الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire
وزارة التعليم العالي و البحث العلمي
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

كلية علوم الطبيعة و الحياة
Faculté des Sciences de la Nature et de la Vie

جامعة الإخوة منتوري قسنطينة
Université Frères Mentouri Constantine 1

Mémoire

Présenté en vue de l'obtention du diplôme de Master
Filière : Sciences Biologiques
Spécialité : Bioinformatique

Intitulé

Nouvelle approche basée sur l'optimisation par essaim de particule pour la détection des ilots CpG dans le génome humain.

Présenté et soutenu :
Le : 30_09_2021
Par :
Bourghoud Yasmine & Moumene Ferial

Devant le jury composé de :

Présidant : Dr. Kamel KELLOU ; Université Frères Mentouri Constantine 1.

Encadreur : Dr. AmiraGHERBOUDJ; Université Frères Mentouri Constantine 1.

Examinateur : Dr. Hamza CHEHILI; Université Frères Mentouri Constantine 1.

Année universitaire : 2020 / 2021

Remerciement

On remercie Dieu le tout puissant de nous avoir donné la santé, la force et la volanté d'entamer et de terminer ce mémoire.

Tout d'abord, ce travail ne serait pas aussi riche et n'avait pas pu avoir le jour sans l'aide de l'encadrement de Mme Gherboudj Amira, on la remercie pour la qualité de son encadrement exceptionnel, pour sa patience, sa rigueur et sa disponibilité durant notre préparation de ce mémoire.

Nous remercions Professeur Mr Kellou Kamel pour avoir accepté de présider nos jury. On remercie également Docteur: Chehili Hamza, membres du jury d'avoir accepté l'éxamination et l'évaluation de ce travail.

On adresse aussi nos vifs remerciements à nos chers pères et à nos chères mères pour leurs encouragements et le soutien affectif et matériel qu'ils nous ont apporté tout au long de notre existence.

Nous remercions aussi mes frères, mes sœurs, mes chers, mes collègues, ainsi que toutes les personnes qui m'ont apporté un soutien moral de loin ou de près.

Nos remerciements s'adressent également à tous nos professeurs pour leurs générosités et la grande patience dont ils ont su faire preuve malgré leurs charges académiques et professionnelles.

Table des matières

Chapitre 3 :Etude experementale

Liste des figures

Liste des tableaux

PSO: L'optimisation par essaim de particules (de l'anglais: ParticleSwarmOptimization).

Kb : kilo base (103 paires de bases).

Mb : Méga base (106 paires de bases.

Pb : paire de base.

TS : Gène Tissu-Spécifique.

CGI : îlot CpG (CpG Islands).

ORI : origine de réplication.

Mb : Méga paire base (106 paires de bases).

Pg : pico grammes.

RBS : site de fixation de ribosome.

ORI : origine de réplication d'initiation.

CpG : Cytosine_Phosphate_Guanine.

CG : Cytosine Guanine.

ApT : Adenine_Phpsphate_Thymine.

TpG : Thymine_Phosphate_Guanine.

CpA : Cytosine_Phosphate_Adenine.

CpGoe : rapport entre le nombre de CpG observés et le nombre de CpGattendus sous une distribution aléatoire des nucléotides.

SWM : Sliding Window Method.

Résumé

Dans ce mémoire nous avons abordé le sujet du cancer Le terme « cancer » englobe un groupe de maladies se caractérisant par la multiplication et la propagation anarchiques de cellules anormales. Si les cellules cancéreuses ne sont pas éliminées, l'évolution de la maladie va mener plus ou moins rapidement au décès de la personne touchée.

Un cancer peut être dû à des facteurs externes (mode de vie, facteurs environnementaux ou professionnels, infections), ou internes (mutations héréditaires, hormones, dérèglement du système immunitaire, etc.). Ces facteurs de risque peuvent agir ensemble ou de façon successive, et enclencher ou favoriser le développement du cancer. Souvent, plusieurs dizaines d'années séparent l'exposition à des facteurs externes et le déclenchement de la maladie.

Un cancer peut être soigné par un ou une combinaison de plusieurs traitements (chirurgie, radiothérapie, chimiothérapie, hormonothérapie, immunothérapie ou traitement ciblé).

Nous avons également établi quelques techniques pour la prédication du cancer avec quelques méthodes de résolutions que nous avons abordé.

Mots clés :cancer, mutations héréditaires, traitement ciblé.

ملخص

في هذه الأطروحة تطرقنا إلى موضوع السرطان. يشمل مصطلح «السرطان » مجموعة من الأمراض التي تتميز بالتكاثر والانتشار غير المنضبط للخلايا غير الطبيعية. إذا لم يتم القضاء على الخلايا السرطانية ، فإن تطور المرض سيؤدي عاجلاً أم آجلاً إلى وفاة الشخص المصاب.

يمكن أن يكون السرطان ناتجًا عن عوامل خارجية (نمط الحياة ، عوامل بيئية أو مهنية ، عدوى) ، أو داخلية (طفرات وراثية ، هرمونات ، اضطراب جهاز المناعة ، إلخ). يمكن أن تعمل عوامل الخطر هذه معًا أو بالتتابع ، وتحفز أو تعزز تطور السرطان. في كثير من الأحيان، يستغرق الأمر عدة عقود بين التعرض للعوامل الخارجية وظهور المرض.

يمكن علاج السرطان بواحد أو مجموعة من العلاجات المتعددة (الجراحة ، العلاج الإشعاعي ، العلاج الكيميائي ، العلاج الهرموني،) العلاج المناعي أو العلاج الموجه.

