si** $F(x_i) > pbest_i$

Mettre à jours $pbest_i = F(x_i)$

Mettre à jours $\boldsymbol{X_{pbest_i}} = \boldsymbol{X_i}$

**Finsi**

**Si** $F(x_i) > vbest_i$

Mettre à jours $vbest_i = F(x_i)$

Mettre à jours $\boldsymbol{X_{vbest_i}} = \boldsymbol{X_i}$

**Finsi**

*Finpour*

*Pour i=1 à nb faire*

Calculer sa vitesse $\boldsymbol{V_i}(t+1)$ **… (selon la formule V.1)**

Mettre à jours sa position $\boldsymbol{X_i}(t+1)$ **(selon la formule V.2)**

*Finpour*

**Jusqu'à** un critère d'arrêt est satisfait

**Figure V.5 : L'algorithme de base de PSO [41].**

Alors pour définir un algorithme PSO nous avons besoins de quatre éléments concernant l'essaim:

1. Le nombre de particules de l'essaim ;
2. la vitesse maximale d'une particule $V_{max}$ ;
3. la topologie et la taille du voisinage d'une particule qui définissent son réseau social ;
4. les paramètres : d'inertie d'une particule et les coefficients de confiance.

### V.3.1. 1. La notion de voisinage :

Le voisinage d'une particule est le sous-ensemble de particules de l'essaim avec lequel il a une communication directe. Ce réseau entre toutes les particules est connu comme la sociométrie, ou la topologie de l'essaim. Il existe deux principaux types de voisinage :

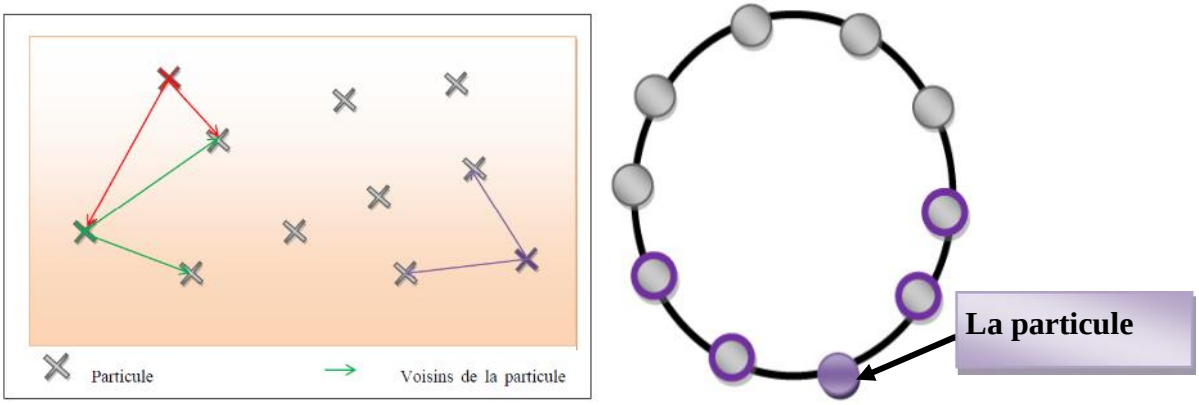
- *Les voisinages géographiques :* les voisins sont considérés comme les particules les plus proches. Cependant, à chaque itération, les nouveaux voisins doivent être recalculés à partir d'une distance prédéfinie dans l'espace de recherche. C'est donc un voisinage *dynamique*. Il est non préférable à utiliser car, d'une part, il s'agirait d'un voisinage trop local, et d'autre

part car il est très lourd en terme de calculs, nécessitant de recalculer le voisinage de chaque particule à chaque itération **[41]**.

- *Les voisinages sociaux :* les voisins sont définis à l'initialisation et ne sont pas modifiés, alors est un voisinage *statique*. C'est le voisinage le plus utilisé, pour plusieurs raisons :

1. Il est plus simple à programmer.
2. Il est moins coûteux en temps de calcul.

Pour ce faire, nous disposons (virtuellement) les particules en cercle ensuite, pour la particule étudiée, nous incluons progressivement dans ses informatrices, d'abord elle-même. Puis, les plus proches à sa droite et à sa gauche, jusqu'à atteindre la taille. La taille de voisinage est donnée dans la partie d'initialisation. Il existe plusieurs topologies pour définir le voisinage, la topologie étoile où chaque particule est reliée à toutes les autres, c'est-à-dire l'optimum du voisinage est l'optimum global ; la topologie en rayon où les particules ne communiquent qu'avec une seule particule centrale…etc. **[41]**.




**(a)** **(b)**

**Figure V.6 : Le voisinage dans la PSO : (a) représente un voisinage géographique, (b) représente le voisinage social (topologie anneau) [41].**

### V.3.1. 2. Les paramètres d'un algorithme PSO :

A première vue, il semblerait que de nombreux paramètres sont à prendre en compte pour l'application de PSO. Les paramètres nécessaires sont :

- *La taille de l'essaim* qui représente la taille de la population c'est la quantité de particules allouées pour résoudre le problème dépend essentiellement à la taille de l'espace de recherche et le rapport entre les capacités de calcul de la machine et le temps maximum de recherche. Il n'y a pas de règle pour déterminer ce paramètre **[41].**
- *Les poids de confiance*, notés précédemment c1 et c2, sont des poids selon lequel une exploration locale ou globale est préférée. La sélection de ces paramètres a un impact sur la vitesse de convergence et la capacité de l'algorithme pour trouver l'optimum global **[32]**. Si

nous choisissons c1 = c2 l'algorithme étudierait également les directions des meilleurs particules locales et globales. Maurice C. et Kennedy J. **[75]** ont proposés une restriction pour ces deux paramètres comme suit :

$$\chi = \frac{2}{\left|2 - \varphi - \sqrt{\varphi^2 - 4\varphi}\right|}$$

Avec : $\varphi = c1 + c2, \varphi > 4$

$\chi$ C'est le facteur de constriction utilisé dans le calcul de la vitesse comme suit :

$$V(t+1) = \chi(a.V(t) + b_1(P_i - X(t)) + b_2(P_g - X(t)))$$

- *La vitesse maximale,* afin d'éviter que les particules ne se déplacent trop rapidement dans l'espace de recherche, passant éventuellement à côté de l'optimum, il peut être nécessaire de fixer une vitesse maximale $V_{max}$ pour améliorer la convergence de l'algorithme. Eberhart Russel et al. **[29]** suggèrent que, dans la pratique, la vitesse devrait limiter entre l'intervalle [$-V_{max}$, $V_{max}$] où $V_{max}$ est habituellement donnée égal à 4. Ils suggèrent aussi que l'utilisation du facteur de constriction $\chi$ donne généralement un meilleur taux de convergence sans avoir à fixer de vitesse maximale.
- Un autre paramètre très important c'est l'inertie, noté précédemment a, ce paramètre détermine la façon dont la vitesse de la particule précédente influe sur la vitesse à la prochaine itération. Une faible inertie favorise la recherche locale et une haute inertie favorise la recherche globale. Supposons que **c1=c2=e**, si a plus grand que e la particule préférera sa propre direction que celle des meilleurs résultats, autrement, la particule préférera de trouver les meilleures particules. Fixer ce facteur, revient donc à trouver un compromis entre l'exploration locale et l'exploration globale **[32]** .Eberhart Russel et al. **[29]** indiquent une meilleure convergences pour $a \in [0.8, 1.2]$. Au-delà de **1.2**, l'algorithme tend à avoir certaines difficultés à converger. Enfin, ils suggèrent qu'il est possible de faire diminuer le facteur d'inertie au cours du temps, un peu à la manière de la température dans un algorithme de recuit simulé. De bons résultats ont été trouvés lorsque a est progressivement changé au cours de la procédure entre 0.9 et 0.4 (0,9 au début des itérations et 0,4 dans les itérations finales), l'objectif ici est de favoriser la recherche globale au début de l'algorithme et la recherche locale plus tard **[41]**.
- *Le critère d'arrêt,* qui est généralement un nombre maximal des itérations.

### V.3.2. Avantages et inconvénients :

Les PSO sont utilisés dans des domaines multiples où elles montrent leur efficacité pour trouver une bonne solution, une analyse sur les avantages et inconvénients de cet algorithme d'optimisation peuvent résumer dans les points suivants :

#### a) Avantages [87] :

- La PSO est basé sur l'intelligence donc elle peut être appliquée à la fois dans la recherche scientifique et l'ingénierie.
- La PSO n'a pas de chevauchement et de calcul de mutation. La recherche peut être effectuée par la vitesse de la particule. Lors de l'élaboration de plusieurs générations, que la particule la plus optimiste peut transmettre des informations sur les autres particules, et la vitesse de la recherche est très rapide
- Le calcul de PSO est très simple. En comparaison avec les autres algorithmes, elle occupe la plus grande capacité d'optimisation et elle peut être remplie facilement

#### b) Inconvénients [107]:

- La méthode souffre facilement de l'optimisme local, ce qui provoque le moins exact à la régulation de la vitesse et la direction.
- La PSO nécessite plusieurs paramètres qui sont déterminés dans l'étape de l'initialisation et qui influent directement sur la convergence vers l'optimum global.
- L'application de la PSO est simple et peut être facilement parallélisées pour le traitement simultané, mais il a une convergence lente et peut être facilement piégée par les optimums locaux.

## V.4. L'algorithme de la sélection clonale :

L'algorithme de la sélection clonale est l'un des algorithmes de système immunitaire artificiel (en anglais Artificial immun System AIS) qui s'inspire par le mécanisme de défense dans le corps humain. Avant d'expliquer cet algorithme il faut parler d'abord sur le système immunitaire humain.

### V.4. 1. Le système immunitaire humain :

Le système immunitaire naturel, est un système complexe de cellules, molécules et des organes, symbolise un mécanisme d'identification capable de percevoir et de lutte contre la dysfonction de nos propres cellules .Ce système protège le corps contre les agents infectieux comme les virus, les bactéries, champignons et autres parasites. Toute molécule qui peut

être reconnue par le système immunitaire adaptatif est connue comme un *antigène*. Le composant de base du système immunitaire est les *lymphocytes* ou *les cellules de globules blancs*. Les lymphocytes existent sous deux formes, *les cellules B* et *les cellules T*. Ces deux types de cellules sont assez similaires, mais ils diffèrent en ce qui concerne la façon dont ils reconnaissent les antigènes et leurs rôles fonctionnels. Les cellules B sont capables de reconnaître les antigènes libres, tandis que les cellules T ont besoin des autres cellules accessoires qui peuvent lui présenter les antigènes. Elles ont des structures chimiques différentes et produits plusieurs *anticorps* de *forme Y* à partir de leurs surfaces pour tuer les antigènes **[72]**.

- ***Les cellules B et les anticorps:*** les principales fonctions des cellules B comprennent la production et la sécrétion d'anticorps **(Ab)** en réponse à des éléments exogènes comme les bactéries, les virus et les cellules tumorales (antigène). Chaque cellule B est programmée pour produire un anticorps spécialisé. Les anticorps sont des protéines spécifiques qui ont la capacité de reconnaitre et lier à un autre élément particulier. La production et la liaison des anticorps est généralement un moyen de signaler aux autres cellules de tuer, d'ingérer ou de supprimer la substance liée **[70]**.
- ***Les cellules T et les lymphokines:*** les cellules T sont appelés T parce qu'elles matures dans le *Thymus*. Leur fonction inclure la réglementation des actions pour des autres cellules et attaque directement les cellules infectée. Les cellules T peuvent être subdivisées en trois sous-classes principales: les cellules T helper (Th), cytotoxique (tueur) et les cellules T suppresseurs **[70]**.

o *Les cellules T helper*, ou simplement cellules *Th*, elles libèrent des substances nommées les lymphokines qui ont un rôle de recrutement pour d'autres cellules immunitaires (l'activation des cellules B).

o *Les cellules T cytotoxiques*, sont capables d'éliminer les microbes envahisseurs, des virus ou des cellules cancéreuses. Une fois activée et liée à leurs ligands, ils injectent des produits chimiques pour provoquer leur destruction.

o *Les cellules T suppresseurs*, sont essentielles pour le maintien de la réponse immunitaire. Elles sont parfois appelées cellules CD8. Sans leur activité, l'immunité perdre le contrôle ce qui provoque des réactions allergiques et des maladies auto-immunes.

Il existe des autres cellules dans le système immunitaire humain nommées les cellules *phagocytaires*. Ce sont les microphages (polynucléaires) et les macrophages (monocytes sanguins). Ces cellules défendent dans le corps par la *phagocytose*. Le macrophage plus elle

réalise la phagocytose elle a la capacité de présenter l'antigène aux cellules pour synthétiser les anticorps. Le processus de la réponse immunitaire pour protéger le corps est résumé dans les étapes suivantes **[26]**:

**1.** Cellules spécialisées présentatrices d'antigène (*antigen presenting cells* (APCs)), telles que les macrophages, parcourent le corps, l'ingestion et la digestion des antigènes trouvés et leur fragmentation en *peptides antigéniques*.

**2.** Des morceaux de ces peptides sont reliés à une molécule nommé *major histocompatibility complex (MHC)* et sont affichées sur la surface de la cellule. Les cellules T ou les lymphocytes T, ont des molécules réceptrices qui permettent à chacun d'eux de reconnaître des différentes combinaisons de *peptide-CHM*

**3.** Les cellules T activées par la reconnaissance, elles sécrètent des lymphokines, ou des signaux chimiques, qui mobilisent les autres composants du système immunitaire

**4.** Les lymphocytes B, qui ont également des molécules réceptrices, répondre à ces signaux. Contrairement aux récepteurs de cellules T, toutefois, celles des cellules B peut reconnaître les pièces d'antigènes libres, sans les molécules *CHM*.

**5.** Lorsqu'ils sont activés, les cellules B se divisent et se différencient en cellules plasmatiques qui sécrètent des protéines anticorps.

**6.** En se liant aux antigènes qu'ils trouvent, les anticorps peuvent les neutraliser,

**7.** Ou précipiter leur destruction. Certaines cellules T et B deviennent des cellules mémoire qui persistent dans la circulation et renforcer la préparation du système immunitaire pour éliminer le même antigène si elle se présente à l'avenir. Car les gènes codant pour des anticorps dans les cellules B subissent fréquemment à des mutations et des modifications alors les anticorps peuvent être amélioré, ce phénomène est appelé *la maturation d'affinité*[4].

### V.4. 2. L'algorithme de la sélection clonale:

Le système immunitaire artificiel a débuté au milieu des années 80, plusieurs définitions dans la littérature, Castro et Timmis **[15]** définis le AIS comme : « *un système adaptatif, inspiré de l'immunologie théorique et les fonctions immunitaires observés, les principes et les modèles, qui sont appliqués à la résolution de problèmes* ».

Le système immunitaire artificiel possède plusieurs algorithmes qui simulent les mécanismes de défense dans le système immunitaire naturel, nous focalisons dans notre travail sur l'algorithme de la sélection clonale.

[4] L'affinité est le degré de liaison entre le récepteur d'une cellule et l'antigène

### V.4. 2. 1. La théorie de la sélection clonale :

La théorie de la sélection clonale dans le système immunitaire est utilisée pour expliquer la réponse de base du système immunitaire adaptative à un stimulus antigénique. La théorie repose sur l'idée que seules les cellules qui sont capables de reconnaître un antigène prolifèrent **[10].** Les principales caractéristiques de la théorie de la sélection clonale sont: la prolifération et la différenciation de cellules qui reconnaissent l'antigène, génération de nouvelles modifications génétiques aléatoires, exprimé à la suite des modèles d'anticorps divers par une forme de mutation somatique accélérée, estimation des lymphocytes nouvellement différenciées transportant des récepteurs antigéniques de faible affinité **[72].** Nous pouvons résumer le processus de la sélection clonale dans les étapes suivante **[10]** :

**1.** Quand un antigène envahit le corps, des cellules immunitaires reconnaissent cet antigène avec des degrés d'affinité différents.

**2.** Ensuite les cellules B répondent par la production des anticorps dont chaque cellule sécrète un seul type d'anticorps qui est relativement spécifique à l'antigène.

**3.** Lorsque l'appariement entre les récepteurs des anticorps et l'antigène est fort, les cellules B sont stimulées (la prolifération ou bien le clonage et la maturation des cellules).

**4.** Le taux de clonage d'une cellule est directement proportionnel à son affinité avec l'antigène, les cellules qui ont les plus grandes affinités seront les plus proliférées et réciproquement. En plus, les lymphocytes qui ont une forte affinité peuvent se différencier en des cellules mémoires.

**5.** Le taux de la maturation d'une cellule est inversement proportionnel à son affinité.

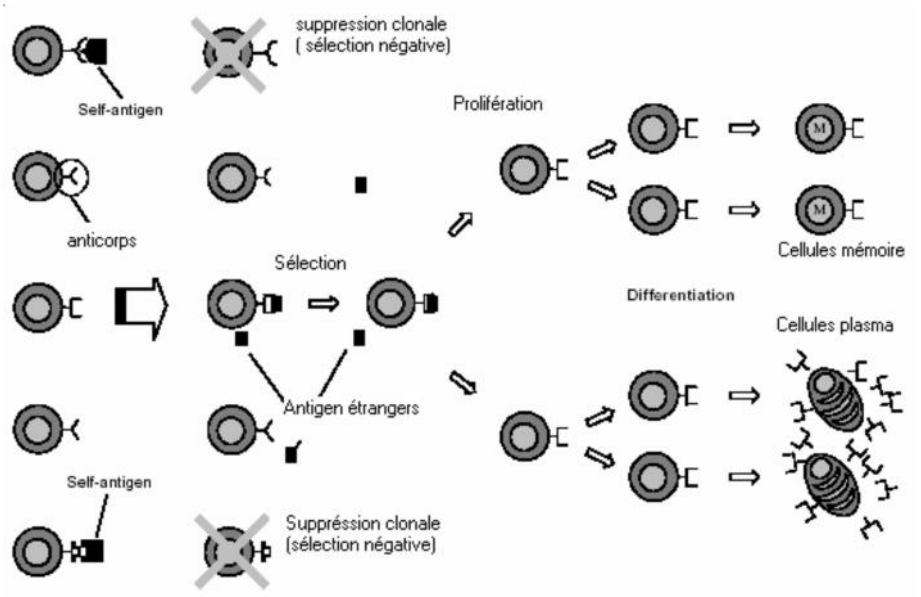



**Figure V.7 : Le principe de la sélection clonale [72].**

### V.4. 2. 2. L'algorithme de la sélection clonale :

En 2000, De Castro et Von Zuben **[26]** ont proposé l'algorithme de la sélection clonale de base nommé CLONALG (CLONal selection ALGorithm). Dans cet algorithme les étapes de base de la sélection clonal sont bien accomplies. L'algorithme est comme suit **[10]** :

1. *Initialisation* : produire une population de solutions (répertoire d'anticorps) de *n* candidats liées au problème à étudier ;
2. *Evaluation*: étant donné un ensemble de modèles pour être reconnu, pour chaque modèle, déterminer son affinité avec chaque élément de la population,
3. *Sélection et clonage*: sélectionner un nombre *n'* des meilleurs éléments de n qui ont l'affinité la plus élevée et de générer des copies de ces individus proportionnellement à leur affinité avec l'antigène,
4. *Hyper mutation*: muter toutes les copies avec taux de maturation inversement proportionnelle à leurs affinités.
5. Ajouter les individus mutés à la population et sélectionner les *n''* individus maturés qui ont la plus grande affinité à l'antigène pour composer le nouveau répertoire ;
6. Sélectionner *d* individus qui possèdent des valeurs d'affinité faible et les remplacer par des individus générés aléatoirement;
7. Répéter les étapes 2 à 6 jusqu'à ce qu'un critère d'arrêt donné soit rencontré.