لقد أنشأنا أيضًا بعض التقنيات للتبشير بالسرطان جنبًا إلى جنب مع بعض طرق الحل التي ناقشناها.

الكلمات المفتاحية: السرطان ، الطفرات الوراثية ، العلاج الموجه.

Summary

In this thesis we touched on the topic of cancer. The term “cancer” includes a group of diseases characterized by the uncontrolled proliferation and proliferation of abnormal cells. If the cancer cells are not eliminated, the disease progression sooner or later will lead to the death of the affected person.

Cancer can be caused by external factors (lifestyle, environmental or occupational factors, infection), or internal (genetic mutations, hormones, immune system disorder, etc.). These risk factors can act together or in sequence, and induce or promote the development of cancer. Often, it takes several decades between exposure to external factors and the onset of disease.

Cancer can be treated with one or a combination of several treatments (surgery, radiotherapy, chemotherapy, hormonal therapy), immunotherapy or targeted therapy.

We have also created some techniques for cancer preaching along with some of the solution methods we have discussed.

Keywords: cancer, genetic mutations, targeted therapy.

Introduction générale

Le cancer est l'une des maladies les plus répandues au monde, il fait partie des causes principales de mortalité actuelle. Selon l'organisation mondiale de la santé, le nombre de personnes qui seront atteintes d'un cancer dans le monde aura doublée 2020, passant à environ 15 millions d'individus touchés.

Traiter toutes les formes de cancers ainsi qu'une meilleure détection des cancers sont les clés pour augmenter les chances de survie d'un individu. Le taux de réussite des traitements offerts, notamment la chimiothérapie et la radiothérapie, dépendent du type de cancer diagnostiqué ainsi que de son stade de détection. Les chances de survie d'un individu atteint d'un cancer dépendent directement de la détection à des stades primaires, soit en début de maladie.

L'une des problématiques de la détection d'un cancer provient de la nécessité d'avoir des cellules cancéreuses en quantités suffisantes. En effet, les diverses techniques qui sont proposées, telles que l'imagerie de résonance magnétique, les frottis et le scanner par exemple, nécessitent la présence d'une masse de cellules ou une tumeur. En ayant la présence d'une tumeur pour obtenir un diagnostic, on diminue déjà les chances de survie d'une personne.

Plusieurs groupes de recherches se sont penchés sur diverses manières de détecter un cancer ou un cancer potentiel en étudiant certaines altérations de l'ADN humain. Ces altérations ne modifient aucunement la séquence de l'ADN mais plutôt la manière de s'exprimer de cette dernière et sont communément appelés facteurs épigénétiques.

La plupart des chercheurs assimilent l'épigénétique à l'étude des systèmes de marquage de la chromatine (méthylation de l'ADN, modification de l'histone et remodelage de la chromatine). La méthylation de l'ADN dans le génome d'un mammifère fait référence à la méthylation de la cytosine dans un site CG. Les îlots CpG sont fortement enrichis en régions nucléotidiques CG. Ces sites CpG sont des segments courts où un nucléotide cytosine est suivi d'un nucléotide guanine dans la direction 5 'à 3'.

Les îlots CpG jouent un rôle important dans la méthylation de l'ADN. La plupart des îlots CpG sont des sites d'initiation de la transcription. La méthylation d'un promoteur de gène semble être étroitement associée à l'inactivation ou l'inhibition de la transcription de ce gène, mais le non méthylation de ces sites promoteurs (îlots CpG non méthylés) d'un gène non transcrit peut induire sa transcription. L'hyperméthylation des îlots CpG situés dans les régions promotrices des gènes suppresseurs de tumeurs est maintenant fermement établie

entant que mécanisme important d'inactivation des gènes. L'hyperméthylation des îlots CpG a été décrite dans presque tous les types de tumeurs. L'élaboration de profils d'hyperméthylation des îlots CpG pour toutes les formes de tumeurs humaines a permis de recueillir des données cliniques pilotes utiles pour la surveillance et le traitement des patients cancéreux. Par conséquent, l'identification des îlots CpG est une tâche très importante en bioinformatique. En effet, cela facilite la tâche aux biologistes en limitant le nombre de biomarqueurs à identifier.

Notre manuscrit est organisé en trois chapitres :

Chapitre1 : présente le contexte biologique de notre travail à savoir, les concepts biologiques nécessaires pour la compréhension du problème traité : la génomique, les îlot CpG, l'épigénétique, la méthylation, les îlots CpG méthylés et leur relation avec les tumeurs.

Chapitre 2 : dans ce dernier nous avons présenté des notions en relation avec le domaine de l'optimisation ainsi que les méthodes de détection des îlots CpG qui ont été présentées dans la littérature.

Chapitre 3 : nous avons proposé notre approche de détection des ilots CpG qui consiste à une hybridation de la méthode CpG Cluster et la métaheuristique d'optimisation par essaim de particule (dite « PSO » : Particule Swarm Optimization). Ensuite, nous avons effectué une comparaison entre nos résultats et les résultats de cinq autres méthodes présentées dans la littérature.

CHAPITRE 1 :

ILOTS CPG DANS LE GENOME HUMAIN

1 INTRODUCTION

Un dinucléotide CpG, parfois appelé site CpG en référence à l'anglais CpG site, est un segment d'ADN de deux nucléotides dont la séquence de bases nucléiques est CG. La notation « CpG » est une abréviation de cytosine–phosphate–guanine destinée à être clairement distinguée de la notation « CG » qui peut également désigner une paire de bases sur deux brins d'ADN distincts et non la séquence d'un brin d'ADN donné. Dans les génomes, les dinucléotides CpG ont une distribution différente de celle d'autres dinucléotides comme CpG, ApT ou TpA, car ils définissent des îlots CpG dans lesquels leur concentration est bien plus élevée et qui jouent un rôle dans la régulation de l'expression génétique.

2 GENOMIQUE

La génomique est une discipline de la biologie moderne. Elle étudie le fonctionnement d'un organisme, d'un organe, d'un cancer, etc. à l'échelle du génome, et non plus limitée à celle d'un seul gène. La génomique se divise en deux branches :

- La génomique structurale, qui se charge du séquençage du génome entier ;
- La génomique fonctionnelle, qui vise à déterminer la fonction et l'expression des gènes séquencés en caractérisant le transcriptome et le protéome.

L'essor de cette discipline a été facilité par le développement des techniques de séquençage des génomes et la Bioinformatique.

- En 1972, le premier véritable séquençage d'un génome est publié, avec la lecture de la séquence ARN du gène du virus bactériophage MS2.
- Elle a été très médiatisée à la fin du XX^{e} siècle avec la compétition entre différentes équipes scientifiques pour la publication de la première carte du génome humain, annoncée conjointement le 26 juin 2000 par Bill Clinton et Tony Blair.
- Depuis, un nombre croissant de génomes complets sont séquencés chez des espèces vivantes très différentes : le ver *Caenorhabditis Elegans* en 1998, la mouche drosophile et la plante *Arabidopsi sthaliana* en 2000 ou encore, le chien en 2005. En septembre 2007, une équipe menée par le biologiste et entrepreneur Craig Venter a publié le premier génome complet d'un individu qui se trouve être Craig Venter lui-même. Le génome du co-découvreur de la structure de l'ADN et ancien directeur du Projet génome humain, James Watson, a aussi été séquencé dans son intégralité à la même période **[1]**.

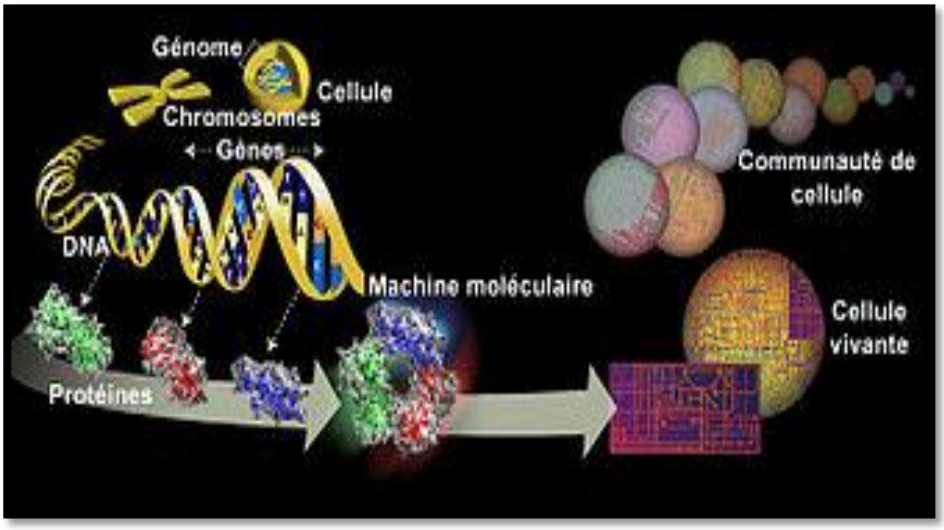


Figure 1:Illustration de l'ADN a la vie, chez l'humain [1].

Le génome est constitué d'un ou plusieurs chromosomes dont le nombre total dépend de l'espèce considérée, chaque chromosome étant constitué d'une unique molécule d'ADN, linéaire chez les eucaryotes et le plus souvent circulaire chez les procaryotes. Chaque chromosome peut être présent en un ou plusieurs exemplaires, le plus souvent deux chez les espèces sexuées, l'un d'origine maternelle et l'autre d'origine paternelle (organisme diploïde).

2.1 Utilité de la Génomique

Connaitre la séquence nucléotidique permet de multitude études :

- Exploration des fonctions associées aux gènes, aux variations alléliques et au polymorphisme génétique.
- Reconstruction d'arbres phylogénétique d'espèces vivantes.
- Analyse de l'histoire évolutive des êtres vivants en lien avec leur écosystème, Compréhension de maladies liées aux gènes**[1]**.

2.2 Post Génomique

La post-génomique est une approche fonctionnelle qui vise à offrir une image globale du fonctionnement des êtres vivants. Elle implique l'étude de l'expression simultanée de l'ensemble des gènes, la description de la totalité des interactions qui se forment entre les produits de ces gènes au niveau cellulaire, mais aussi au niveau des organes ou d'un organisme. Complémentaire de la génomique, la post-génomique va plus loin dans la compréhension du fonctionnement de la cellule. Elle cherche à savoir quand et dans quelles

conditions un gène s'exprime pour enclencher la fabrication de protéines, et quelle est leur activité**[1]**.

3 GENOME

Le génome est l'ensemble du matériel génétique d'un individu ou d'une espèce codé dans son ADN (à l'exception de certains virus dont le génome est porté par des molécules d'ARN). Il contient en particulier toutes les séquences codantes (transcrites en ARN messagers, et traduites en protéines) et non-codantes (non transcrites, ou transcrites en ARN, mais non traduites).