Un CLONALG donc possède cinq paramètres doivent être définis par l'utilisateur **[10]:**

- *La taille de la population des anticorps (n),* indique le nombre total d'anticorps doit être maintenue par le système. Liée au problème à résoudre.
- *La taille du pool de sélection (n')*, désigne le nombre total d'anticorps avec la plus grande affinité pour les tirer de la population d'anticorps et les cloner. Des valeurs plus petites peuvent réduire la diversité de la population en veillant à ce que les meilleurs anticorps sont clonés.
- *Taille de remplacement (d),* qui représente le nombre d'anticorps d'affinité plus faible pour les remplacer par des autres anticorps aléatoire.
- *Le facteur de clonage (β),* est défini comme étant un facteur de pondération pour le nombre de clones créés pour les anticorps sélectionnés. Plus la valeur de β est petite, plus la recherche dans la zone est effectuée par l'algorithme.
- *Nombre de générations (G),* indique le nombre total d'itérations de l'algorithme à effectuer. Des valeurs élevées de G peut provoquer le problème d'être coincé sur un optimum local.

#### V.4.2.3. Avantages et inconvénients :

Dans cette section nous introduisons en quelques points les avantages et les inconvénients de CLONALG,

**a) Avantage [66]:**

- Le CLONALG est relativement possède une complexité faible et dispose un petit nombre de paramètres de l'utilisateur par rapport à d'autres systèmes AIS tels que le système de reconnaissance immunitaire artificiel (SRIA) proposé par Watkins en 2004 **[10]**
- Il a beaucoup de flexibilité en combinaison avec d'autres algorithmes de recherche stochastiques**.**
- La recherche sur l'optimum global quelles que soient les valeurs des paramètres initiaux.
- Le CLONALG a une convergence rapide.

**b) Inconvénient [66] :**

- Le CLONALG est lent lorsque le problème nécessite une taille de la population élevée surtout quand la fonction d'évaluation des anticorps pour la sélection est compliquée (deux évaluation par itération).
- La sensibilité de l'algorithme aux paramètres.
- La maturation est inversement proportionnelle à l'affinité, la nécessité de déterminer la formule qui calcule le facteur de maturation.

## V.5. L'optimisation multiobjectif :

La fonction objectif nommée aussi la fonction de coût ou de critère, c'est la fonction *f* que l'algorithme d'optimisation vise à l'optimiser (trouver la solution optimale). Cette fonction peut être monoobjectif c'est-à-dire constituée d'un seul objectif à optimiser. Tandis que, dans certains contextes décisionnels, la prise en compte d'un objectif unique est insuffisante, à cette raison l'optimisation multiobjectif a été introduite. Comme le suggère le nom, une fonction d'optimisation multiobjectif consiste à optimiser plusieurs fonctions objectifs simultanément qui sont, en général, *contradictoires* ou *conflictuels*.

Un problème d'optimisation multiobjectif (MOP) peut alors être posé pour une minimisation sous la forme générale suivante **[80]** :

$$\text{MOP} \begin{cases} minimiser \ f(x) = \left(f_1(x), f_2(x), \ldots, f_k(x)\right) \\ x \in E \end{cases}$$

Où $k \geq 2$ et $x = (x_1, x_2, \ldots, x_n)$ représente le vecteur de décision avec $x_i$ les variables du problème et *n* le nombre des variable.

$f(x) = (f_1(x), f_2(x), \ldots, f_k(x))$ le vecteur de k fonctions objectifs $f_i$ et k le nombre d'objectifs

E est l'ensemble non vide des solutions réalisables, c'est-à-dire celles qui respectent les contraintes du problème **[80]** :

- L'ensemble $R^n$ qui contient E est dit *espace de décision* ;
- L'ensemble $R^k$ qui contient F dit *espace des critères ou espace des objectifs* ;
- L'ensemble *F=f(E)* est la projection de l'espace E sur l'espace des objectifs

Le principe d'une optimisation multiobjectif est différent de celle d'une optimisation monoobjectif. Le but principal dans la première est de trouver *la solution optimale globale* ou l'optimum global qui résulte en la meilleure valeur (plus petite ou plus grande) de la fonction monoobjectif. Par contre dans le cas d'optimisation multiobjectif il n'existe pas une solution optimale globale, il y a plus qu'une fonction objectif ($k \geq 2$), chaque fonction objectif pouvant avoir une solution optimale différente. Le but d'un problème multiobjectif est de trouver *le bon compromis* plutôt qu'une seule solution. Afin de résoudre ce type de problème, plusieurs méthodes ont été proposées dans la littérature, elles sont regroupées en deux familles : les méthodes agrégées et les méthodes de Pareto **[80]**.

### V.5. 1. Les méthodes agrégées :

Ce type de méthode est le plus simple consiste à transformer un problème multiobjectif à un problème simple objectif par la combinaison des objectifs. Elles sont utilisées seulement pour les objectifs commensurables c'est-à-dire qui sont exprimés dans la même unité (tous les objectifs sont quantitatifs ou bien tous les objectifs sont qualitatifs). Une des méthodes utilisées pour combiner les solutions c'est la méthode de *la somme pondérée*. Dans cette méthode, en affectant à chaque objectif un poids dépend à leur importance dans le problème, alors la fonction d'objectif devient sous la forme :

$$\sum_{i=1}^{k} w_i f_i$$

Avec $\sum_{i=1}^{k} w_i = 1$

Où k représente le nombre d'objectifs, le $w_i$ c'est le poids affecté au *ième* objectif. Cette méthode est très efficace et simple à mettre en œuvre mais les difficultés principales résident d'une part dans la détermination du poids de chaque objectif et d'autre part dans l'expression des interactions entre les objectifs **[1]**.

### V.5. 2. Les méthodes de Pareto :

Basées sur le principe suivant : «Il existe un équilibre tel que nous ne pouvons pas améliorer un critère sans détériorer au moins un des autres critères». Elles sont des méthodes fondées sur les deux notions *la dominance au sens de Pareto* et *les fronts de Pareto*. La première utilisation de cette notion a été proposée en 1989 par Goldberg **[40]** qui montre l'efficacité de la dominance au sens de Pareto pour trouver un ensemble de compromis optimaux entre les objectifs.

*a) La dominance de Pareto :c*

Soit les deux solutions **A** et **B,** nous disons que la solution A *domine* la solution B si et seulement si : $\forall i \in \{1, 2, \dots . k\} : f_i(A) \leq f_i(B)$ et $\forall j \in \{1, 2, \dots . k\} : f_j(A) < f_j(B)$.

Avec $k$ le nombre des objectif et $f_i$ la ième fonction objectif.

Cette définition est pour un problème de minimisation, elle est utile aussi au problème de maximisation $(\geq, >)$. Si la solution A domine la solution B, nous disons que la solution *B est dominée par A* ou bien *A est non dominée par B* ou entre les deux nominations nous pouvons dire *A est la solution non dominée* **[80]**.

*b) L'optimum de Pareto :*

Définit comme suit : soit la solution **C**, nous disons que **C** est *une solution optimale au sens de Pareto* si elle n'est dominé par aucune autre solution. Toutes les solutions optimales au sens de Pareto, sont incomparables les uns par rapport aux autres. Elles sont appelées aussi les solutions *non inférieures* ou *non dominées*.

*c) Le front de Pareto :*

La représentation de solutions non dominées (Pareto optimales) dans l'espace des objectifs est appelée *le front de Pareto* . La figure **(Figure V.8)** montre un exemple d'un front de Pareto pour un problème de minimisation de deux objectifs. Les points en blanc représentent le front de Pareto.

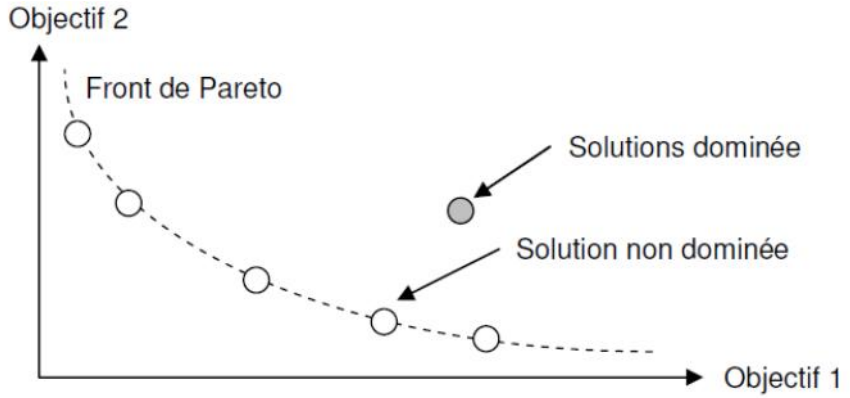



**Figure V.8 : Exemple d'un front de Pareto [46].**

## V.6.La sélection de caractéristiques par les algorithmes d'optimisation :

L'application des algorithmes d'optimisation dans la sélection de caractéristiques n'est pas récente, plusieurs travaux sont proposés dans ce contexte. Ils sont utilisés dans le cadre des approches *Wrapper* où la sélection de sous ensemble des solutions candidate est toujours liée à les performances d'un algorithme de classification par exemple un algorithme génétique avec le classificateur SVM pour construire l'hybride *GA-SVM*, un algorithme d'optimisation par l'essaim particulaire avec le SVM, *PSO-SVM*…etc. Quelques soit l'algorithme utilisé GA, PSO ou CLONALG pour la sélection de caractéristique, le principe de sélection est le même. Le principe est illustré dans la figure (**Figure V.9 )**

***i) Le codage:***

Le codage utilisé pour représenter les individus est généralement le codage binaire {0,1}. Le « 0 » utilisé pour indiquer l'absence de caractéristique dans la population courante et le « 1 » indique leur présence. Longueur de l'individu égal au nombre total de caractéristiques et la taille de la population posée par l'utilisateur dépend au problème à résoudre où chaque individu de la population représente une solution potentielle. Ce type de codage permet d'éviter de sélectionner une caractéristique plusieurs fois dans un même individu, c'est l'avantage principal qui rend ce codage très utilisable. Un des problèmes qui se pose est le nombre de « 1 » dans chaque individu. En général, la valeur de chaque caractéristique d'un individu est déterminée aléatoirement avec une probabilité de 0.5 pour que la valeur soit « 0 » ou « 1 », ce qui signifie qu'il y a toujours environ 50% de caractéristiques d'un individu qui ont la valeur « 1 » et ce qui conduit probablement à une sélection d'environ 50% des caractéristiques à la fin. Il y a plusieurs autres propositions dans ce contexte afin de minimiser le nombre de caractéristiques sélectionnées (proposé d'autres probabilités pour avoir un

nombre plus faible de « 1 »). Une autre possibilité pour minimiser le nombre de caractéristiques sélectionnées est d'intégrer cet objectif dans la fonction *fitness* **[33].**

***ii) La fonction de fitness :***

La fonction le fitness utilisée est toujours dépendante aux performances d'une classification. Elle peut être monoobjectif ou multiobjectif. Dans le cas monoobjectif nous trouvons la fonction égale à l'accuracy. Dans le cas multiobjectif, elle est liée à l'accuracy et des autres critères comme la taille de sous ensemble sélectionné.

♣ L'application des AGs dans le domaine de la sélection de caractéristiques a été débutée en 1993 par Ferri qui ont montré que l'utilisation des AGs est bien adaptée pour sélectionner dans un temps raisonnable les meilleures caractéristiques sur des ensembles de caractéristiques de taille moyenne de 20 à 49 caractéristique. Ensuite, Kudo et Sklansky ont montré la possibilité d'utiliser les AGs pour la sélection sur des ensembles de grande taille (50 caractéristiques et plus) **[46].** L'application des algorithmes génétiques pour l'analyse des données d'expression génique est très vaste, plusieurs approches sont proposées dans ce contexte **[17] [18] [31] [33] [99]** et plusieurs améliorations sont observées.

→ Cheng-San Yang et al. **[17]** ont proposé un hybride *GA-KNN* avec une adaptation de l'AG nommé *l'AG chaotique*. Dans cet algorithme, la probabilité de mutation est calculée à chaque génération en utilisant une carte chaotique nommée la carte de Gauss. L'objectif de cet algorithme est augmenté la diversité par une exploration chaotique.

- Sultan H.-A. et Mohammed E. **[99]** ont proposé un système hybride pour classer le cancer GA-DT. Dans ce système, un algorithme GA standard est utilisé pour explorer l'espace de recherche et les arbres de décision (DT) utilisés pour évaluer les résultats de cette recherche.
- EL AKADI, A et al. **[31]** ont proposé aussi un hybride GA, MRMR et SVM, dans un premier temps le MRMR est utilisé pour un filtrage c'est-à-dire éliminer les bruits et la redondance. Ensuite, l'hybride GA-SVM utilisé pour sélectionner les gènes les plus discriminants.

Nous remarquons après la synthèse des travaux qui utilisent les GAs pour découvrir les biomarqueurs que toujours le critère de stabilité est ignoré et les hybrides qui appliquent un filtrage avant l'optimisation fournis des résultats plus performants au niveau d'accuracy et le nombre de gènes sélectionné.

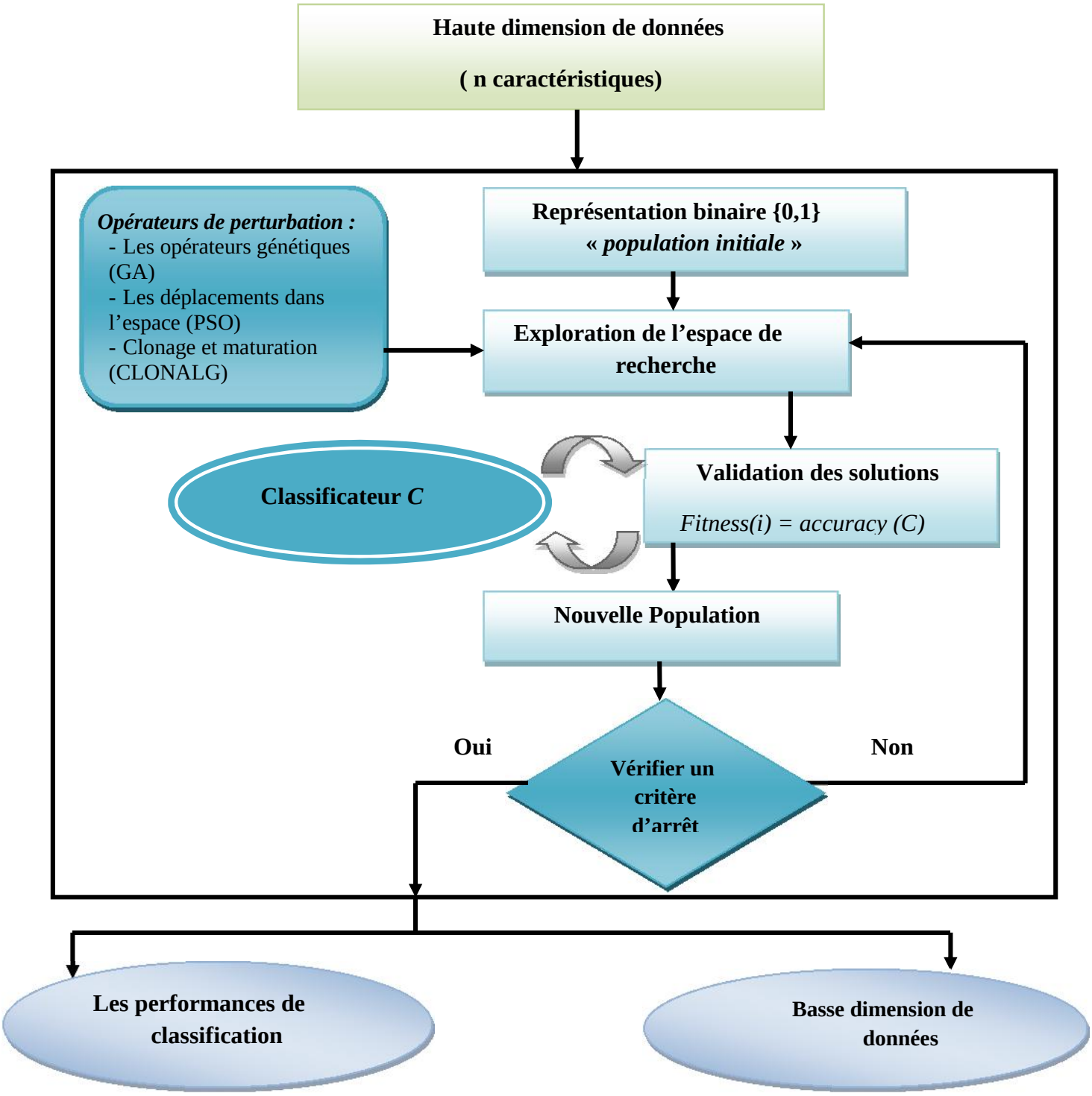



**Figure V.9 : Le processus de sélection par les algorithmes d'optimisation**

♣ Les PSOs sont récemment utilisées dans le domaine de la sélection de caractéristique, *PSO for feature selection* débutée par l'utilisation d'un PSO binaire (sbPSO) **[29]**, où la dimension de l'essaim égal au nombre total de caractéristique et la valeur de la vitesse calculée est utilisée pour décider si la caractéristique est sélectionnée ou non. Ensuite, une amélioration (IBPSO) de la sbPSO a été proposée par Chuang et al. **[19],** l'approche IBPSO

propose de réinitialiser le $X_{vbest}$ d'une particule quand elle est piégée dans un optimum local pour un nombre prédéfini des itérations consécutives. Ils espèrent que par cette adaptation, les particules vont continuer à chercher d'autres positions ce qui permet d'obtenir de meilleurs résultats.

L'application de PSO pour l'analyse des données biologiques et l'extraction des données les plus informatives est récente. Plusieurs travaux sont réalisés dans ce contexte **[9] [11] [32] [33],** une grande amélioration dans les résultats a été remarquée.

→ Enrique Alba et al**.[33]** ont proposés une version adaptée de la PSO nommée la PSO géométrique (GPSO). La question clé de la GPSO est le concept de mouvement des particules. Dans cette approche, à la place de la notion de vitesse ajoutée à la position, un opérateur nommé *three-parent mask-based crossover* (3PMBCX) est appliqué sur chaque particule afin de la déplacer dans l'espace. L'opérateur 3PMBCX est un opérateur de croisement entre trois parents en utilisant un masque de la même taille de parents, les parents dans ce cas sont : la position courante de la particule, la meilleure position global de la particule et la meilleure position trouvée historiquement.

→ Emmanuel Martinez et al. **[32]** ont proposés une adaptation de la sbPSO nommée cuPSO, dans cette méthode au lieu de mettre à jour les positions de toutes les particules à chaque itération, la cuPSO mettre à jour seulement les positions d'un sous ensemble de particules.

→ Barnali Sahu et Debahuti Mishra **[9]** ont proposés une nouvelle approche basée sur le sbPSO (K-means -SNR-PSO). L'approche proposée est divisée en deux stages. Dans le premier l'ensemble de données est regroupé utilisant l'algorithme K-means , le score SNR est appliqué pour classer les gènes dans chaque cluster. Les gènes ayant un score élevé de chaque cluster sont rassemblés et un sous-ensemble de nouveaux gènes est généré. Le deuxième stage, consiste à introduire le nouveau sous ensemble de gènes comme une population initiale de l'algorithme PSO.

Cette approche parmi les approches qui appliquent un filtrage avant une étape d'optimisation, ces approches fournies des résultats avec hautes performances mais toujours la stabilité est ignorée. Ainsi nous remarquons que les résultats obtenus par les PSO sont mieux que plusieurs autres algorithmes d'optimisation.

♣ L'application de CLONALG pour la sélection de caractéristiques est aussi récente, plusieurs approches et méthodes sont développées **[13] [57] [72] [74]**.

→ Jmal. Y et al. **[57]** ont proposé un système hybrides CLONALG-SVM dans ce système le CLONALG est l'algorithme de recherche et le classificateur SVM comme un outil d'évaluation. Le système proposé est appliqué sur des données microarray et montre leur efficacité pour trouver la bonne solution.

→ Boyun Z**.[13]** est proposé un système de sélection de caractéristiques qui est utilisé *l'approximation de Markov Blanket* pour sélectionner les caractéristiques pertinente, le réseaux bayésien pour l'apprentissage et le CLONALG comme une procédure de recherche. Cette approche a été fournie des résultats très performants par apport à plusieurs autres approches proposées dans ce contexte.

Finalement, nous pouvons déduire quelques points forts et faibles concernant l'application des algorithmes d'optimisation dans la sélection de caractéristiques et les résumés dans le tableau suivant (**Table V.1)**:

| Les points forts | Les points faibles |
|---|---|
| - Les performances augmentées : la pluparts des travaux montrent l'augmentation des performances avec ce type des algorithmes<br>- La taille de la signature biologie dans le cas d'analyse des données d'expression génique est minimale<br>- La facilité d'hybridation GA-SVM, GA-PSO, GA-DT, PSO-KNN…etc. | - Lente, pour chaque individu dans la population une évaluation de leur fitness est nécessaire<br>- La plupart des approches ne prend pas en considération le critère de la stabilité |

**Table V.1 : Les points forts et les faibles pour la sélection de caractéristiques par les algorithmes d'optimisation**

## V.7. Conclusion :

Ce chapitre avait pour objectif d'introduire dans un premier temps, les trois algorithmes d'optimisation que nous avons utilisé dans notre travail, les algorithmes génétiques, l'optimisation par l'essaim particulaire et l'algorithme de la sélection clonale. Pour chacun des algorithmes, nous avons détaillé le principe de base avec une simulation algorithmique de ce principe pour terminer avec quelques avantages et inconvénients connus pour chacun. Puis un aperçu sur l'optimisation multiobjectif a été introduit où nous avons expliqué le principe de l'optimisation multiobjectif et les méthodes proposées dans la littérature pour résoudre ce type de problèmes. La fin de ce chapitre a été consacrée à l'application des algorithmes d'optimisation dans le domaine la sélection de caractéristiques.

Dans la suite, nous s'inspirons de ces techniques et les techniques précédentes où nous proposons une nouvelle approche de sélection de caractéristiques pour découvrir les biomarqueurs dans le but de limiter les inconvénients des méthodes existantes pour les deux critères, la performance et la stabilité

## VI.1.Introduction :

Comme nous avons déjà vu, les approches de sélection de caractéristiques sont principalement de trois types : filter, wrapper et embedded, chacune possède des avantages et des limitations. Pour les approches de type filter, elles sont caractérisées par leur rapidité mais présentent des limitations comme l'indépendance au classificateur et l'ignorance des interactions entre les caractéristiques dans le cas où la sélection est univariée. Les approches de type wrapper soufrent de leurs complexité très élevée ainsi qu'une dépendance aux classificateurs utilisés pour l'évaluation mais elles permettent d'obtenir des résultats de haute performances. Tandis que, les approches de types embedded, sont des approches intermédiaire entre la rapidité (plus rapide que les approches wrapper) et la dépendance au classificateur utilisé pour l'évaluation, mais elles souffrent toujours de la complexité. La plupart des algorithmes de sélection proposés dans la littérature pour analyser les données d'expression génique sont caractérisés par leur *instabilité (*la plupart du temps cette notion est ignorée*)*, plusieurs facteurs provoquent cette limitation mais le facteur principal est « *la grande dimensionnalité avec un petit nombre des échantillons* ». Afin de résoudre ce problème deux approches sont proposées, basée ensemble et basée groupe. La première consiste à réaliser un sous échantillonnage de la base initiale afin d'obtenir des sous bases perturbées (échantillons), un algorithme de sélection est appliqué sur chacune de ces bases ensuite un mesure de similarité entre les sous-ensembles sélectionnés est calculé. La deuxième a le même principe que la première, mais la différence est qu'au lieu d'utiliser un sous échantillonnage de la base initiale, nous pouvons utiliser un clustering, les sous bases obtenues sont des groupes (clusters) pas des échantillons aléatoire.

L'utilisation des algorithmes d'optimisation pour la sélection est très fréquente, ils sont utilisés comme des approches wrapper où l'algorithme joue un rôle d'un moteur de recherche (dans l'espace des solutions) et le classificateur lié à cet algorithme comme un outil d'évaluation, malgré leur complexité les résultats obtenus avec ces algorithmes sont remarquable (performance très élevée). Dans ce contexte, plusieurs systèmes hybrides sont proposés GA/SVM, PSO/SVM, CLONALG/SVM…etc.

Dans le but d'exploiter les avantages des approches filter, wrapper et embedded, maximiser la stabilité et utiliser les hybrides *algorithmes d'optimisation/ système de classification* grâce à leur performances, nous proposerons une nouvelle approche composées de trois étapes pour sélectionner et découvrir les biomarqueurs.

## VI.2. La contribution :

L'approche que nous avons proposée vise à sélectionner une signature biologique qui représente le bon biomarqueur caractérisée par leur taille minimale, exactitude maximale et une stabilité remarquable. Notre approche est composée de trois étapes séquentielles clustering, filtrage et optimisation.

### VI.2. 1. Le processus général de sélection :

Soit un ensemble **G** = $g_1; g_2; \dots ; g_n$ composé de *n* gènes et un ensemble d'apprentissage $X = x_1; x_2; \dots ; x_n$ composé de m échantillons (samples) où chaque $x_i = (g_{i1}; g_{i2}; \dots ; g_{in})$ représente l'*ième* échantillon c'est un vecteur dont les composantes sont les valeurs qui représentent les niveaux d'expression des *n* gènes dans cet échantillon ($g_{ij}$: $le\ niveau\ d'expression\ de\ j\ ième\ gène\ dans\ l'ième\ échantillon$), nous avons aussi un ensemble des classes (label) Y = $y_1; y_2; \dots ; y_n$. Le processus de sélection se déroule comme suit :

1. *La première étape*, les gènes sont *regroupés (clustering)* en utilisant le concept de l'approximation d'une couverture de Markov basé sur les deux métriques de la théorie de l'information, l'information mutuelle et l'entropie conjointe.
2. *La deuxième étape*, étape intermédiaire vise à diminuer la dimension de l'espace, c'est l'étape de *filtrage*, nous avons utilisé dans cette étape une méthode de type filter ou embedded sur chaque cluster et les gènes sélectionnés de chaque cluster sont les gènes non redondants et les plus pertinents.
3. *L'étape finale*, l'entrée de cette étape est le nombre des gènes réduit, un système wrapper est lancé constitué de trois algorithmes d'optimisation GA, CLONALG et PSO avec les caractéristiques suivantes :
   - La taille de la population égale au nombre de groupes,
   - les gènes sélectionnés de chaque groupe sont les gènes présentés dans chaque individu de la population,
   - la fonction de fitness est multiobjectif.

   Ces trois algorithmes d'optimisation sont en coopération, la nature de cette coopération c'est une *migration* des meilleures solutions.

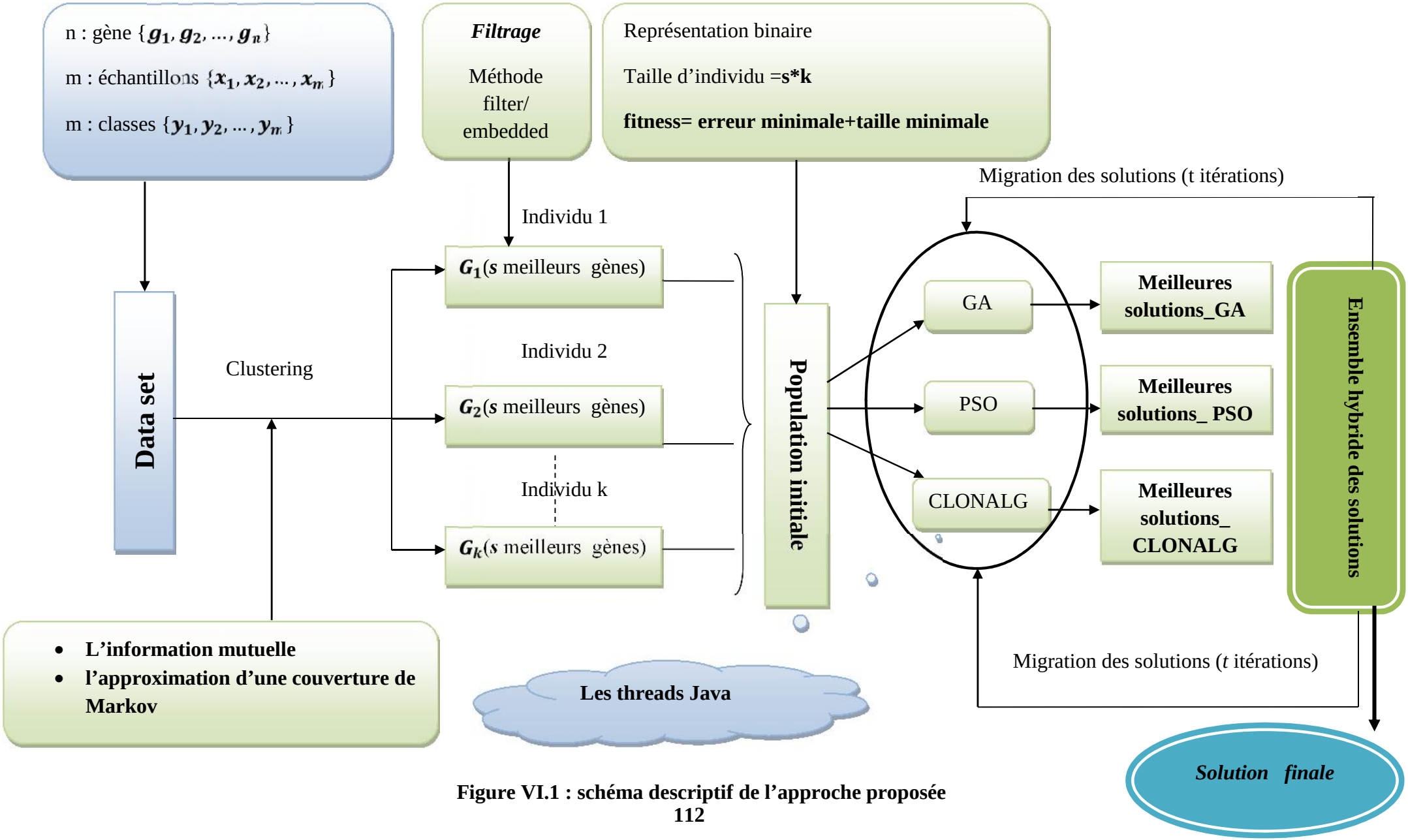



**Figure VI.1 : schéma descriptif de l'approche proposée**

### VI.2. 2. La première étape : le clustering :

Cette étape vise à grouper les données avec une similarité maximale au sein d'un même groupe et dissimilarité maximale entre les groupes, dans le contexte de notre travail, deux gènes sont similaires, s'ils ont la même pertinence au problème à résoudre. Dans ce cadre, nous avons proposé un nouvel algorithme pour grouper les gènes, l'algorithme est non supervisé c'est-à-dire ne demande pas le nombre de cluster comme paramètre initial pour distribuer les gènes sur les clusters selon certain métriques. L'algorithme commence par un seul groupe contient un seul gène (centre), ensuite un processus itératif est lancé pour évaluer la pertinence des gènes avec le centre de ce groupe, par conséquent une insertion à ce groupe ou création d'un nouveau groupe est réalisé. Les deux objectifs principaux de cette étape sont

- Éliminer les redondances dans les étapes suivantes.
- Maximiser la stabilité de la sélection.

Ce clustering est basé sur le concept de *l'approximation d'une couverture de Markov* en utilisant les deux métriques de la théorie de l'information qui sont l'information mutuelle et l'entropie conjoint.

#### a) L'approximation d'une couverture de Markov :

En 1966 Koller et Sahami ont proposé une technique basée sur l'entropie, connu sous le nom la Couverture de Markov (en anglais Markov Blanket) pour identifier les caractéristiques redondantes et inutiles pour les supprimées, ce concept connu une utilisation très fréquente dans le domaine de la sélection des caractéristiques grâce à leur utilité pour trouver des bon résultats. Il est défini comme suit :

- Pour chaque caractéristique/gène $\boldsymbol{g_i}$, soit **M** un sous ensemble de **G** ne contient pas $\boldsymbol{g_i}$,**C** le vecteur de classes, **M** couverture de Markov de $\boldsymbol{g_i}$ *si et seulement si* : $\boldsymbol{g_i}$ indépendante conditionnellement de $\boldsymbol{G - M - \{g_i\}}$ sachant que **M**, signifie formellement :

$$\boldsymbol{P(G - M - \{g_i\}, C | g_i, M) = P(G - M - \{g_i\}, C | M)}$$

→ Deux caractéristiques A et B sont indépendantes conditionnellement sachant que X**,** si $\mathbf{P(A|X,B) = P(A|X)}$**,** c'est-à-dire B ne donne aucune information sur A au-delà de ce qui est déjà dans X. Si une caractéristique $g_i$ a une couverture de Markov **M** au sein de sous-ensemble sélectionné actuellement, il suggère que $g_i$ ne donne pas plus d'information au-delà de **M** sur C et d'autres caractéristiques sélectionnées, donc $g_i$ pourrait être éliminée **[111]**.

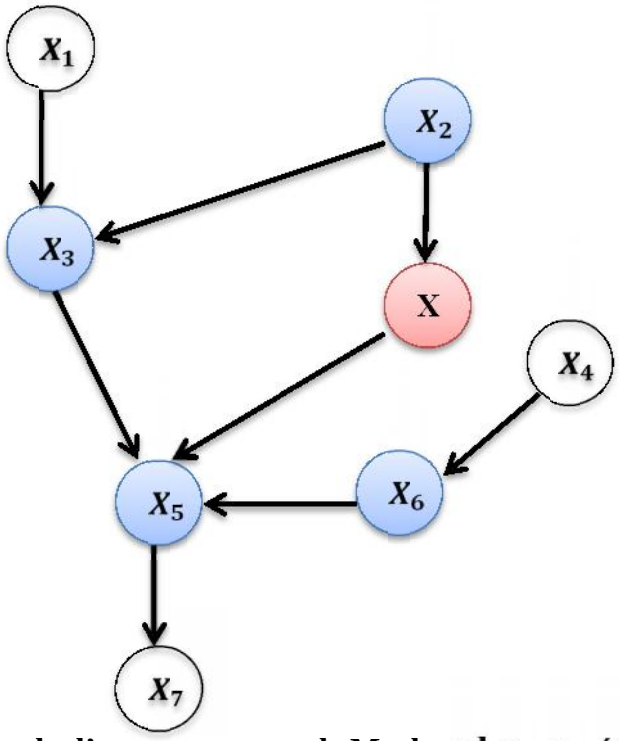



**Figure VI.2 : Exemple d'une couverture de Markov dans un réseau Bayésien, La couverture de Markov de X, $M(X)=\{X_2, X_3, X_5, X_6\}$.**

Toutefois, étant donné la complexité de calcul pour déterminer l'indépendance conditionnelle de caractéristiques est généralement très élevée, Yu et Liu **[111]** ont proposé l'utilisation d'une seule caractéristique pour approximer la couverture de Markov de $g_i$:

- *L'approximation d'une couverture de Markov :*

L'approximation d'une couverture de Markov que nous avons manipulée dans notre travail basée sur un nouveau coefficient nommé le *coefficient de corrélation de l'information* qui est défini comme une normalisation de l'information mutuelle par l'entropie conjointe comme suit :

$$ICC(X,Y) = \frac{MI(X,Y)}{H(X,Y)} \quad \textbf{... (VI.1)}$$

Avec $MI(X,Y)$ est l'information mutuelle qui permet de mesurer la dépendance entre deux variables (X et Y). Estimé empiriquement par :

$$MI(X,Y) = H(X) + H(Y) - H(X,Y) \quad \textbf{... (VI.2)}$$

Où H(X) et H(Y) sont des entropies et H (Y, X) est l'entropie conjointe entre X et Y.