Est souvent comparé à une encyclopédie dont les différents volumes seraient les chromosomes. Les gènes seraient les phrases contenues dans ces volumes et ces phrases seraient écrites dans un langage génétique représenté par quatre bases (adénine, guanine, cytosine et thymine) abrégées en AGCT .Les génomes dans le monde vivant :

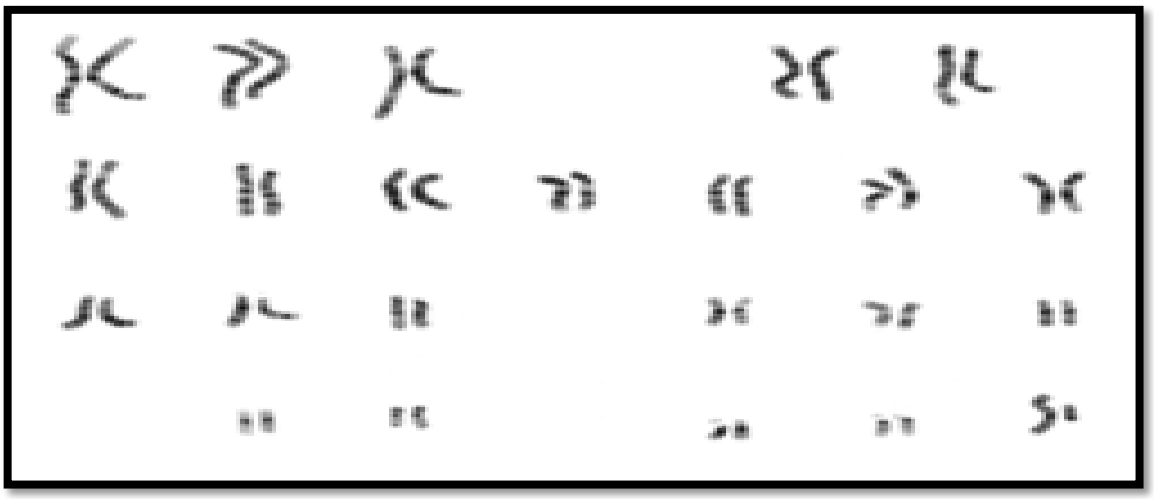


Figure 2: Les 46 chromosomes qui forment le caryotype du génome humain[2].

Chez les virus, le génome est contenu soit dans une (ou plusieurs) molécule(s) d'ADN ou d'ARN, à simple ou double brin.

Chez les procaryotes (bactéries et archées), le génome est généralement contenu dans une molécule d'ADN circulaire. Peut aussi exister un génome extra chromosomique, contenu dans des plasmides et des épisomes.

Chez les eucaryotes, on distingue :

Le génome nucléaire, contenu dans le noyau qui caractérise les eucaryotes. C'est de ce génome dont on parle en général quand on parle du génome d'un eucaryote (animal, plante, champignon, etc.).

Les génomes non-nucléaires, contenus dans des organites sont :

- le génome mitochondrial, contenu dans les mitochondries, chez laquais totalité des eucaryotes.

- le génome chloroplastique, contenu dans les chloroplastes, chez les algues et les plantes supérieures.

Chez quelques eucaryotes (par exemple la levure) sont aussi présents des plasmides (de taille réduite).

Chez l'homme en particulier (organisme eucaryote), le génome nucléaire est réparti sur 46 chromosomes, soit 22 paires d'autosomes et deux gonosomes (XX chez la femme, XY chez l'homme). Il ne faut pas confondre le génome et le caryotype, qui caractérise les chromosomes**[2]**.

3.1 L'annotation des génomes

L'annotation d'un génome consiste à traiter l'information brute contenue dans la séquence dans le but de :

Prédire, le contenu en gènes, la position des gènes à l'intérieur d'un génome (le début, la fin, et chez les eucaryotes, les introns et les exons), ainsi que leur organisation (gènes uniques ou en opéron, avec des séquences promotrices, des terminateurs, des sites de fixation ribosomaux (RBS) …). Dans ce cas, on parle d'annotation structurale.

Prédire la fonction potentielle de ces gènes (leur attacher une étiquette, portant leur nom probable, leur fonction probable, leurs interactions probables). Dans ce cas on parle d'annotation fonctionnelle **[2]**.

3.2 Taille du génome

La taille du génome se mesure en nombre de nucléotides, ou bases. La plupart du temps, on parle de Pb (pour *paire de bases*, puisque la majorité des génomes est constituée de doubles brins d'ADN ou bien d'ARN). On emploie souvent les multiples kb (pour kilo-base) ou Mb (méga-base), qui valent respectivement 1 000 et 1 000 000bases. La taille du génome peut aussi être exprimée en pg (pico-grammes), ce qui correspond à la masse d'ADN (haploïde) par cellule. 1 pg représente environ 1 000 Mpb.

La taille du génome peut varier de quelques kilo-bases chez les virus à plusieurs centaines de milliers de Mb chez certains eucaryotes. La quantité d'ADN, contrairement à ce

qui a été longtemps supposé, n'est pas proportionnelle à la complexité d'un organisme ; ainsi, l'amibe Amoebadubia, un organisme unicellulaire, a un génome environ 200 fois plus grand qu'Homosapiens. Ce constat est fréquemment appelé paradoxe de la valeur C **[2]**.

4 EPIGENETIQUE

L'épigénétiques est l'étude de la relation entre génotype (l'information du génome d'un individu) et phénotype (l'ensemble des caractéristiques observables de l'organisme de l'individu), C'est aussi la discipline de la biologie qui étudie la nature des mécanismes modifiant de manière réversible, transmissible (lors des divisions cellulaires) et adaptative l'expression des gènes sans en changer la séquence nucléotidique (ADN).Alors que la génétique correspond à l'étude des gènes, l'épigénétiques s'intéresse à une « couche » d'informations complémentaires qui définit comment ces gènes vont être utilisés par une cellule ou ne pas l'être **[4]**. », « C'est un concept qui dément en partie la « fatalité » des gènes **[5]**. »

4.1 Les mécanismes épigénétiques

Pour chaque être vivant, l'information génétique est portée par l'ADN, dont la séquence est identique dans toutes les cellules d'un même organisme. Elle est codée par l'enchaînement spécifique des quatre bases nucléiques :

Adénine, Thymine, Cytosine et Guanine A-T-C-G.