→ *L'entropie* permet de mesurer l'incertitude dans une variable aléatoire X avec une probabilité de distribution marginale p(x) comme suit :

$$H(X) = -\sum\sum p(x) log\, p(x) \quad \textbf{... (VI.3)}$$

→ Et $H(X,Y)$ *l'entropie conjointe* qui permet de mesurer l'information contenue dans un système de deux variables aléatoires par exemple X et Y, elle est calculée en utilisant la probabilité jointe **p(x,y)** comme suit :

$$H(X,Y) = -\sum\sum p(x,y)\log p(x,y) \quad \text{... (VI.4)}$$

Soit : $icc(g_i, C)$le coefficient de corrélation qui représente la pertinence entre la caractéristique/gène $g_i$ et le vecteur des classes C,

$icc(g_j, C)$ le coefficient de corrélation qui représente la pertinence entre la caractéristique/gène $g_j$ et le vecteur des classes C ,

$icc(g_i, g_j)$ le coefficient de corrélation qui représente la pertinence entre les deux caractéristiques/gènes $g_i$ et $g_j$.

$g_i$ est une approximation d'une couverture de Markov de $g_j$ si et seulement si **[110]:**

- $icc(g_i, C) \geq icc(g_j, C)$
- $icc(g_i, g_j) \geq icc(g_j, C)$

**... (VI.5)**

Si $g_i$ est une approximation d'une couverture de Markov de $g_j$ alors ils ont la même pertinence au problème à résoudre, pour la sélection des caractéristiques nous pouvons éliminer $g_j$ et dans le cas d'un clustering nous pouvons les regrouper dans le même groupe **[110].**

### b) L'algorithme de clustering :

Notre algorithme de clustering est simple, constitué de deux étapes, comme il est illustré dans le pseudo code suivant (**Figure VI.3 )**, dans la première les gènes sont ordonnés d'une manière descendante selon *leurs information mutuelle avec le vecteur des classes* en calculant pour chaque gène leur information mutuelle avec le vecteur des classes dont l'objectif de déterminer les gènes qui ont une couverture de Markov à l'avance et les affecter comme des centres des clusters, la deuxième étape vise à grouper les gènes *selon le principe de l'approximation d'une couverture de Markov* dans le même groupes ou dans des groupes différents.

**Sortie** :

k groupes $\{Gr_1, Gr_2, \dots, Gr_k\}$, k centres Cr (vecteur des centres)

1. Pour i=1 à n faire
   a) Calculer l'information mutuelle entre chaque vecteur gène $g_i$ et le vecteur de classe

      **MI (i)= MI ($g_i$, C)** … **(**selon la formule **VI.2)**

   b) Ordonner d'une manière descendante les valeurs de MI, nous obtenons la liste **G'** des gènes ordonnés selon leur information mutuelle avec le vecteur de classes

   Fin pour

2. Assigner le premier gène de la liste **G'** au premier groupe $Gr_1$, et le mettre comme centre de ce groupe $Cr(1) = G'(1)$ .
3. k=1
4. Pour i=2 : longueur (G')
   a) Pour chaque gène de la liste G', s'il y a un groupe $Gr_k$ où le centre de ce groupe est leur approximation d'une couverture de Markov alors regrouper G'(i) à $Gr_k$ **…** (selon les formules **VI.1 et VI.5)**
   b) Sinon mettre G'(i) dans un nouveau groupe et le mettre comme leur centre

   c) k=k+1

   d) $Cr(k) = G'(i)$

   Fin pour

5. Donner les k groupes $\{Gr_1, Gr_2, \dots, Gr_k\}$ et vecteur Cr de k centres

**Figure VI.3: L'algorithme de Clustering basée sur l'approximation d'une couverture de Markov**

### VI.2. 3. La deuxième étape : étape de filtrage :

Après la construction des groupes, l'étape suivante consiste à lancer un système de filtrage, en utilisant deux méthodes la première est de type Filter mRMR et la deuxième de type embedded SVM-RFE. Cette étape vise à réduire l'espace de recherche afin de diminuer la complexité. Les deux méthodes données des listes ordonnées de gènes selon leur pertinence,

avec une taille de signature précisée par l'utilisateur notée ***s***, ensuite nous sélectionnons les **s-top** gènes (sont les gènes les plus pertinents). Pour la méthode SVM-RFE l'algorithme est arrêté lorsque la taille de l'ensemble de données courant égal à ***s***. A la fin de cette étape, la dimensionnalité de l'espace diminué à ***n'=k*s*** gènes c'est-à-dire à ***(n'*100 /n )% ,*** seulement les gènes non redondants et pertinents sont restés.

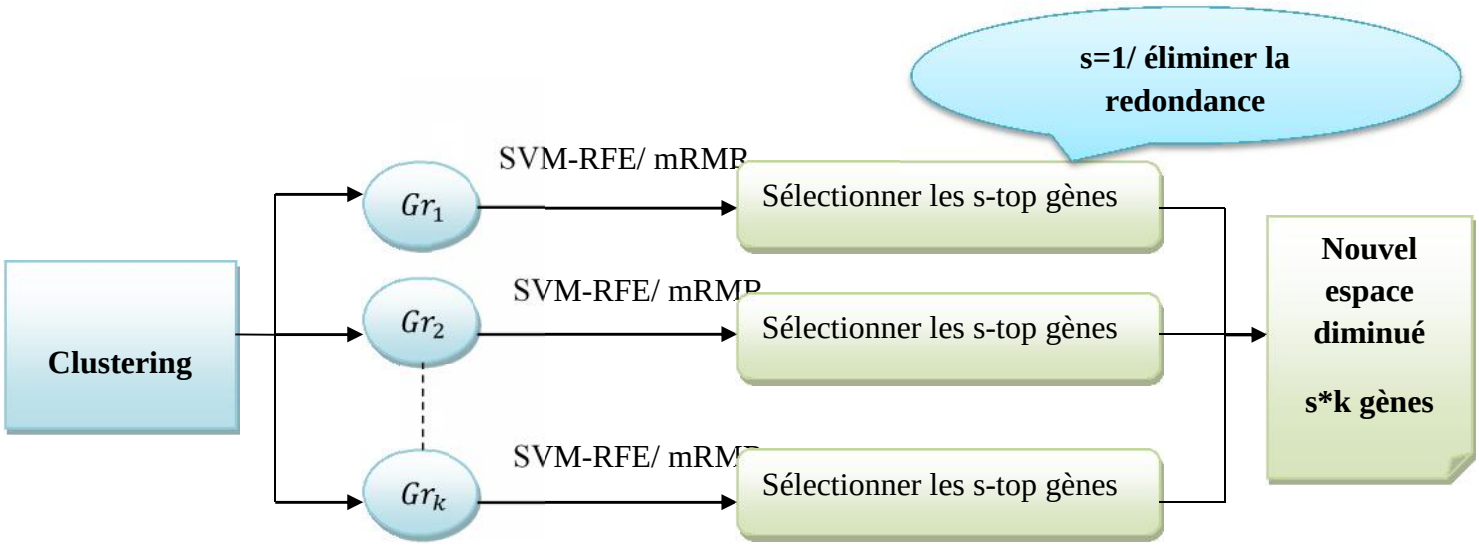



**Figure VI.4: Le processus de filtrage**

## VI.2. 4. La troisième étape : l'étape d'optimisation

L'objectif principal de cette étape consiste à trouver la bonne solution (le bon compromis), dans notre contexte un sous ensemble minimal des gènes (biomarqueur) qui discrimine bien entre les classes des échantillons (l'exactitude de la classification). Cette étape nommée l'étape *d'optimisation* car elle utilise trois algorithmes d'optimisation : les algorithmes génétique, l'optimisation par essaim particulaire et l'algorithme de la sélection clonale. Ces trois algorithmes sont en coopération, à chaque itération un tampon de taille $\mathbf{3} \times \boldsymbol{k'}$ est construit contient seulement les **k'** meilleures solutions présentées par chaque algorithme. Nous avons utilisé ces trois algorithmes pour exploiter les avantages de chacun afin d'obtenir la solution la plus performante. Comme il est illustré dans la figure suivante **(Figure VI.5)** qui montre les résultats de l'erreur et la taille de la signature biologique obtenus pendant 300 itérations par chaque algorithme d'optimisation sur la base côlon (concernant le cancer du côlon), nous pouvons constater que les résultats sont moins performants par apport aux résultats obtenus par le système d'optimisation coopératif **(GA/ CLONALG /PSO)** dans *la section VI.3.2*.

| La base | GA | CLONALG | PSO |
|---|---|---|---|
| Côlon | Erreur/Nombre de gènes | | |
| | 0.11/02 | 0.14/01 | 0.09/10 |

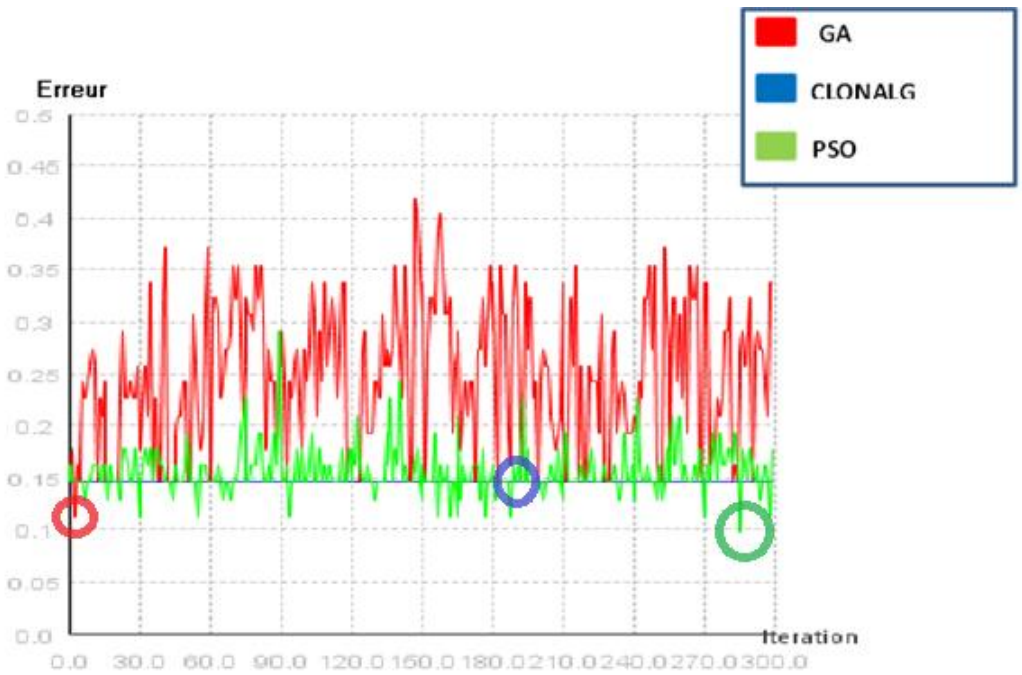



**Figure VI.5: Les résultats obtenus par chaque un des algorithmes (GA, CLONALG, PSO) sur la base concernant le cancer du côlon**

Dans la pratique, nous avons utilisé la notion des *threads Java*. Où chaque algorithme est affecté à un thread dont l'objectif est de garantir la *rapidité* et le *parallélisme d'exécution*. Le thread Java est défini comme un processus léger ou bien une unité d'exécution autonome qui peut effectuer des tâches en parallèle avec d'autres threads. Les avantages des processus légers par rapport aux processus système sont : la rapidité de lancement et d'exécution, le partage des ressources système du processus englobant.

Nous avons programmé les deux étapes précédentes en Matlab et cette dernière étape en Java, pour assurer la communication entre ces deux parties nous avons activé le serveur qui permet de lier Matlab et Java, cette activation réalisée à l'aide de la commande « *com.jamal.server.MatlabServer* **».** Avant l'utilisation de cette commande il faut réaliser les deux opérations suivantes :

- Ajouter le fichier jamal.jar à la partie codée en Java (*External JARs*).

- Modifier le fichier *classpath.txt* en Matlab**.** La modification consiste à ajouter le chemin du jamal.jar à la dernière ligne de ce fichier. La communication est commencée lorsque nous obtenons le message « *Jamal::MatlabServer is ready* »

### a) Le codage et l'initialisation

Le codage que nous avons utilisé dans notre travail est le codage binaire classique. Comme déjà vu, ce codage consiste à coder chaque solution (individu) possible par une chaine binaire de taille égale au nombre total de gènes sélectionnés dans l'étape précédente ***(k*s)***. Un gène d'indice **i** a pour valeur « 1 » si l'*ième* gène de l'ensemble de départ est présent dans le sous-ensemble courant et « 0 » dans le cas contraire. Nous avons utilisé ce codage grâce à sa simplicité dans la programmation et il permet d'éviter la sélection d'un gène plusieurs fois dans un même individu (pas de redondance).

### b) La fonction de fitness

Comme nous avons déjà vu, l'objectif de notre approche proposée est de trouver un sous-ensemble minimal des gènes qui discrimine bien entre les échantillons à étudier (exactitude maximale), à cette raison la fonction de fitness que nous avons utilisé est multiobjectif, le premier objectif consiste à minimiser l'erreur d'une classification et le deuxième consiste à minimiser le nombre de gènes sélectionnés dans le sous ensemble final (signature biologique). Les classificateurs que nous avons utilisés pour valider l'exactitude de la discrimination sont : KNN, SVM et Bayes. Pour évaluer la classification, nous avons utilisé la technique de la validation croisée *k-fold* (*K-foldCV*) avec k=10 grâce à leur efficacité dans le cas où le nombre des échantillons est petit, à chaque itération l'apprentissage se fait sur *neuf* échantillons et le test sur l'échantillon qui reste.

Pour résoudre la fonction de fitness nous avons utilisé les méthodes agrégées : « *la somme pondérée* », qui transforment un problème multiobjectif en un problème simple avec un seul objectif. En affectant à chacun des objectifs un coefficient de poids. Ce coefficient représente l'importance attribuée à l'objectif, le problème devient sous la forme :

$$\sum_{i=1}^{n} w_i \square f_i$$

Avec **n** le nombre d'objectifs (n=2), $\mathbf{w_i}$ le poids affecté au *ième* objectif ($f_i$) et $\sum_{i=1}^{n} \mathbf{w_i} = \mathbf{1}$. La fonction de fitness utilisée dans notre travail est définie comme suit :

$$f = \alpha * \text{l'erreur de la classification} + \beta * \text{nombre de gènes sélectionnés} \quad \text{... (VI.6)}$$

$\alpha$ et $\beta$ sont des poids de pondération

### VI.2. 4.1. Les algorithmes génétiques

L'algorithme génétique utilisé dans notre approche est l'algorithme de base pour la sélection des caractéristiques avec une représentation binaire des chromosomes et une fonction de fitness multiobjectif (deux objectifs à minimiser). La population initiale subie à un ensemble des opérateurs génétiques, *sélection à roulette*, *mutation* et le *croisement à un point*. Les points de croisement et de mutation sont sélectionnés d'une manière aléatoire, comme suit :

Si **rand()** < $p_m$ ($p_c$) alors

*Point_à_muter (le point de croisement)* = round (rand*lg)

Où lg : longueur d'un individu

Sinon rien à faire

… **(VI.7)**

Où $p_m$ et $p_c$ sont les probabilités de mutation et de croisement respectivement. La population initiale dans un premier temps est constituée des individus résultants dans la deuxième étape, à partir de la deuxième évolution la population initiale sera constituée des solutions migrées. Les solutions migrées sont les $\mathbf{k'}$ meilleures solutions et l'algorithme accepte seulement **k** solutions si la taille de la population en entrée $\mathbf{k'' > k}$.

**Entrées**

**k** : la taille de la population (c'est la taille acceptée par l'algorithme)

$\mathbf{p_m}$ : la probabilité de mutation,

$\mathbf{p_c}$ : la probabilité de croisement,

**Num_evol** : le nombre maximal des évolutions (le nombre des migrations),

**itr :** le nombre de générations de l'algorithmes

$\mathbf{k'}$ : le nombre de meilleures solutions à migrer $(\mathbf{k' < k})$

**Sortie**

Les $\mathbf{k'}$ meilleures solutions à migrer,

**Sol** : meilleure solution

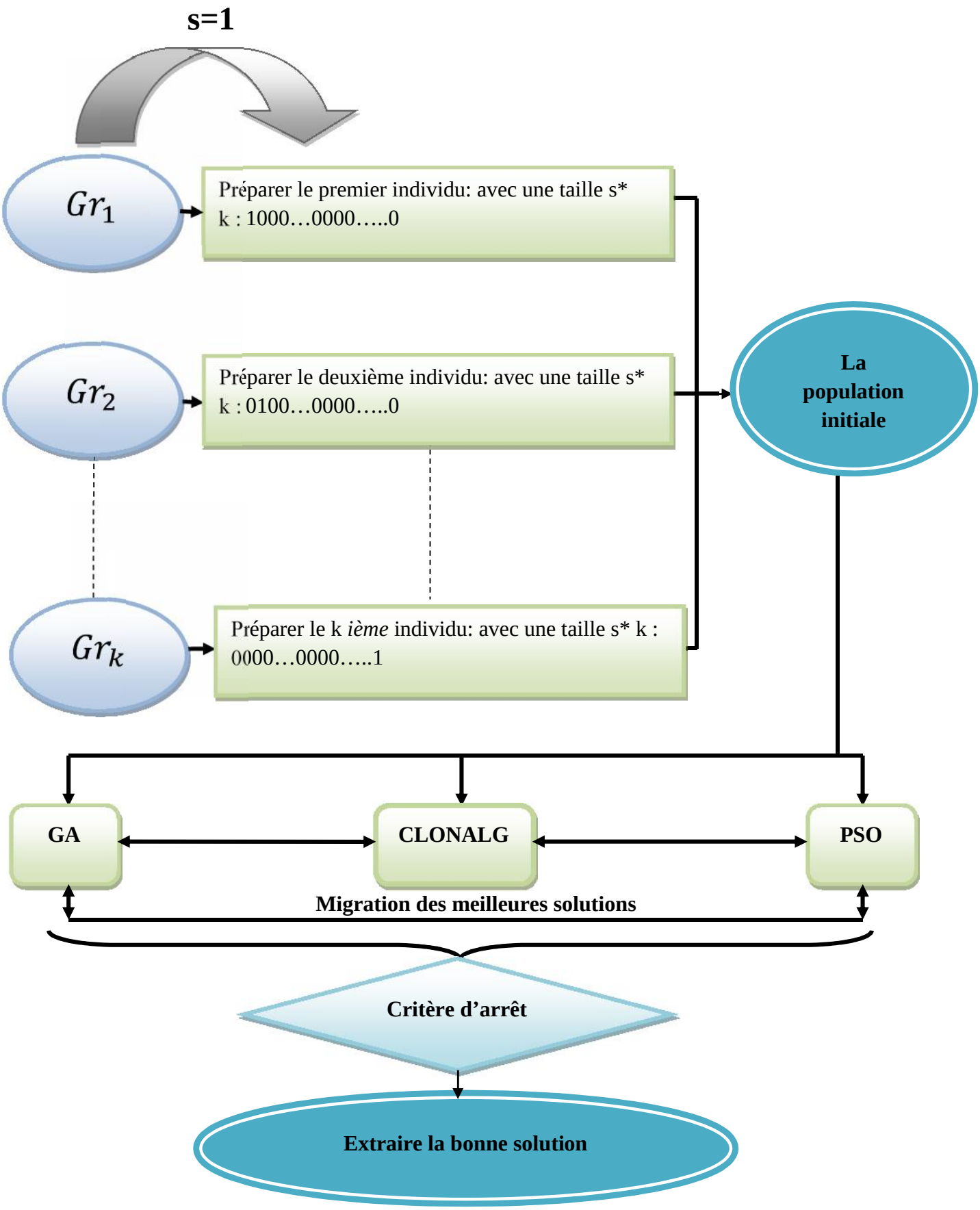



**Figure VI.6: La migration des meilleures solutions entre les systèmes d'optimisation**

**Pour** j=1 à Num_evol **faire**

**a)** Calculer la taille de la population courante $k''$

**Si** $k'' > k$ alors

Sélectionnr les **k** meilleures solutions

**Sinon** prendre tous les individus courants

**b)** Pour i=1 à itr faire

**i)** Evaluer la population …**(** selon la formule **VI.6)**

**ii)** Appliquer les opérateurs génétiques sur la population courante (de taille k)

- La sélection à roulette
- Le croisement avec une probabilité $p_c$ **( à un point)**
- La mutation avec une probabilité $p_m$ **(aléatoire à un point) … (**selon la formule **VI.7)**

**iii)** Nouvelle population, retournr à l'étape (i)

Fin pour (b)

**c)** Evaluer la population courante et sélectionner les $k'$ meilleures solutions et les migrées … **(** selon la formule **VI.6)**

**d)** Extraire la meilleure solution **Sol**

**e)** Récupérer les meilleures solutions produites par les autres algorithmes (migrées)

**f)** Hybrider l'ensemble des solutions avec les $k'$ meilleures solutions de GA

**Fin pour**

**Figure VI.7: L'algorithme génétique pour la sélection de gènes**

### VI.2. 4.2. L'algorithme de la sélection clonale

L'algorithme de la sélection clonale utilisé dans notre travail c'est l'algorithme **CLONALG.** L'algorithme possède trois opérateurs, la sélection, le clonage et la maturation, avec ces opérateurs nous pouvons évoluer vers la solution la plus performante. Le clonage des $k'$ meilleures solutions sélectionnées se fait proportionnellement à l'affinité, le nombre de clones **C** générés est défini comme suit:

$C = \sum_{i=1,k'} round(\beta \cdot N / (i \cdot fat))$ / **fat :** un facteur de multiplication **…(VI.8)**

La maturation de **C** clones est réalisée avec un taux inversement proportionnel à l'affinité, le facteur de maturation est calculé pour *l'ième* solution comme suit :

$$facteur\_maturation = \frac{Fitness(i)}{5} * exp(-0.01 * Fitness(i)) \quad \text{…(VI.9)}$$

Ce facteur est utilisé ensuite pour muter les C clones, les points de mutation sont sélectionnés d'une manière aléatoire comme suit :

**Si rand () <** $facteur\_maturation$ **lors**

**Point_à_muter = round (rand*lg)/ lg : longueur d'un clone** **…(VI.10)**

**Sinon rien à faire**

Comme les algorithmes génétiques, la population initiale dans la première évolution constituée des individus résultants dans la deuxième étape, à partir de la deuxième évolution la population initiale sera constituée des solutions migrées.

**Entrées**

**k** : la taille de la population (la taille acceptée par l'algorithme)

$\mathbf{s_r}$ : Le nombre d'éléments pires pour les éliminer,

**Num_evol** : le nombre maximal des évolutions (le nombre des migrations),

**itr :** le nombre de générations de l'algorithmes

$\mathbf{k'}$ : le nombre de meilleures solutions à sélectionner et migrer ($\mathbf{k'} < k$)

**Sortie**

Les $\mathbf{k'}$ meilleures solutions à migrer,

**Sol** : meilleure solution

**Pour** j=1 à Num_evol **faire**

**a)** Calculer la taille de la population courante $\mathbf{k''}$

**Si** $\mathbf{k''}$**>k** alors

Sélectionner les **k** meilleures solutions

**Sinon** prendre tous les individus courants

**b) Pour** i=1 à itr **faire**

**i)** Evaluer la nouvelle population et sélectionner $\mathbf{k'}$ meilleures solutions **…** **(**selon la formule **VI.6)**

**ii)** Cloner proportionnellement à l'affinité ***(C)*** **…(** selon la formule **VI.8)**

**iii)** Maturation proportionnelle inversement à l'affinité (facteur de maturation) **…(** selon les formules **VI.9** et **VI.10)**

**iv)** Réévaluer les clones maturés et sélectionner $\mathbf{k'}$ meilleures solutions et les migrées **…(** selon la formule **VI.6)**

**v)** Remplacer les premières solutions sélectionnées par les nouvelles solutions

**vi)** Sélectionner les $\mathbf{s_r}$ mauvaises solutions et les remplacer par des autres d'une manière aléatoire

**Fin pour**

**c)** Fournir la meilleure solution **Sol**

**d)** Récupérer les meilleures solutions produites par les autres algorithmes (migrées)

**e)** Hybrider l'ensemble des solutions avec les $\mathbf{k'}$ meilleures solutions de **CLONALG**

**Fin pour**

**Figure VI.8: L'algorithme CLONALG pour la sélection de gènes**

### VI.2. 4.3. L'optimisation par essaim particulaire

Dans notre travail, nous avons utilisé le PSO binaire standard pour la sélection des caractéristiques (The standard binary PSO ou bPSO). L'exploration de l'espace de recherche se fait par des mutations au niveau des particules. A chaque itération la nouvelle vitesse **v** calculée est utilisée pour compter le facteur de mutation ***S*** comme suit :

Si ($v_i < V_{min}$)

$v_{,i} = V_{min}$

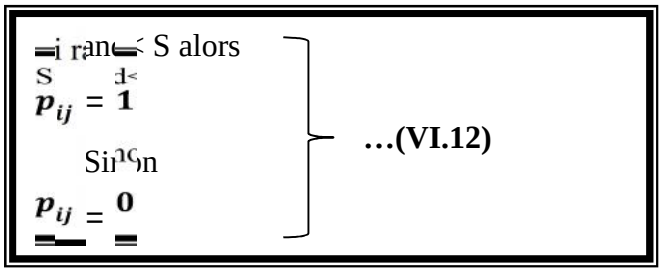
Si ($v_i > V_{max}$)

$v_i = V_{max}$

**S=1/1+exp** $(-v_i)$

**…(VI.11)**

Où $v_i$ la vitesse calculée pour l'ième particule, $V_{min}$ et $V_{max}$ représentent la vitesse minimale et maximale respectivement et la mutation réalisée selon la formule suivante :




**…(VI.12)**

Avec p : la particule, i : *l'ième* particule, j : la *jème* caractéristique. Les valeurs de $V_{min}$ et $V_{max}$ sont données par l'utilisateur, avec $V_{min} = -V_{max}$. Comme les algorithmes génétiques et la sélection clonale, la population initiale dans la première évolution constituée des individus de la deuxième phase, à partir de la deuxième évolution la population initiale constituée des solutions migrées.

**Entrées**

**k** : la taille de la population (la taille de la population acceptée par l'algorithme)

$c1$: le poids cognitif, **c2** : le poids social, **a** : Inertie

$V_{min}$,$V_{max}$ : La vitesse minimale et maximale.

**Num_evol** : le nombre maximal des évolutions (le nombre des migrations).

**itr :** le nombre de générations de l'algorithmes

$k'$ : le nombre de meilleures solutions à migrer ($k' < k$)

**Sortie**

Les $k'$ meilleures solutions à migrer,

**Sol** : meilleure solution

**Pour** j=1 à Num_evol **faire**

**a)** Calculer la taille de la population courante $k''$

**Si** $k''>k$ alors

Sélectionner les **k** meilleures solutions

**Sinon** prendre tous les individus

**b)** /* Initialisation */

***i)*** Définir le voisinage : la topologie étoile (L'optimum du voisinage= l'optimum global).

***ii)*** Initialiser la position et la vitesse des particules aléatoirement.

**c)** **Pour** i=1 à itr **faire**

***i)*** Pour chaque particule, calculer *local_best_fitness*…**(** selon la formule **VI.6)**

***ii)*** Mettre à jour local_best_position

***iii)*** Calculer global_best_fitness (minimum local_best_fitness)

***iv)*** Mettre à jour global_best_fitness

***v)*** Mettre à jour global_best_position

***vi)*** Calculer la vitesse $V_i$ … **(**selon la formule **V.1)**

***vii)*** Calculer le facteur de mutation **S …(** selon la formule **VI.11)**

***viii)*** Muter la particule **…(** selon la formule **VI.12)**

***ix)*** Calculer la nouvelle position $X_i$**…(** selon la formule **V.2)**

**Fin pour**

**d)** Evaluer la population finale et sélectionner les $\mathbf{k}'$ meilleures solutions et les migrées **… (** selon la formule **VI.6)**

**e)** Fournir la meilleure solution **Sol**

**f)** Récupérer les meilleures solutions produites par les autres algorithmes

**g)** Hybrider l'ensemble des solutions avec les $\mathbf{k}'$ meilleures solutions

**Fin pour**

**Figure VI.9: L'optimisation par l'essaim particulaire pour la sélection de gènes**

Pour extraire *la meilleure solution*, à chaque évolution de migration nous gardons la solution la plus bonne fournie entre les trois algorithmes d'optimisation, en comparant avec la solution de l'évolution précédente ; à la fin nous obtenons un vecteur des bonnes solutions trouvées pendant toutes les évolutions et immédiatement nous pouvons extraire la meilleure solution (solution finale).

## VI.2. 5. Mesure de la stabilité :

Notre approche de sélection permet d'obtenir une signature biologique (le bon biomarqueur) *robuste* c'est-à-dire quand nous réalisons des perturbations dans la base initiale

nous obtenons des signatures avec une similarité maximale. Afin de mesurer la stabilité nous avons suivi les étapes suivantes :

1. Une perturbation de l'ensemble de données initial, la perturbation consiste à supprimer des instances de façon aléatoire à partir d'un ensemble de données original afin de créer une ou plusieurs ensembles de données réduits.
2. Appliquer sur chacun des différents ensembles de données perturbés l'approche de sélection que nous avons proposé afin d'obtenir les différentes signatures biologiques.
3. La dernière étape consiste à mesurer la stabilité entre ces signatures par une mesure de similarité. Dans le cadre de notre travail, nous avons utilisé les indices de similarité **Dice**, **Jaccard** et **Tanimoto**

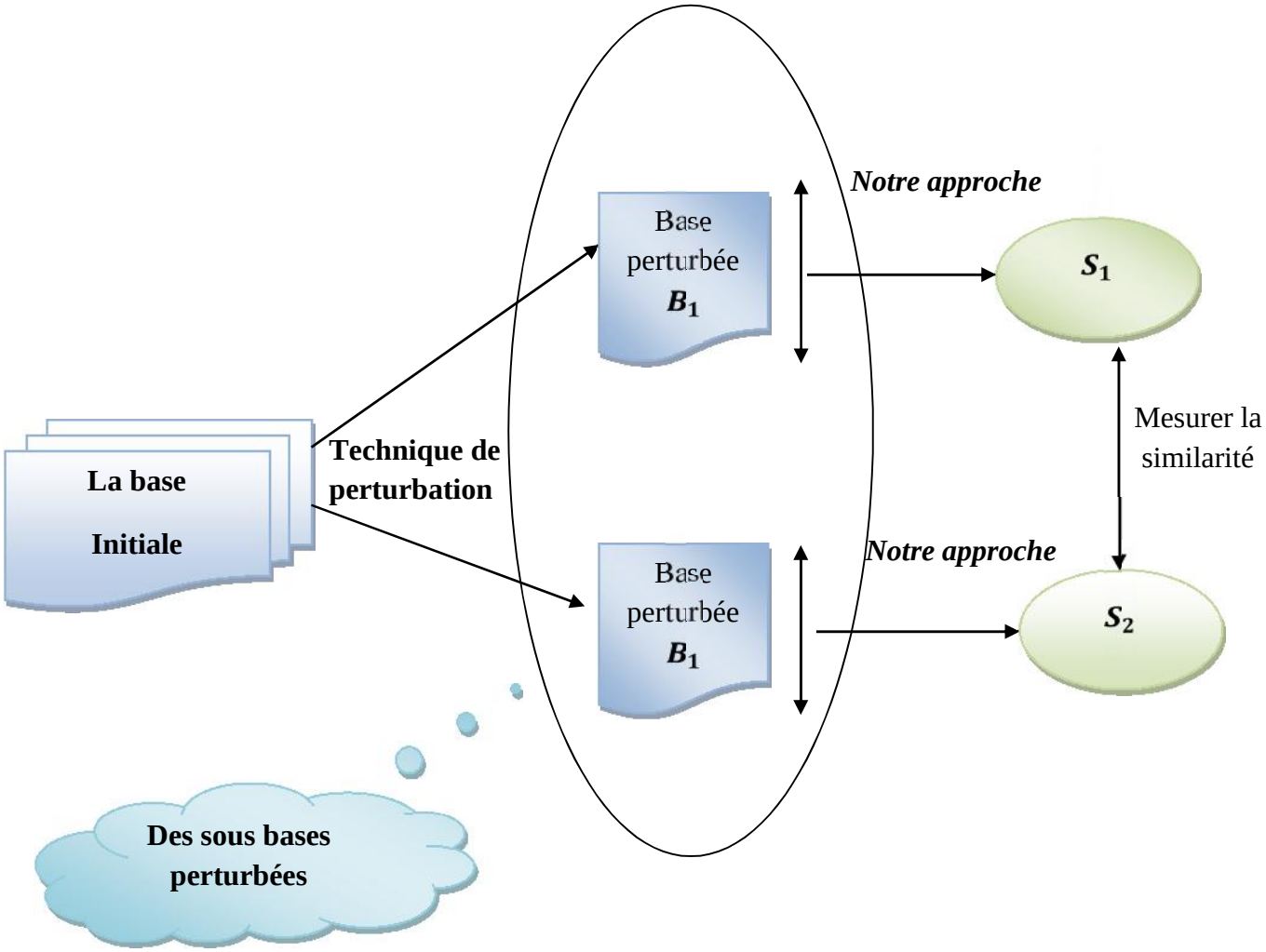



**Figure VI.10: Le principe pour mesurer la stabilité**

### VI.2. 5. 1. La construction des bases perturbées :

La technique de perturbation que nous avons utilisé dans notre travail c'est la technique basée sur *les chevauchements fixes* entre les sous bases perturbées (entre chaque deux bases), nous réalisons cette perturbation ***l*** fois avec un *chevauchement de 80%* comme suit :

**Entrée**

***X*** : la base de données initiale $X = \{x^1, x^2, \dots x^n\}$ avec m échantillons et n gènes.

C : vecteur des classes

***l*** : le nombre des perturbations.

**Sortie**

***sim*** : la stabilité (la moyenne des similarité calculées).

1. **Pour** i=1 à ***l*** **faire**
   a) Calculer la taille de chevauchement $m' = round(m \times 80)/100$
   b) Meme_base= Sélectionner aléatoirement $m'$ échantillons de différentes classes
   c) Reste_base1= Sélectionner *m''* parmi les *m-m'* échantillons restants
   d) Reste_base2= Constituer des échantillons restants *( (m-m' )- m'')*

   Reste_base1 ≠ Reste_base2
   e) Créer les deux bases perturbées avec un **chevauchement** entre eux de *m'échantillons*

   $B_1$=*[Meme_base ; Reste_base1]*, $B_2$=*[Meme_base ; Reste_base2]*
   f) Appliquer notre approche de sélection sur les deux bases $B_1$ et $B_2$, nous obtenons les deux signatures $S_1$ et $S_2$
   g) Calculer la similarité entre les deux signatures, $sim_i(S_1, S_2)$

   **Fin pour**

2. Calculer la moyenne des **stab …(** selon la formule **VI.13)**

**Figure VI.11: L'algorithme de mesure de la stabilité**

### VI.2. 5. 2. Le calcul de similarité :

Comme nous avons déjà dit, une mesure de stabilité nécessite une mesure de similarité entre deux sous ensemble des gènes sélectionnées. De nombreux indices de similarité existants dans la littérature, dans notre travail nous avons utilisé les trois indices : **Dice**, **Jaccard** et **Tanimoto**.