Le décryptage de ce code à quatre lettres ne permet pas d'expliquer comment une même succession peut donner autant de combinaisons. Un des exemples les plus parlants est qu'à partir d'une cellule souche unique naît un organisme entier composé de cellules différentes, tant au niveau de leur structure que de leur fonction. Ces cellules ayant toutes la même origine possèdent exactement le même code génétique, qui est interprété différemment grâce à une batterie de mécanismes épigénétiques.

Chaque cellule exploite plusieurs types de mécanismes épigénétiques indépendants, dont le mode d'action fait intervenir l'ADN et des protéines.

En effet, pour que le noyau d'une cellule de 5 à 6 micromètres de diamètre puisse contenir deux mètres d'ADN, plusieurs niveaux de compaction sont nécessaires.

Certains mécanismes épigénétiques agissent sur la structure de la chromatine, en la faisant passer d'un état condensé à un état décondensé, ou inversement, selon qu'un gène a besoin d'être exprimé ou réprimé.

D'autres interviennent directement au niveau de séquences régulatrices de l'ADN au voisinage des gènes.

Ces séquences particulières ne codent pour aucune protéine, mais contrôlent où et quand les gènes sont exprimés.

La double hélice d'ADN s'enroule autour d'octamères de protéines, appelées histones, pour former des nucléosomes.

Puis, ces structures sont organisées dans l'espace pour former des fibres de chromatine plus ou moins denses.

L'état très compact est appelé hétéro chromatine et empêche la transcription des gènes. L'état le moins condensé, l'euchromatine, contient la portion active du génome.

5 ILOTS CPG : STRUCTURES ATYPIQUES DES GENOMES DE MAMMIFERES

Les génomes eucaryotes présentent différents patrons de méthylation. Certaines régions du génome sont méthylées, au milieu, des fois tout le génome ou presque est méthyle. Le ‘presque’ consistes-en de courtes régions génomiques (environ 1 kb), riches en dinucléotides CpG, qui échappent à la méthylation dans de nombreux tissus, dont la lignée germinale.

Ces îlots CpG ont été très étudiés depuis leur découverte il y a une trentaine d’années,

Particulièrement chez les mammifères, du fait de leurs implication potentielle dans la régulation de l’expression des gènes, et parce qu’une méthylation aberrante de ces îlots CpG a été observée dans certains cancers.

La détection de courtes séquences particulières enchâssées dans un génome globalement méthylé n’est pas forcément triviale, et nous présenterons dans cette partie de différents moyens de détection d’îlots CpG. Nous discuterons ensuite des rôles potentiels des îlots CpG, notamment de leur lien avec la transcription et la réplication **[3]**.

5.1 Découverte et détection des îlots CpG

Cooper, Bird et collègues[4]ont mis en évidence au début des années 1980 qu’unepetite fraction (environ 1%) des génomes de six vertébrés (poulet, homme, souris, couleuvre, xénope, et truite) étaient particulièrement riche (environ quinze fois plus que le reste du génome) en sites de coupure de l’enzyme HpaII, dans les tissus somatiques comme dans le sperme. HpaII est une enzyme de restriction qui coupe l’ADN sur les sites CCGG quand le C n’est pas méthylé, et permet donc d’estimer la fraction de dinucléotides CpG non méthylés.

Une étude approfondie de certains fragments a permis de confirmer que ces régions HTF (HpaIItiny fraction) étaient effectivement non méthylées. Ces 'îlots HTF', riches en G et C, et surtout en dinucléotides CpG, ont été appelés par la suite îlots CpG**[3]**.

5.1.1 Détection des îlots CpG

Les méthodes les plus classiquement utilisées s'inspirent d'une étude réalisée dans laquelle les auteurs identifient les îlots CpG sur la base de trois critères : longueur des séquences supérieure à 200 paires de bases (pb), taux de G + C supérieur à 50%, et rapport CpG observé/CpGattendu, noté CpGo/e, supérieur à 0, 6. Cependant, l'utilisation de ces critères entraîne la détection d'éléments transposables riches en G et C, et en CpG. Par exemple, chez l'homme, les séquences détectées correspondent notamment à des séquences Alu, insérés relativement récemment dans le génome. Leur richesse en CpG indiquerait qu'ils dérivent d'un élément source riche en CpG. Ponger & Mouchiroud**[5]**proposent de considérer les îlots CpG de longueur supérieure à 500 pb, et d'effectuer la détection d'îlots sur des séquences masquées pour les éléments répétées. Le logiciel que ces auteurs proposent, CpG ProD, est particulièrement dédié à la détection d'îlots CpG promoteurs de gènes. C'est pourquoi ils associent en plus à chaque îlot CpG détecté un premier score, calculé à partir de caractéristiques de séquences, qui donne la probabilité pour que cet îlot CpG soit associé à un promoteur, ainsi qu'un second score qui indique son sens de transcription probable.

Ces restrictions de critères permettent de diminuer la proportion d'ET définis comme îlots CpG, et d'augmenter la proportion de 'vrais' îlots CpG, c'est-à-dire ayant un rôle régulateur, notamment les îlots CpG promoteurs**[6]**.