- **Le coefficient de Dice** : Le coefficient est toujours compris entre 0 et 1. Il se définit comme le double de l'intersection de deux échantillons divisé par l'union de ces deux lots comme suit :

$$\mathbf{sim}(S_1, S_2) = 2 \frac{|S_1 \cap S_2|}{|S_1| + |S_2|}$$

- **Indice de Jaccard :** ce coefficient est toujours compris entre 0 et 1. Mesure du coefficient de similarité de Jaccard entre deux ensembles de données est le résultat de la division de l'intersection de ces deux ensembles sur l'union de ces derniers selon la formule suivante :

$$sim(S_1, S_2) = \frac{|S_1 \cap S_2|}{|S_1| + |S_2|}$$

- **Indice de Tanimoto :** la métrique de distance Tanimoto mesure la quantité de chevauchement entre les deux ensembles de cardinalité quelconque. ***sim*** prend ses valeurs dans [0,1] avec 0 signifiant qu'il n'y a pas de chevauchement entre les deux ensembles, et 1 que les deux ensembles sont identiques.

$$sim(S_1, S_2) = \frac{|S_1 \cap S_2|}{(2 \times |S_1 + S_2| - |S_1 \cap S_2|)}$$

Nous avons utilisé ces indices car ce type d'indice ne nécessite pas une comparaison entre des ensembles ont la même taille. Parce que dans notre approche la taille de la signature biologique obtenue est variée. Après le calcul de similarité entre les sous-ensembles des gènes sélectionnés, la stabilité peut être calculée selon la formule suivante :

$$stab = \frac{\sum_{i=1}^{l} sim_i(S_1, S_2)}{l} \quad \text{...(VI.13)}$$

## VI.3. Les résultats expérimentaux :

Dans cette section, nous présentons les expérimentations que nous avons réalisées afin de valider notre approche pour la découverte de biomarqueurs. Nous commençons par décrire les bases de données que nous avons utilisées (bases liées à la maladie du cancer). Nous passons après à l'étude des paramètres posés pour présenter finalement les résultats de sélection par notre approche et leur comparaison avec d'autres méthodes proposées dans le même contexte.

### VI.3.1. Les paramètres et les bases de validation :

Pour valider notre approche nous avons utilisé neuf bases de différents types du cancer, tumeurs ou pour des effets thérapeutiques: colon, leucémie avec deux autre bases concernant la même maladie leucémie 1 et 2, sein, DLBCL, la prostate, SRBCT et tumeurs de cerveau.

- *Le cancer du côlon :* la première base que nous avons utilisée pour tester notre approche concernant le cancer du côlon, cette base est constituée de 62 expériences de puces à ADN rassemblés à partir des patients qui ont cette maladie avec 2000 niveaux d'expression de gènes. Parmi ces 62 échantillons, 40 sont des biopsies tumoraux et 22 sont des biopsies normaux à partir des parties saines des côlons des mêmes patients. « *http://datam.i2r.a-star.edu.sg/datasets/krbd/* »
- *La leucémie :* la deuxième base que nous avons utilisée pour tester l'approche c'est la base ALL-AML leucémie, se compose de 72 expériences de puces à ADN rassemblés à partir des patients qui ont la leucémie avec 7129 niveaux d'expression de gènes. Deux classes de distinction : leucémie myéloïde aiguë (en anglais Acute Myeloid Leukemia AML) et la leucémie aiguë lymphoblastique (en anglais Acute Lymphoblastic Leukemia ALL). La base complète constituées de 25 échantillons de AML et 47 d'ALL.

  « *http://datam.i2r.a-star.edu.sg/datasets/krbd/*»

  Pour la même maladie, la leucémie, nous validons deux autres bases, *Leucémie1* constituée de 72 échantillons qui représente trois classes de cette maladie ALL B-cell (38 échantillons), ALL T-cell (9 échantillons ), AML (25 échantillons) avec 5327 niveaux d'expression de gènes et *leucémie2* constituée de 72 échantillons pour trois classes AML (28 échantillons), ALL (24 échantillons), MLL (20 échantillons ) avec 11225 niveaux d'expression de gènes.

  « *http://www.gems-system.org/* »
- *Le DLBCL* : la troisième base concernant Lymphomes diffus à grandes cellules B (en anglais Diffuse large B-cell lymphomas DLBCL), la base est constituée de 77 échantillons avec 5469 niveaux d'expression de gènes pour deux classes, 58 échantillons pour DLBCL et 19 pour lymphome folliculaire (en anglais Follicular Lymphoma FL).

  « *http://www.gems-system.org/* »
- *Le cancer de la prostate :* la quatrième base concernant le cancer de la prostate, elle est constituée de 136 expériences de puces à ADN avec 12600 niveaux d'expression de gènes. Il faut différencier entre deux classes, 77 échantillons avec la tumeur et 59 sans tumeur.

« *http://datam.i2r.a-star.edu.sg/datasets/krbd/* »

- *Le cancer du sein :* cette base définie pour le cancer du sein, c'est une base pour les effets thérapeutiques, elle est constituée de 97 échantillons avec 24481 niveaux d'expression génique. Les patients étudiés montrent deux classes de diagnostic le premier nommé « rechute » pour 46 patients qui ayant développés des métastases à distance dans les 5 ans et le deuxième nommé « non- rechute » pour 51 patients qui sont restés sains de la maladie après le diagnostic initial à l'intervalle d'au moins 5 ans.

  *« http://datam.i2r.a-star.edu.sg/datasets/krbd/ »*
- *SRBCT :* la sixième base étudiée définie pour les tumeurs à petites cellules rondes Bleu (en anglais Small, Round Blue Cell Tumors SRBCT). Ces tumeurs appartiennent à quatre catégories de diagnostic distinctes. Les SRBCT sont des tumeurs malignes qui ont un aspect caractéristique au microscope, ces tumeurs sont plus généralement observées chez les enfants que chez les adultes. Ils représentent généralement des cellules indifférenciées. La base SRBCT que nous avons utilisé est constituée de 83 échantillons avec 2308 niveaux d'expression de gènes représentent quatre types de ces tumeurs EWS (29 échantillons), RMS (25 échantillons), BL (11 échantillons) et NB (18 échantillons).

  *« http://www.gems-system.org/ »*
- *Tumeurs cérébrales:* cette base est constituée de 90 échantillons avec 5920 niveaux d'expression de gènes, les échantillons représentent cinq types de tumeurs embryonnaires du système nerveux central (SNC), Medulloblastoma (60 échantillons), Malignant glioma (10 échantillons), AT/RT (10 échatillons), Normal cerebellum (4 échantillons), PNET (6 échantillons) qui représentent un groupe hétérogène de tumeurs qui ont une apparence morphologique similaire ce qui difficulté le diagnostic .

  *« http://www.gems-system.org/ »*

| La base | Le nombre des échantillons | Le nombre des gènes | Le nombre des classes |
|---|---|---|---|
| **Côlon** | **62** | **2000** | **2** |
| **Leucémie** | **72** | **7129** | **2** |
| **DLBCL** | **77** | **5469** | **2** |
| **Prostate** | **136** | **12600** | **2** |
| **Sein** | **97** | **24481** | **2** |
| **Leucémie 1** | **72** | **5327** | **3** |
| **Leucémie 2** | **72** | **11225** | **3** |

| | | | |
|---|---|---|---|
| **SRBCT** | **83** | **2308** | **4** |
| **Tumeurs cérébrales** | **90** | **5920** | **5** |

**Table VI.1 : La description des bases utilisées pour la validation**

Les paramètres que nous avons utilisé pour exécuter l'approche proposée sont illustrés dans le tableau suivant (**Table VI.2)**, pour l'inertie *a* nous avons l'initialisée à 0.9 et elle est progressivement changée au cours de l'algorithme jusqu'à 0.4 dans les itérations finales comme suit :

***/ i=1…itr avec facteur=0.5/itr;***

L'objectif principal de cette diminution progressive est de favoriser la recherche globale au début de l'algorithme et la recherche locale plus tard pour la bonne exploration de l'espace de recherche.

| **Le paramètre** | **La valeur** |
|---|---|
| **Num_evol** | **25** |
| **itr** | **30** |
| | **k/2** |
| | **0.1** |
| | **0.9** |
| | **round (k*5/100) (5%)** |
| / | **4/-4** |
| **c1=c2** | **2** |
| **a** | **0.9** |
| **s** | **1** |
| **K pour le KNN** | **5** |
| | **0.98** |
| | **0.02** |
| ***l*** | **20** |

**Table VI.2 : Les paramètres de validation**

Le choix de ces paramètres est important, il influe directement sur les résultats obtenus. Pour les probabilités de mutation et de croisement de l'AG que nous avons utilisé sont égal à **0.1** et **0.9** respectivement (un peu grande) afin de maximiser l'exportation de l'espace de recherche (l'espace de solutions). Les valeurs des poids de pondération $\alpha$ et $\beta$ sont proposées avec une grande importance donnée à l'erreur (**0.98**) car nous intéressons d'abord à une signature qui discrimine bien entre les échantillons étudiés avec une erreur de discrimination la plus minimale puis à la taille de cette signature où le poids donné égal à **0.02**. La valeur de ***s*** qui représente les gènes sélectionnés à partir de chaque groupe (top-gènes) est donnée égal à **1** pour éviter la redondance parce que les gènes au sein d'un même groupe ont la même pertinence.

Les valeurs données pour les autres paramètres comme le K de l'algorithme de classification K-NN, le nombre des itérations pour chaque algorithme **(itr)**, le nombre des évolutions **(Num_evol)**, les meilleures solutions à migrer **(k'),** le nombre des itérations pour mesurer la stabilité **(*l*),** le nombre des solutions à remplacer dans CLONALG **($s_r$)** et la vitesse minimale et maximale des particules sont les valeurs qui nous permettons d'obtenir des bons résultats.

### VI.3.2. Les résultats de clustering :

Dans cette section nous représenterons les résultats obtenus dans la première étape du clustering, l'application de l'algorithme proposé pour le Clustering sur les bases de validation fournis les résultats montrés dans le tableau suivant (**Table VI.3)**:

| La base | Le nombre de groupes |
|---|---|
| Côlon | 54 |
| Leucémie | 116 |
| DLBCL | 80 |
| Prostate | 219 |
| Sein | 465 |
| Leucémie1 | 103 |
| Leucémie2 | 94 |
| SRBCT | 288 |
| Tumeurs cérébrales | 73 |

**Table VI.3 : Les résultats de clustering**

Nous pouvons constater que les résultats obtenus et qui sont illustrés dans la table (**Table VI.3 )** montrent l'efficacité de cet algorithme de clustering afin de grouper les gènes qui ont la même pertinence (similaires) dans le même groupe et différencier les gènes qui n'ont pas la même pertinence (dissimilaire) où le taux du groupement atteint jusqu'à **119** gènes par groupe pour la base *leucémie2*, **81** gènes par groupe pour la base des *tumeurs cérébrales*, **68** pour la *DLBCL*, **61** pour la *leucémie*, **57** pour la base de prostate et **52** pour la base du *cancer du sein*,. Tandis que, avec les base *côlon* et *SRBCT* le taux est faible atteint **37** pour la première et très faible pour la deuxième où il est estimé par **8** gènes dans un groupe ce qui signifie qu'il y a une grande différenciation dans l'expression des gènes dans cette base (SRBCT) et ce qui influe ultérieurement sur la taille de la signature biologique sélectionnée.

### VI.3.3. Mesure de performances :

Dans les deux tableaux suivant (**Table VI.4, Table VI.5)** nous illustrons les résultats obtenus (l'erreur, le nombre de gènes et la sensibilité) sur les neuf bases avec les trois algorithmes de classification SVM, KNN et Bayes et les deux méthodes dans le système de filtrage mRMR et SVM-RFE.

| La base | SVM | | KNN | | Bayes | |
|---|---|---|---|---|---|---|
| | ***Erreur / Nb gènes*** | ***sensibilité*** | ***Erreur / Nb gènes*** | ***sensibilité*** | ***Erreur / Nb gènes*** | ***sensibilité*** |
| Côlon | **0.00/01** | **0.92** | **0.00/02** | **0.92** | **0.01/03** | **0.97** |
| Leucémie | **0.00/02** | **1.00** | **0.00/03** | **1.00** | **0.00/02** | **0.97** |
| DLBCL | **0.00/02** | **1.00** | **0.00/03** | **0.96** | **0.00/04** | **1.00** |
| Prostate | **0.00/02** | **0.97** | **0.00/03** | **0.92** | **0.01/03** | **0.96** |
| Sein | **0.00/02** | **0.96** | **0.01/03** | **0.98** | **0.02/03** | **0.95** |
| Leukémia1 | **0.00/03** | **1.00** | **0.00/04** | **0.97** | **0.01/03** | **0.98** |
| Leukémia2 | **0.00/02** | **1.00** | **0.00/04** | **0.96** | **0.00/05** | **0.96** |
| SRBCT | **0.00/04** | **0.98** | **0.00/05** | **0.98** | **0.00/07** | **1.00** |
| Tumeurs cérébrales | **0.04/04** | **1.00** | **0.06/02** | **0.93** | **0.04/04** | **1.00** |

**Table VI.4 : Mesure de performance avec le SVM-RFE**

| La base | SVM | | KNN | | Bayes | |
|---|---|---|---|---|---|---|
| | *Erreur / Nb gènes* | *sensibilité* | *Erreur / Nb gènes* | *sensibilité* | *Erreur / Nb gènes* | *sensibilité* |
| Côlon | **0.00/02** | **0.97** | **0.00/02** | **0.92** | **0.01/05** | **0.95** |
| Leucémie | **0.00/03** | **0.97** | **0.00/03** | **1.00** | **0.00/03** | **1.00** |
| DLBCL | **0.00/03** | **0.98** | **0.00/03** | **0.96** | **0.00/04** | **1.00** |
| Prostate | **0.00/03** | **0.96** | **0.00/02** | **0.95** | **0.01/03** | **0.96** |
| Sein | **0.01/03** | **0.98** | **0.00/04** | **0.92** | **0.02/04** | **0.92** |
| Leucémie1 | **0.00/04** | **0.99** | **0.02/04** | **0.97** | **0.04/03** | **0.92** |
| Leucémie2 | **0.00/02** | **0.96** | **0.00/05** | **1.00** | **0.02/03** | **1.00** |
| SRBCT | **0.00/03** | **0.98** | **0.00/05** | **0.98** | **0.00/08** | **0.96** |
| Tumeurs cérébrales | **0.02/05** | **1.00** | **0.02/03** | **0.98** | **0.03/06** | **1.00** |

**Table VI.5: Mesure de performance avec mRMR**

Nous pouvons constater que les résultats obtenus pour les neuf bases en utilisant la méthode SVM-RFE sont meilleurs que ceux obtenus en utilisant la méthode mRMR à cause de leur liaison avec le système de classification dans le filtrage, mais nous pouvons dire aussi que dans les deux cas et avec les deux algorithmes de filtrage les résultats obtenus présentent un taux d'erreur de classification faible entre 0 et 0.06 et des sous-ensembles avec 8 et moins de 8 gènes sélectionnés, nous pensons que l'initialisation des paramètres utilisée dans notre travail contribue à la performance de l'approche d'une manière significative pour trouver des petites signatures biologiques avec une grande précision de la classification (faible erreur). Nous remarquons que les résultats pour les bases qui ont deux classes des échantillons sont mieux ceux obtenus pour les bases où le nombre de classes supérieur à deux. Aussi, les résultats obtenus avec les classificateurs SVM et KNN sont mieux qu'avec le classificateur Bayes où le taux d'erreur atteint jusqu'à 0% avec 2 à 3 gènes sélectionnés (Prostate, Côlon) pour les algorithmes KNN et SVM où l'algorithme de filtrage est SVM-RFE tandis que, pour l'algorithme de Bayes et avec les même bases le taux d'erreur estimé par 1% avec 3 à 5 gènes sélectionnés. Finalement, nous pouvons dire que nous avons réussi avec notre approche à améliorer les taux de classification tout en diminuant le nombre de gènes sélectionnés. La

figure suivante illustre un exemple des changements de l'erreur avec le nombre de gènes pendant les vingt-cinq migrations.

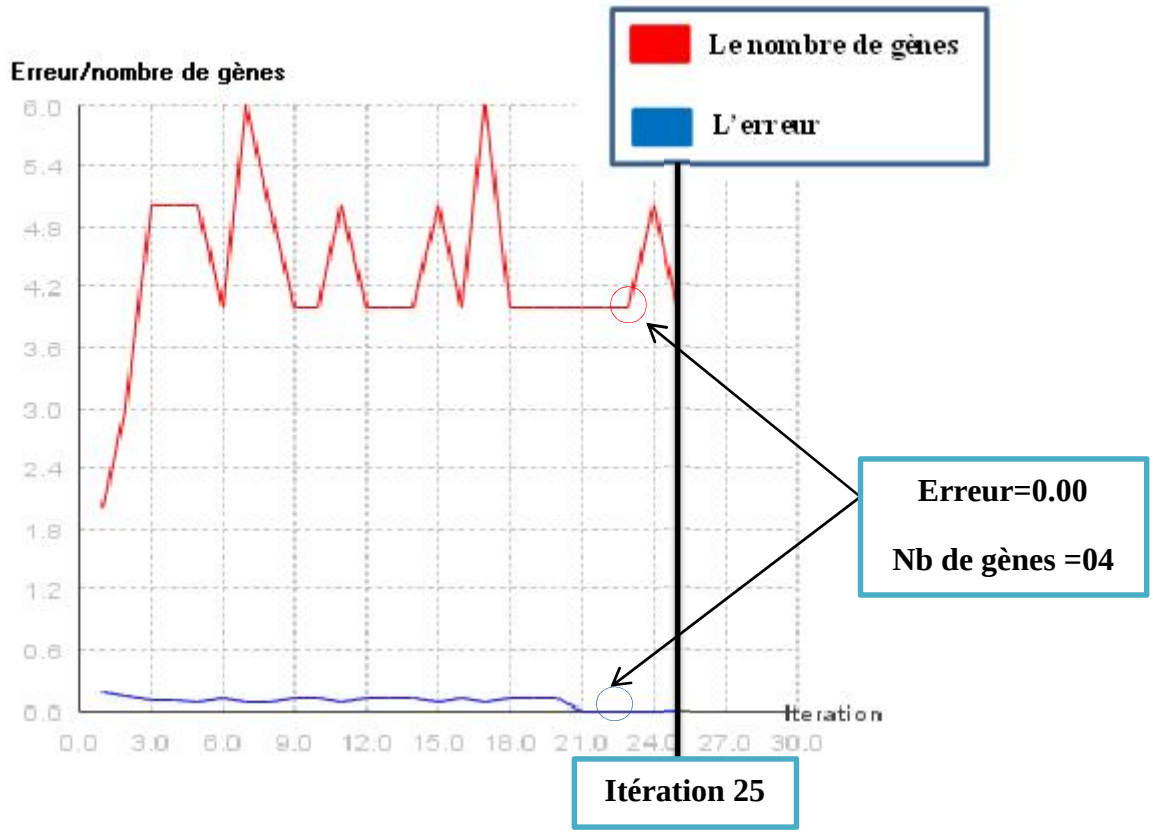



**Figure VI.12: Les valeurs de l'erreur et nombre de gènes, la base de la leucémie2, le filtrage : SVM-RFE, le classificateur :KNN**

### VI.3.4. Mesure de stabilité :

Dans cette section nous présenterons les résultats obtenus lorsque nous mesurons la stabilité en utilisant les indices de similarité de Dice, Jaccard et Tanimoto sur les neuf bases après les perturbations de la base initiale, les résultats sont présentés dans le tableau suivant (**Table VI.6)**. Dans ce tableau nous avons montré les moyennes des similarités calculées pour chaque base et avec chaque indice de similarité, la dernière colonne du tableau présente la moyenne des trois mesures pour chaque base.

| Les bases | La stabilité calculée | | | Moyenne |
|---|---|---|---|---|
| | Dice | Jaccard | Tanimoto | |
| Colon | 0.92 | 0.91 | 0.91 | 0.91 |
| Leucémie | 0.94 | 0.92 | 0.91 | 0.92 |
| DLBCL | 0.94 | 0.93 | 0.92 | 0.93 |
| Prostate | 0.94 | 0.91 | 0.89 | 0.91 |
| Sein | 0.85 | 0.79 | 0.75 | 0.78 |
| Leucémie1 | 0.92 | 0.89 | 0.87 | 0.89 |
| Leucémie2 | 0.87 | 0.83 | 0.80 | 0.83 |
| SRBCT | 0.95 | 0.92 | 0.88 | 0.91 |
| Tumeurs cérébrales | 0.90 | 0.86 | 0.83 | 0.86 |

**Table VI.6 : Mesure de stabilité avec les trois indices : Dice, Jaccard et Tanimoto**

Les résultats obtenus et illustrés dans le tableau au-dessus démontrent que l'approche proposée permet d'obtenir une signature biologique avec une stabilité remarquable (signature robuste), nous pouvons observer que toutes les mesures fournies par nos algorithmes présentent un taux de similarité élevé plus que **75%**. Aussi, les mesures obtenus avec l'indice de Dice sont mieux que les deux autres indices de Jaccard et Tanimoto où le taux de similarité atteint jusqu'à **95%**. A partir de ces résultats nous pouvons classer notre approche comme une approche robuste. La figure suivante montre un exemple sur l'évolution des valeurs des similarités pendant les vingt itérations pour la base *Leucémie*.