5.2 Localisation génomique des îlots CpG

Les études précédentes soulignent que les îlots CpG se distinguent par des caractéristiques particulières (richesse en GC, richesse en CpG, non méthylation), mais également par une localisation dans des régions spécifiques.

5.2.1 Ilots CpG et promoteurs

En effet, loin d'être répartis de manière uniforme, les îlots CpG sont associés étroitement aux séquences promotrices, et que ces derniers portaient préférentiellement des marqueurs de chromatine activement transcrite par rapport au reste du génome. Si environ la moitié des îlots CpG sont associés aux promoteurs de gènes, la réciproque est également vraie, c'est à dire que plus de la moitié des gènes sont associés à des îlots CpG, les gènes dont l'expression est

restreinte à un ou quelques tissus sont moins souvent associés à des îlots CpG. Différentes études estiment qu'environ 40% des gènes TS sont concernés par cette association, leur calcul se basait sur une estimation du nombre d'îlots CpG dans le génome à partir du nombre de fragments de restriction par digestion (environ 45 000 îlots CpG). Cependant, leur étude ne prenait pas en compte le fait que tous les îlots CpG n'étaient pas associés à des gènes. Donc quel est le rôle de ces îlots CpG associés aux promoteurs, s'ils en ont un ? Plusieurs études ont montré que la méthylation d'un îlot promoteur entraînait la répression de l'expression du gène associé. La méthylation des îlots CpG est notamment associée in vivo à la répression de l'expression des gènes sur le chromosome X inactivé et de l'expression de nombreux gènes soumis à l'empreinte. Il serait donc tentant de considérer que les îlots CpG servent de régulateurs fins de la transcription, c'est-à-dire qu'ils sont non méthylés lorsque le gène correspondant est activement transcrit, et méthylés pour réprimer la transcription. Dans ce sens, ils montrent que la plupart des gènes testés qui sont spécifiques de la lignée germinale ont un îlot promoteur méthylé dans les tissus somatiques. Cependant, la corrélation globale est relativement faible entre l'hyperméthylation des îlots promoteurs et la transcription ou non des gènes correspondants. D'une part, la régulation de la transcription d'un gène peut être faite par un îlot CpG qui ne se situerait pas sur le promoteur. De plus, les chercheurs proposent que la méthylation soit un phénomène qui intervient de manière stochastique sur l'ADN des promoteurs inactifs, qui ne sont donc plus protégés de la méthylation**[7]**.

5.2.2 Ilots CpG et origines de réplication

En plus de leur association particulière avec les promoteurs de gènes, les îlots CpG semblent particulièrement liés aux origines de réplication. La réplication de l'ADN, qui précède chaque division cellulaire (à l'exception de la seconde division de méiose) se fait pour les génomes eucaryotes par l'activation de différents points d'initiation, appelés origine de réplication (ORI). L'analyse des intermédiaires de réplication pour trois gènes de hamster et un gène humain a montré que les îlots CpG étaient présents dans les brins naissants très courts. Ces auteurs **[8]**ont donc proposés que ces îlots CpG soient des origines de réplication. Par la suite, ont proposé un modèle selon lequel les îlots CpG seraient la trace de promoteurs associés à des origines de réplication dans les cellules de l'embryon précoce. Selon leur modèle, la fourche de réplication serait initiée de manière unidirectionnelle, le long d'un promoteur actif, avant son élongation bidirectionnelle **(voir. figure 3).** De plus, ont proposé un jeu d'origines de réplication couvrant 0,4% du génome de la souris, et ont montré que les ORI les plus efficaces étaient celles qui Co localisés avec les îlots CpG promoteurs. Le

nombre croissant de données confirme donc une association étroite entre îlots CpG et ORI, même si on ne peut pas dire pour le moment si tous les îlots CpG sont ou non associés aux origines de réplication. Cependant, les mécanismes responsables de cette association sont peu clairs**[9]**.

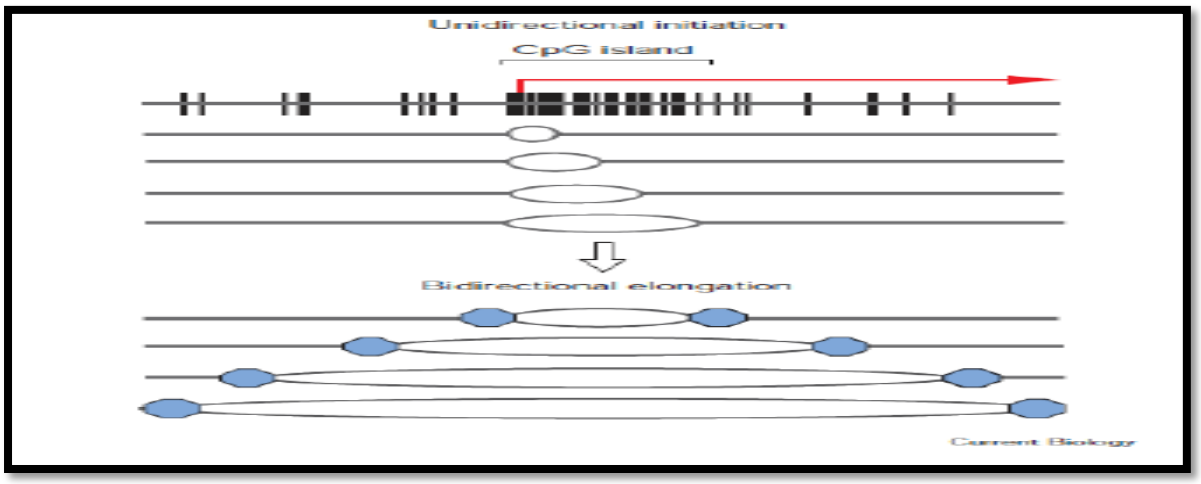


Figure 3 : Modèle de réplication unidirectionnelle sur les ilots CpG[10].