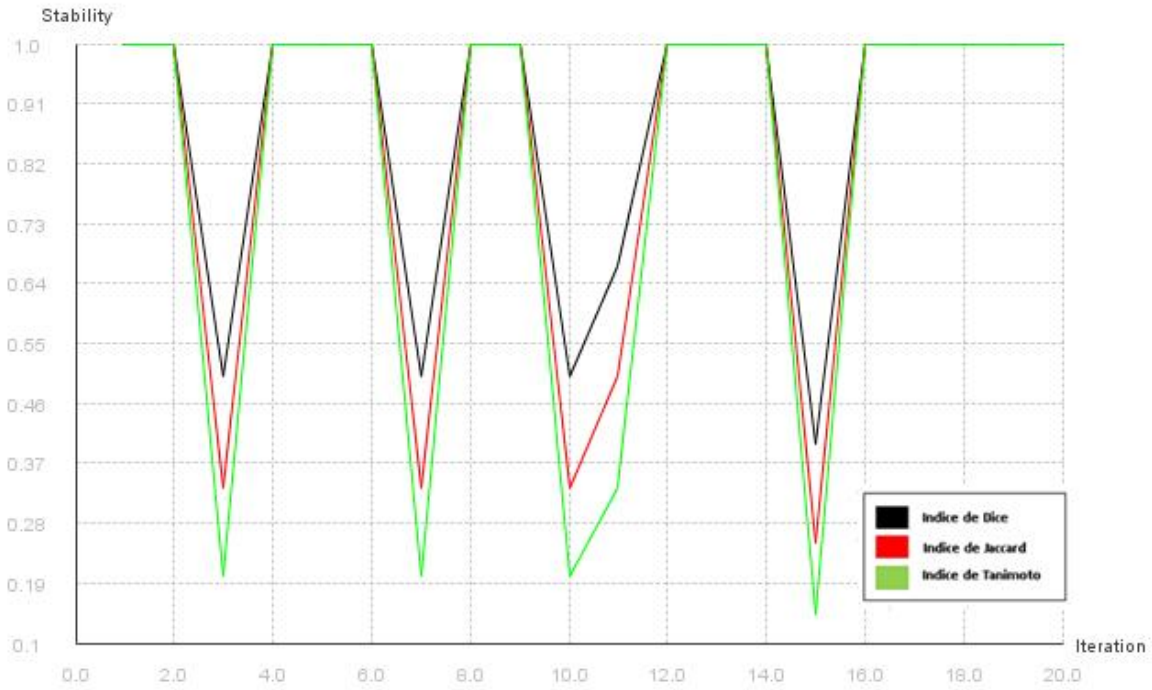



**Figure VI.13: Les valeurs des similarités pendant les vingt itérations, la base de la leucémie2**

## VI.4. Une étude comparative :

Dans cette section nous avons comparé les résultats de notre travail aux résultats des travaux déjà réalisés dans le même contexte et avec les mêmes bases. La comparaison se fait au trois niveaux, le taux d'erreur, le nombre de gènes sélectionné et la stabilité. Les travaux que nous avons sélectionnés pour cette comparaison sont des travaux qui ont utilisés les algorithmes d'optimisation pour la sélection et des autres qui ont visé à optimiser la stabilité.

| La base | Le travail | L'erreur | Le nombre de gène | La stabilité |
|---|---|---|---|---|
| Colon | [9] | 0.025 | 5 | / |
| | [34] | 0.02 | 100 | 0.89 |
| | [90] | 0.21 | / | [0.45 -0.99] |
| | [31] | 0.15 | 10 | / |
| | [88] | 0.21 | 50 | 0.47 |
| | [42] | 0.21 | 10 | 0.89 |
| Leucémie | [9] | 0.009 | 10 | / |
| | [34] | 0.05 | 20 | 0.82 |

| | | | | |
|---|---|---|---|---|
| | **[90]** | **0.14** | **/** | **[0.54 -0.99]** |
| | **[31]** | **0.02** | **10** | **/** |
| | **[88]** | **0.1** | **50** | **0.68** |
| | **[42]** | **0.06** | **10** | **0.89** |
| **DLBCL** | **[9]** | **0.009** | **13** | |
| | **[32]** | **0.00** | **3** | **/** |
| | **[34]** | **2.60** | **120** | **0.84** |
| | **[88]** | **0.14** | **50** | **0.68** |
| | **[92]** | **0.00** | **03** | **/** |
| **Sein** | **[9]** | **0.00** | **20** | |
| | **[88]** | **0.40** | **50** | **0.33** |
| **Prostate** | **[34]** | **0.14** | **10** | **0.84** |
| | **[88]** | **0.22** | **50** | **0.47** |
| | **[92]** | **0.00** | **02** | **/** |
| | **[42]** | **0.06** | **10** | **0.71** |
| **SRBCT** | **[32]** | **0.00** | **5** | **/** |
| **Tumeurs cérébrales** | **[32]** | **0.02** | **12** | **/** |
| | **[88]** | **0.28** | **50** | **0.47** |

**Table VI.7 : Les résultats de quelques travaux dans le contexte**

A partir le tableau au-dessus nous pouvons déduire que les approches qui prennent en considération le critère de stabilité comme les travaux **[34] [88] [90]** sont moins performants (un taux d'erreur élevé) et la taille de la signature est très grande, si nous comparons avec notre approche, ces travaux ont utilisés en général l'approche basée ensemble (sous échantillonnage aléatoire) pour augmenter la stabilité, alors nous pouvons constater que, avec cette approche d'amélioration de stabilité les performances sont diminuées.

Tandis que les approches qui ont utilisés les algorithmes d'optimisation pour la sélection soit sur la base entière, où après une étape de filtrage comme les travaux **[9] [31] [32] [92]** les résultats obtenus montrent un taux d'erreur faible entre 0 et 0.15, avec un nombre de gènes sélectionnés varié entre 2 et 50, mais la stabilité est toujours ignorée dans ces méthodes. A partir de cette comparaison, nous pouvons dire que notre méthode de découverte a aidée à trouver une signature biologique minimale qui représentent le bon biomarqueur avec un taux

d'erreur faible atteint jusqu'à 0 % et une stabilité élevée atteint jusqu'à 95%, c'est dire nous avons réussi à rassembler dans ce travail entre *la performance* et *la stabilité.*

## VI.5. Conclusion

Dans ce chapitre nous avons détaillé notre contribution qui est proposée dans le cadre de ce travail qui concerne la découverte de biomarqueurs pour le diagnostic du cancer. Notre approche est composée de trois étapes successives, la première nommée l'étape de Clustering qui vise à éliminer les redondances dans les étapes suivantes avec une maximisation de la stabilité, la deuxième c'est le système de filtrage qui vise à diminuer l'espace de recherche et sélectionner les gènes les plus pertinents au problème à résoudre (les classes des échantillons), la dernière concernant l'optimisation de la sélection qui est composé lui-même de trois algorithmes d'optimisation **GA**, **CLONALG** et **PSO**. Les algorithmes sont en coopération la nature de cette coopération c'est une migration des meilleures solutions pendant certain nombre des évolutions déterminé par l'utilisateur. Notre approche est testée sur neuf bases représentent des différents types du cancer pour deux mesures, l'erreur / le nombre de gènes sélectionnés et la stabilité. Les résultats montrent l'efficacité de notre méthode de découverte et pour justifier bien cette efficacité nous avons réalisé une étude comparative avec des travaux déjà réalisés dans le même contexte. Enfin, nous pouvons dire qu'avec ce travail nous avons réussis à hybrider entre la découverte de bon biomarqueur qui discrimine bien les échantillons étudiés et la robustesse de ce biomarqueurs ou bien cette signature biologique.

# *Conclusion générale et perspectives*

Actuellement, la bioinformatique comme un domaine multidisciplinaire qui fait appel à plusieurs disciplines telles que la biologie moléculaire, l'informatique, les mathématiques et les statistiques possède plusieurs axes de recherche. Parmi ces axes nous nous intéressons dans notre travail à la découverte de biomarqueurs transcriptomiques pour le diagnostic du cancer. Une des technologies à haut débit de la biologie moléculaire la plus utilisée pour la découverte de ce type de biomarqueurs c'est la technique des puces à ADN. C'est une technique qui permet d'obtenir un ensemble de données de haute dimension, des milliers de gènes exprimés simultanément, ces expressions sont mesurées afin de déterminer le niveau d'expression de chaque gène. Ces données peuvent contenir des gènes bruyants, redondants et non pertinents, l'objectif d'un processus de découverte de biomarqueurs est de sélectionner un ensemble minimal de gènes les plus informatifs et qui sont exprimés différentiellement. Ces biomarqueurs peuvent être utilisés à la suite pour le diagnostic, le pronostic, la prédiction, le criblage, les essais cliniques…etc. Dans ce contexte, nous avons utilisé les techniques de l'intelligence artificielle, notamment les techniques d'apprentissage automatique et nous avons défini le problème de découverte de biomarqueurs comme un problème de sélection de caractéristiques avant une étape d'apprentissage supervisé (la classification). La sélection de caractéristiques peut être définie comme un processus d'élimination des caractéristiques/gènes bruitées, redondantes et non pertinentes et la classification pour désigner si les gènes sélectionnés discriminent bien entre les échantillons à étudier (gènes exprimés différentiellement) ou non.

La synthèse des techniques de sélection existantes nous a permis de déduire un certain nombre de faiblesses parmi lesquelles la complexité très élevée pour les approches wrapper et l'indépendance au classificateur pour les approches Filter ce qui conduit à des résultats moins performants. Une autre limitation connue pour les méthodes de sélection en générale c'est l'instabilité, la plupart des méthodes existantes ne prennent pas en considération l'aspect de la stabilité malgré le fait qu'il est très important dans le domaine biomédical.

Dans le but de limiter ces inconvénients notre contribution a été développée. Nous avons proposé une nouvelle approche performante et robuste pour la découverte de biomarqueurs.

Cette approche est composée de trois étapes. La première nommée *l'étape de clustering*, les deux objectifs principaux de cette étape sont : éliminer la redondance et garantir la stabilité. Dans ce contexte, nous avons proposé un nouvel algorithme de clustering basé sur l'approximation d'une couverture de Markov. La deuxième étape nommée *l'étape de filtrage*, dans cette étape nous avons utilisé deux types de méthodes filtre et embedded (mRMR et SVM-RFE) pour filter les gènes non pertinents ce qui conduit immédiatement à diminuer la dimensionnalité de l'espace de données. La dernière étape c'est *l'étape d'optimisation* dans laquelle nous avons exploité trois algorithmes d'optimisation : les algorithmes génétiques, l'optimisation par essaim particulaire et l'algorithme de la sélection clonale avec une fonction multiobjectif constituée de deux objectifs de minimisation, le taux d'erreur et la taille de l'ensemble sélectionné. Nous avons utilisé les méthodes agrégées (la somme pondérée) pour résoudre cette fonction. Les trois algorithmes sont en coopération via *une migration des meilleures solutions* afin d'exploiter les avantages de chacun pour trouver à la fin des évolutions la bonne solution c'est-à-dire une signature robuste de taille minimale et performante.

Nous avons validé notre approche sur des bases pour des différents types du cancer. La validation se fait par deux critères, la performance et la stabilité. Pour évaluer les performances, nous avons utilisé la technique de validation croisée *k folds* avec *k=10* grâce à leur efficacité quand le nombre des observations est petit, pour la stabilité nous avons utilisé une perturbation avec des chevauchements fixes de 80% échantillons et les indices de Dice, Jaccard et Tanimoto pour mesurer la similarité entre les sous-ensembles de gènes sélectionnés à partir des sous bases perturbées, cette opérations est réalisée pendant vingt itérations et la moyenne des similarités est calculée, cette moyenne désigne le taux de la stabilité . Les résultats obtenus ont montré que nous avons réussi à réduire le nombre de gènes sélectionnés avec un taux d'erreur très faible et une stabilité très satisfaisante.

Nous terminons ce manuscrit par donner quelques perspectives et prolongements naturels de ce travail:

*A court terme :*

- Tester notre approche sur autres bases avec un nombre de classes supérieur à cinq.
- Utiliser des autres méthodes de type filtre comme reliefF,ttest,..etc., dans la deuxième étape pour étendre l'approche.

*A long terme :*

- Proposer d'autres techniques d'optimisation dans la troisième phase afin d'améliorer les résultats.
- Proposer un nouveau protocole de coopération plus compliqué entre les algorithmes d'optimisation afin d'exploiter mieux les performances de chacun.
- Utiliser l'ensemble de classificateurs ou méta-ensemble au lieu d'un seul classificateur afin d'obtenir des résultats de la classification plus stable.
- Utiliser d'autres algorithmes de clustering supervisé dans la première étape afin de comparer les résultats avec celui obtenus par notre approche.
- Remplacer la première étape de clustering par un sous échantillonnage aléatoire ou par des techniques d'échantillonnage comme le Bagging, Bootstrapping pour comparer aussi les résultats de la stabilité avec celui obtenus.

A la fin de ce travail, le domaine de la découverte de biomarqueurs en utilisant les techniques de la sélection de caractéristiques et l'apprentissage supervisé (la classification) reste toujours ouvert et connait des développements très importants dont le but principal des différentes recherches est identifié un sous ensemble minimal de gènes qui distinguent bien entre les échantillons ( le bon biomarqueur) et qui est utilisé à la suite dans des différentes applications comme le développement des médicaments, la médecine clinique…etc.

# *Bibliographie:*

**[1]** Alain, B. (2008). Algorithme évolutionnaire pour l'optimisation multiobjectif. Séminaire – LAAS.

**[2]** Anant, N.- B., Rohit, M., Abdullah, F., Amit, V. et Dwarakanath, B.-S.(2010). Cancer biomarkers - Current perspectives. Indian J Med Res 132, pages: 129-149

**[3]** Andrade et Sander. (1997). From genome data to biological knowledge, Current Opinion in Biotechnology. Journal of Bioinformatics.

**[4]** Anil, K.- M., Pranita,J. et Shailendra, K.- S. (2012). Protein structure prediction using support vector machine. International Journal on Soft Computing ( IJSC ) Vol.3, No.1.

**[5]** Antoine, C., Laurent, M. et Yves, K. (2003). Apprentissage Artificiel- Concepts et algorithms. Ouvrage, http://www.e-booksland.com/Algorithmique-et-programmation/Apprentissage-artificiel-Concepts-et-algorithmes.html.

**[6]** Arnaud,F.(2009). Classification d'ARN codants et d'ARN non-codants. Thèse de doctorat, Université des Sciences et Technologies de Lille.

**[7]** Arpad, K., Ajith, A. et Yuehui, C. (2008). Computational Intelligence in Bioinformatics. Studies in Computational Intelligence, Volume 94, Springer

**[8]** Baker, M. (2005). In biomarkers we trust. Nature Biotechnology 23(3): 297‑304.

**[9]** Barnali, S., Debahuti, M. (2012). A Novel Feature Selection Algorithm using Particle Swarm Optimization for Cancer Microarray Data. International Conference on Modeling Optimization and Computing (ICMOC-2012).

**[10]** Berna, H.-U. et Sadan, K.-K. (2011). A review of clonal selection algorithm and its applications. Artif Intell Rev, Springer.

**[11]** Bing X., Mengjie Z., Will N.- B. (2012). Multi-Objective Particle Swarm Optimisation (PSO) for Feature Selection. ACM 978-1-4503-1177-9/12/07.

**[12]** Biomarkers Definitions Working Group.(2001). Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. PubMed, Clin Pharmacol Ther. 2001 Mar;69(3), pages: 89-95.

**[13]** Boyun Z. (2009) .Retracted: Using Bayesian Network and AIS to Perform Feature Subset Selection. 5th International Conference on Intelligent Computing, ICIC 2009 Ulsan, South Korea.

**[14]** Breiman, L. (1996). Bagging predictors. Machine Learning, 24, pages:123-140.

**[15]** Castro, L.N.d., et Timmis, J.I. (2003).Artificial Immune Systems as a Novel Soft Computing Paradigm, Soft Computing - A Fusion of Foundations, Methodologies and Applications, 7, (8), pages : 526-544.

**[16]** Céline, B.-A. (2012). Introduction à la bioinformatique. Université Claude Bernard, Lyon 1, Laboratoire de Biométrie et Biologie Evolutive (UMR 5558)

**[17]** Cheng-San, Y. Li-Yeh, C. J.L.et Cheng,H. (2008).Information Gain with Chaotic Genetic Algorithm for Gene Selection and Classification Problem. IEEE, Man and Cybernetic, pages : 1128-1133

**[18]** Cho, H. S. (2003) . DNA microarray data based classification of cancers using neural networks and genetic algorithms. Nanotech, 1, 28–31.

**[19]** Chuang, L.-Y., Chang, H.-W., Tu, C.-J., Yang, C.-H.( 2008). Improved binary pso for feature selection using gene expression data. Computational Biology and Chemistry 32 (1),pages : 29–38.

**[20]** Claverie, J.-M., Audic, S., et Abergel, C. (1999). La Bioinformatique: une discipline stratégique pour l'analyse et la valorisation des génomes. *Conference Proceeding:Rencontres de Luminy.*

**[21]** Cosmin, L., Jonatan, T., Stijn, M., David, S., Alain, C., Colin, M., Virginie, S., Robin, D., Hugues, B. et Ann, N. (2012).A Survey on Filter Techniques for Feature Selection in Gene Expression Microarray Analysis. IEEE/ACM transactions on computational biology and bioinformatics, VOL. 9, NO. 4.

**[22]** Darius, M.-D. (2010). Data mining for genomics and proteomics : Analysis of Gene and Protein Expression Data. Wiley.

**[23]** David, D., Blaise, H.et Jean-Daniel, Z. (2012). Évolution de la stabilité de la sélection de variables en fonction de la taille d'échantillon et de la dimension. CAp 2012

**[24]** Davis, L. (1991). The genetic Algorithm Handbook. Ed.New-York : Van Nostrand Reinhold, ch.17.

**[25]** David, W.-M.(2004). Bioinformatics: Sequence and genome analysis. Cold Spring Harbor Laboratory Press.

**[26]** De Castro.L.N et Von Zuben.F.J. (2000). Clonal selection algorithm with engineering Application. Proc GECCO's, Las Vegas, NV, pp. 36–37.

**[27]** Dettling M. (2004). BagBoosting for tumor classification with gene expression data. Bioinformatics; 20(18):3583–3593.

**[28]** Dhananjaya, P.-S., Ratna, P., Anil, R. et Dilip, K.-A. (2012). Bioinformatics-Assisted Microbiological Research: Tasks, Developments and Upcoming Challenges. American Journal of Bioinformatics 1 (1): 10-19.

**[29]** Eberhart, R.C., Shi, Y., Kennedy, J.( 2001). Swarm Intelligence. Morgan Kaufmann, San Francisco, CA, USA.

**[30]** Ein-Dor, L., Zuk, O. et Domany, E. (2006). Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. Proceedings of the National Academy of Sciences, 103(15), 5923–5928.

**[31]** EL AKADI, A. , AMINE , A., EL OUARDIGHI , A. et ABOUTAJDINE, D. (2009) .A New Gene Selection Approach Based on Minimum Redundancy-Maximum Relevance (MRMR) and Genetic Algorithm (GA). Computer Systems and Applications, 2009. AICCSA 2009. IEEE/ACS International Conference, pages : 69-75.

**[32]** Emmanuel, M. Mario, M.-A. et Victor, T.(2010). Compact cancer biomarkers discovery using a swarm intelligence feature selection algorithm. Computational Biology and Chemistry 34, Elsevier, pages : 244–250

**[33]** Enrique, A., José, G.-N., Laetitia, J. et El-Ghazali, T. (2007). Gene Selection in Cancer Classification using PSO/SVM and GA/SVM Hybrid Algorithms. Congress on Evolutionary Computation, Singapor

**[34]** Feng, Y. et Mao, K.- Z. (2011). Robust Feature Selection for Microarray Data Based on Multicriterion Fusion. IEEE/ACM transactions on computational biology and bioinformatics, vol. 8, no.4.

**[35]** Francisco, A. (2010). Bioinformatics and Biomarker Discovery. Wiley

**[36]** Freund, Y. et Schapire, R. (1996). Experiments with a new boosting algorithm, Machine Learning: Proceedings of the Thirteenth International Conference, pp. 148-156.

**[37]** Gang, W., Frederick, H.- L. et Qiang, Y. (2004). Feature Selection with Conditional Mutual Information MaxiMin in Text Categorization. ACM

**[38]** Giorgio, V. (2011). Machine learning methods for gene/protein function prediction. University de gliStudi di Milano.

**[39]** Gokhan, G., Zehra, C. et Lei, Y. (2009). Stable and Accurate Feature Selection. ECML PKDD '09 Proceedings of the European Conference on Machine Learning, and Knowledge Discovery in Databases: Part I, pages 455 – 468.

**[40]** Goldberg et David, E. (1989). Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1st edition.

**[41]** Guillaume, C. (2009). Optimisation par essaim particulaire. Ecole d'ingénieurs en informatique, France.

**[42]** Guoyin, W., Juan, G. et Feng, H. (2013). A stable gene selection method based on sample weighting. 26th IEEE Canadian Conference Of Electrical And Computer Engineering (CCECE).

**[43]** Guyon, B., Weston, S., Barnhill, V., et Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. Machine Learning, 46(1–3), 389–422.

**[44]** Guyon, I. et Elisseeff A. ( 2003). An introduction to feature and variable selection. Journal of Machine Learning Research, vol. 3, p. 1157-1182.

**[45]** Hans-Joachim, B. (2007). Algorithms Aspects of bioinformatics. Ouvrage, Natural Computing Series, Springer.

**[46]** Hassan, C. (2011). Sélection de caractéristiques: méthodes et applications. Thèse de doctorat, Université Paris Descartes.

**[47]** Hassan, C., Florence, C. et Nicole, V. (2014). Combination of Single Feature Classifiers for Fast Feature Selection. Springer International Publishing Switzerland

**[48]** Haury, A.-C., Gestraud, P. et Vert J.-P. (2011). The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. PLoS ONE, 6(12), e28210.

**[49]** Helman, P. (2004). A bayesian network classification methodology for gene expression data. Journal of Computational Biology, 11(4), 581–615.

**[50]** Hervé, F.-B. (2012) . Machines à Vecteurs Support. Ecole supérieure d'électricité.

**[51]** Housset,C. et Raisonnier, A.(2009). Biologie Moléculaire. Université Pierre et Marie Curie.

**[53]** Huawen, L., Lei, L. et Huijie, Z. (2010). Ensemble gene selection by grouping for microarray data classification. Journal of Biomedical Informatics 43 (2010) 81–87

**[54]** Indyk, P. et Motwani, R. (1998). Approximate nearest neighbors: Towards removing the curse of dimensionality. Pages: 604-613.

**[55]** Jérôme, A.(2012). Prédiction d'Interactions et Amarrage Protéine-Protéine par combinaison de classifieurs. UNIVERSITÉ PARIS-SUD.

**[56]** Jialei, W., Peilin, Z., Steven, C.-H., Hoi, Member, IEEE, et Rong, J. (2014). Online Feature Selection and Its Applications. IEEE transactions on knowledge and data engineering, VOL. 26, NO. 3, 698-710.

**[57]** Jmal, Y., Talbi, E-G. et Mellouli,K.(2010). Artificial Immune System for Feature Selection

**[58]** John, G.-H., Kohavi, R. et Pfleger, K. (1994). Irrelevant features and the subset selection problem. In machine learning : proceedings of the eleventh international, pages: 121-129. Morgan Kaufmann.

**[59]** John, G.- H. (1997). Enhancements to the data mining process. Thèse de doctorat, Stanford, CA, USA.

**[60]** Julie, P. (2005) .Analyse statistique des données issues des biopuces à ADN. Thèse de doctorat, Université JOSEPH FOURIER.

**[61]** Kamal, T. (2008). Développement de nouvelles phases stationnaire monolothiques pour la nano-chromatographie et l'analyse protéomique. Thèse de doctorat, université des sciences et technologies de LILLE.

**[62]** Kardinal, C., Yarbro, J.(2012). A conceptual history of cancer. *Semin Oncol*. 1979;6:396–408

**[63]** Kennedy,J. et Eberhart, R.(1995). Particle Swarm Optimization. IEEE International Conference on Neural Networks, vol. 4, pages : 1942–1948.

**[64]** Kevin, R., Coombes, J. - W. et Keith, A. - B. (2004). The tail-rank statistic for finding biomarkers from microarray data, with application to prostate cancer. Anderson Cancer Center.

**[65]** Khalid, R. (2010). Application of Data Mining In Bioinformatics. Indian Journal of Computer Science and Engineering .Vol 1 No 2, 114-118

**[66]** Khamseh,A., Alinejad-Beromi, Y. (2011). Hybrid CLONAL Selection Algorithm with SA for Solving Economic Load Dispatch with Valve-Point Effect. Canadian Journal on Electrical and Electronics Engineering Vol. 2, No. 10, pages: 463:467

**[67]** Kohavi, R. et John, G.- H. (1997). Wrappers for feature subset selection. Artif. Intell., 97:273-324.

**[68]** Ladha, L. et Deepa, T. (2011). Feature selection methods and algorithms. International Journal on Computer Science and Engineering (IJCSE), Vol. 3 No. 5, pages: 1787: 1797.

**[69]** Laurent, N. (2012). Bioinformatique et données biologiques. www.lifl.fr/~noe/enseignement/m1-genpro/.../bioinfo_bio1-2x3.pdf.

**[70]** Leandro, N,-C. et Fernando, J.-V.-Z. (1999). Artificial Immune Systems: Part I – basic theory and applications. Technical Report.

**[71]** Lemeur, N. (2005). De l'Acquisition des Données de Puces à ADN vers leur Interprétation : Importance du Traitement des Données Primaires. Thèse de doctorat, Ecole Doctorale CHIMIE BIOLOGIE

**[72]** Liangpei, Z., Yanfei, Z., Jianya, G. et Pingxiang, L.(2007). Dimensionality Reduction Based on Clonal Selection for Hyperspectral Imagery. IEEE, TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, VOL. 45, NO. 12, pages : 4172 :4186

**[73]** Lior, R. (2009). Ensemble-based classifiers. Springer Science+Business Media B.V.

**[74]** Luiz O.- V.-B.-O., Rodrigo L.- M.- M. et Dante A.- C.-B. (2012).Clonal Selection Classifier with Data Reduction: Classification as an Optimization Task WCCI 2012 IEEE World Congress on Computational Intelligence,Brisbane, Australia .

**[75]** Maurice, C. et Kennedy,J. (2002). The particle swarm - explosion, stability, and convergence in multidimensional complex space. IEEE , Evolutionary Computation, 6(1):58-73.

**[76]** Miguel, R., Rui, M., Paulo, M., Daniel, G.-P. et Florentino, F.-R. (2007). A Platform for the Selection of Genes in DNA Microarray Data using Evolutionary Algorithms. ACM

**[77]** Miller, B. L. et Goldberg, D. E. (1995). Genetic algorithms, tournament selection, and the effects of noise. Complex Systems, 9:193-212.

**[78]** Nancie, R. (2004). Bioinformatique des puces à ADN et application à l'analyse du transcriptome de Buchnera aphidicola. Thèse de doctorat, L'institut national des sciences appliquées de Lyon

**[79]** Nawin, M. (2010). Introduction to proteomics: Principles and Applications. John Wiley & Sons, Inc.

**[80]** Ouiza, Z. (2013). L'optimisation non linéaire Multiobjectif. Thèse de doctorat, Université de TIZI-OUAZOU.

**[81]** Ouzounis, C. et Valencia, A. (2003). Early bioinformatics: the birth of a discipline-a personal view. Bioinformatics, vol. 19, n°17, pp. 2176-2190.

**[82]** Paul, G.-H. et Teresa, K.-A. (2005). Bioinformatics and molecular evolution. Blackwell Science Ltd, Blackwell Publishing company

**[83]** Pavel, A.- P.(2006). Bio-informatique moléculaire (Une approche algorithmique). Ouvrage, Université de California, San Diego, Springer-Verlag France, Paris.

**[84]** Pedro, D. (2012). A Few Useful Things to Know about Machine Learning. Communications of the ACM, Volume 55 Issue 10, pages : 78-87.

**[85]** Peng, H., Long, F. et Ding, C. (2005). Feature selection based on mutual information : criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27:1226-1238.

**[86]** Piyushkumar, A.-M. et Jagath C.-R. (2010). SVM-RFE With MRMR Filter for Gene Selection. IEEE TRANSACTIONS ON NANOBIOSCIENCE, VOL. 9, NO. 1, pages : 31-37.

**[87]** Qinghai, B. (2010). Analysis of Particle Swarm Optimization Algorithm. Computer and Information Science, Vol. 3, No. 1, pages: 180-184.

**[88]** Randall, W., Taghi, K. et David, D. (2012). A New Fixed-Overlap Partitioning Algorithm for Determining Stability of Bioinformatics Gene Rankers. 11th International Conference on Machine Learning and Applications.

**[89]** Saeys, Y., Inza, I. et Larran˜aga, P. (2007). A Review of Feature Selection Techniques in Bioinformatics. Bioinformatics, vol. 23, no. 19, pages: 2507-2517.

**[90]** Saeys, Y., Thomas, A. et Van de Peer, Y. (2008). Robust Feature Selection Using Ensemble Feature Selection Techniques. Machine Learning and Knowledge Discovery in Databases, Lecture Notes in Computer Science Volume 5212, 2008, pp 313-325, Springer

**[91]** Sajid, N. et Dhruba Kr. B. (2013). Classification of microarray cancer data using ensemble approach. Network Modeling Analysis in Health Informatics and Bioinformatics 2(3). DOI:10.1007/s13721-013-0034-x, Springer.

**[92]** Salam, S.-S., Salwani, A., Mohd, Z.- A.- N., Malek, A. (2013). Hybridizing Relief, mRMR filters and GA wrapper approaches for gene selection. Journal of Theoretical and Applied Information Technology, Vol. 47 No.3.

**[93]** Sanghamitra, B., Saurav, M. et Anirban, M. (2013). A Survey and Comparative Study of Statistical Tests for Identifying Differential Expression from Microarray Data. IEEE transactions on computational biology and bioinformatics.

**[94]** Sean,D.-M. , Jessica, D.-T. et Russ,B.- A.(2014). Bioinformatics. E.H. Shortliffe, J.J. Cimino (eds.), Biomedical Informatics, 695 DOI 10.1007/978-1-4471-4474-8_24, Springer-Verlag London.

**[95]** Soha, A., Mengjie, Z. et Lifeng, P. (2013). Enhanced Feature Selection for Biomarker Discovery in LC-MS Data using GP. IEEE Congress on Evolutionary Computation June 20 23, Cancún, México, pages: 584-591.

**[96]** Sophie, L. (2013). Etude du rôle du gène PROX1 dans le diabète de type 2. Thèse de doctorat, université du droit et de la sante LILLE 2.

**[97]** Souquet, A. et Radet, F.-G. (2004). Algorithmes génétiques. TE de fin d'année, Tutorat de Mr Philippe Audebaud.

**[98]** Stone, M. (1997). An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion. Journal of the Royal Statistical Society B, vol. 39, no. 1, pages 44-47.

**[99]** Sultan H.-A. et Mohammed E. (2009). Bio-inspired Machine Learning in Microarray Gene Selection and Cancer Classification. Signal Processing and Information Technology (ISSPIT), 2009 IEEE International Symposium, pages : 339-343.

**[100]** Thomas, A., Thibault,H., Yves, V.-P, Pierre, D.et Saeys, Y. (2009). Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. Vol. 26 no. 3 2010, pages 392–398 doi:10.1093/bioinformatics/btp630.

**[101]** Timothée, L. et Vincent, C. (2005). Introduction à la Biologie Moléculaire. Cours, *pimprenelle.lps.ens.fr/biolps/sites/default/files/teaching/.../my_biomol.pdf*

**[102]** Tom, F. (2006) . An introduction to ROC analysis. Pattern Recognition Letters 27 ,pages: 861–874.

**[103]** Wael, A., Taghi, M.- K., David, D., Randall, W. et Amri, N. (2012). A Review of the Stability of Feature Selection Techniques for Bioinformatics Data. Information Reuse and Integration (IRI), 2012 IEEE 13th International Conference, pages : 356-363.

**[104]** Wai-Ki, Y., Samir, B.-A., et Cheng, L. (2011). A Survey of Classification Techniques for Microarray Data Analysis. Springer Handbooks of Computational Statistics ,pages : 193-223

**[105]** Wang, H., Khoshgoftaar, T.-M., Wald, R. et Napolitano,A. (2012). A novel dataset-similarity-aware approach for evaluating stability of software metric selection techniques. In Proceedings of the IEEE International Conference on Information Reuse and Integration (IRI 2012).

**[106]** Wei, F., Xingzhi, C., Xiaoquan, S., Jian, X., Deli, Z. et Kang, N. (2012). A machine learning framework of functional biomarker discovery for different microbial communities based on metagenomic data. IEEE 6th International Conference on Systems Biology (ISB).

**[107]** Werner, D., Olaf, W., Kwang-Hyun, C. et Hiroki, Y. (2013). Genetic Algorithm. Springer Science+Business Media LLC, pages : 788 :873

**[108]** Westeel,V., CHU et Besançon.(2012). Les Rencontres de la Cancérologie Française. ATELIER : médicine personnalisée : état des lieux et perspectives

**[109]** Xin, Z. et David, P.-T. (2007). MSVM-RFE: extensions of SVM-RFE for multiclass gene selection on DNA microarray data. Vol. 23 no. 9 2007, pages 1106–1114, doi:10.1093/bioinformatics/btm036

**[110]** Yuan, T. et Zhifa, L. (2013). Feature selection and prediction with a Markov blanket structure learning algorithm. Proceedings of the 12th Annual UT-ORNL-KBRIN Bioinformatics Summit 2013. http://www.biomedcentral.com/1471-2105/14/S17/A3

**[111]** Zexuan, Z., Yew-Soon, O. et Manoranjan, D. (2006). Markov Blanket-Embedded Genetic Algorithm for Gene Selection. Elsevier Science.

**[112]** Zhang, X. (2006) . Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data. BMC Bioinformatics, 7, 197.

**[113]** http://www.news-medical.net/health/What-is-Molecular-Biology.aspx, consulter le 25/02/2014

**[114]** http://tpe-nano-1s3.weebly.com/la-puce-agrave-proteacuteines.html