5.2.3 Ilots CpG et recombinaison

Han et ces collègues**[11]**ont effectué une analyse comparative des îlots CpG dans les génomes de dix mammifères, en utilisant l'algorithme de Jones et Takai[12]pour la détection de ces îlots CpG. Ils ont observé que la densité en îlot CpG était corrélée positivement avec le nombre de chromosomes, et inversement proportionnellement avec leur longueur. La densité en îlot CpG est également corrélée positivement au taux de recombinaison (lui-même lié aux facteurs précédents). Enfin, les îlots CpG sont plus souvent localisés à l'extrémité des chromosomes, ces régions étant par ailleurs connues pour être plus recombinantes. Les îlots CpG semblent donc être associés à des régions de recombinaison élevée **[3].**

5.3 Origine des îlots CpG

Les îlots CpG sont des régions majoritairement non méthylées, dans la plupart des tissus. Leur richesse en CpG indique qu'ils sont moins méthylés dans la lignée germinale et/ou dans les cellules embryonnaires précoces, accumulant ainsi moins de substitutions sur les sites CpG. Par ailleurs, le fait que les îlots CpG soient non méthylés dans la plupart des tissus suggère que le patron de non méthylation de ces îlots CpG pourrait être mis en place avant la différentiation des lignées cellulaires. Les génomes de mammifères subissent une phase de déméthylation / reméthylation durant le développement précoce de l'embryon, puis dans les

cellules précurseur de la lignée germinale. Le statut de non méthylation des îlots CpG doit donc être ré-établi à ces deux moments, protégeant les îlots CpG contre la reméthylationglobale. L'état de non méthylation se perpétue, et les îlots ne subissent pas de perte massive de dinucléotides CpG due à la méthylation. Par ailleurs, l'ablation de ces sites de fixation facilite la méthylation d'îlot promoteur du gène **[13].**

5.3.1 Devenir des îlots CpG

Si l'origine des îlots CpG n'est pas encore totalement élucidée, on constate donc qu'ils continuent d'évoluer. Ceci est particulièrement visible avec les études comparatives des îlots promoteurs dans les génomes de l'homme et de la souris. En effet, si l'on suppose que les îlots CpG ont été mis en place tôt au cours de l'évolution des mammifères, alors si on trouve un îlot CpG dans le promoteur d'un gène humain, on s'attend à en retrouver un dans le promoteur du gène murin orthologue. Les chercheurs**[14]**ont montré que c'était effectivement le cas pour certains gènes, mais ils ont également observé que pour d'autres gènes l'îlot CpG semble avoir été perdu dans une des deux lignées. Ces pertes sont plus importantes dans la lignée murine que dans la lignée humaine. En mesurant l'excès de dinucléotide TpG ou CpA par rapport aux CpG, ou en comparant les taux de substitution entre les séquences de l'homme et du chien d'une part et celles de la souris et le chien d'autre part. Ces îlots CpG sont donc en train de perdre peu à peu leurs caractéristiques d'îlots : on parle d'une érosion des îlots CpG**[3]**.

5.3.2 Ilots CpG et recombinaison

Han et ces collègues[11]ont effectué une analyse comparative des îlots CpG dans les génomes de dix mammifères, en utilisant l'algorithme de Jones et Takai **[12]** pour la détection de ces îlots CpG. Ils ont observé que la densité en îlot CpG était corrélée positivement avec le nombre de chromosomes, et inversement proportionnellement avec leur longueur. La densité en îlot CpG est également corrélée positivement au taux de recombinaison (lui-même lié aux facteurs précédents). En fin, les îlots CpG sont plus souvent localisés à l'extrémité des chromosomes, ces régions étant par ailleurs connues pour être plus recombinantes. Les îlots CpG semblent donc être associés à des régions de recombinaison élevée**[3]**.

6 METHYLATION DES ILOTS CPG

Dans le règne animal les cytosines méthylées se trouvent presque invariablement dans le dinucléotide CpG. Chez les vertébrés, ces sites méthylables suivent une distribution non uniforme : il existe des domaines, appelés îlots CpG, où ce dinucléotide est plus fortement représenté. Les îlots CpG correspondent fréquemment au promoteur et au premier exon des gènes et ils ne sont en général pas méthylés dans les cellules saines. En revanche dans les cellules cancéreuses, il est fréquent d'observer des méthylations aberrantes des îlots CpG, corrélées à la répression transcriptionnelle des gènes qui leur sont associés. Même si l'origine du phénomène est débattue, l'importance de la méthylation dans la progression des cancers est bien établie, il y a donc un enjeu important dans la compréhension de la répression des gènes imposée par la méthylation de l'ADN**[15].**

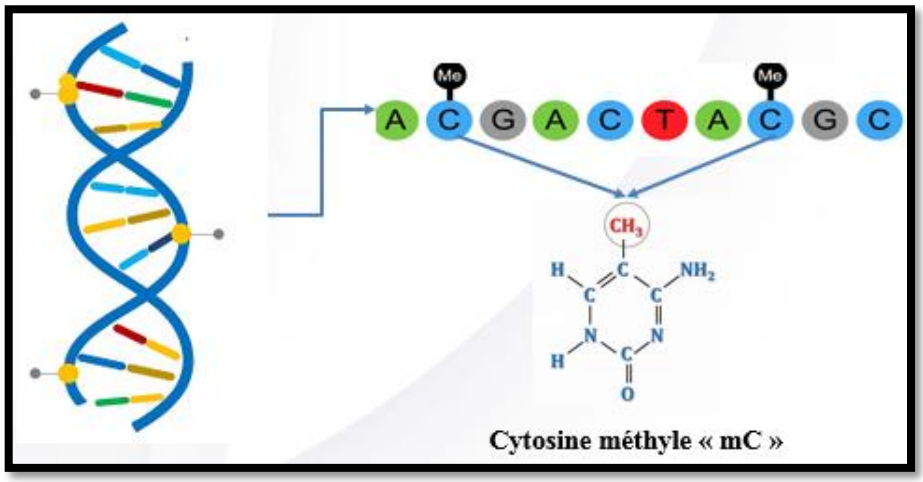


Figure 4 : Illustration du phénomène de méthylation[15].

7 RELATION DES ILOS CPG AVEC LES TUMEURS

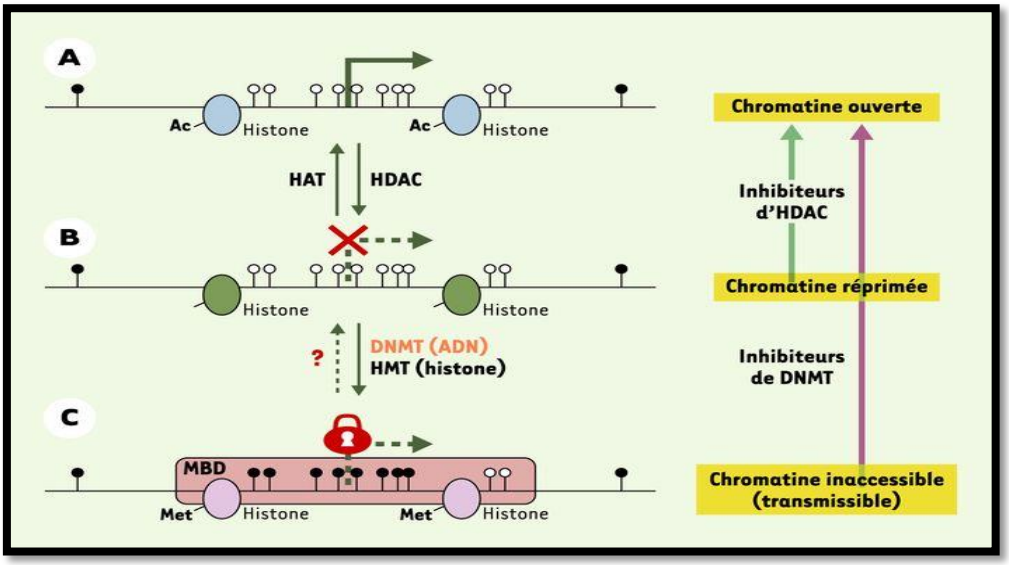


Figure 5:Aspects de la chromatine au niveau de gène suppresseurs de tumeurs, dans les situations normale et tumorale [16].

Lors de la transformation maligne, les cellules accumulent des anomalies épigénétiques comme la méthylation de l'ADN, qui n'affectent pas la séquence d'ADN mais qui sont transmissibles au cours des divisions. La méthylation des îlots CpG a lieu au niveau du promoteur des gènes suppresseurs de tumeur et entraîne une inhibition de leur transcription**(voir figure 5).** Le nombre et le type de gènes méthylés varient en fonction de la localisation tumorale et du type histologique, permettant ainsi de définir un profil de méthylation. Les techniques d'analyse de la méthylation, basées sur un traitement au bisulfite de sodium suivi d'une amplification par PCR ou d'un séquençage, sont suffisamment sensibles et spécifiques pour permettre de détecter ces anomalies non seulement au niveau de la tumeur, mais aussi au niveau des liquides biologiques qui drainent la tumeur ou au niveau de l'ADN circulant.

L'analyse de ces anomalies semble prometteuse pour la détection précoce d'un cancer chez des sujets à risque ou le diagnostic rapide d'une rechute ainsi que pour établir un pronostic ou évaluer la réponse à un traitement. Enfin, la méthylation de l'ADN est réversible sous l'effet d'agents déméthylants, ce qui ouvre de nouvelles perspectives thérapeutiques, en association avec les traitements conventionnels**[16].**

8 CONCLUSION

Le génome est couvert d'un patron de méthylation de l'ADN sauf des régions remarquables les îlots CpG, Ces régions particulièrement riches en dinucléotides CpG, sont associées en partie aux régions clés impliquées dans certaines fonctions cellulaires, comme les promoteurs géniques ou les origines de réplication. Toutefois, on ne sait pas encore clairement si les îlots CpG sont la cause ou la conséquence de cette association.

CHAPITRE 2 :

METHODES D'OPTIMISATION ET DE DETECTION DES ILOTS CPG

1 INTRODUCTION

L'optimisation est une branche des mathématiques cherchant à modéliser, analyser et résoudre analytiquement ou numériquement des problèmes. Elle utilise les techniques de la recherche opérationnelle afin de minimiser ou maximiser une fonction sur un ensemble.

L'idée de combiner des méthodes exactes et / ou des méthodes approchées pour créer de nouvelles méthodes a donné naissance à un pseudo classe de méthodes. C'est la classe de la méthode hybride. Ces dernières ont construit une tendance qui a suscité l'intérêt de plusieurs communautés de chercheurs.

Le principe du théorème «No free lunch » stipule qu'aucune méthode n'est efficace pour tous les types de problèmes. En fait, chaque méthode propose des avantages dont on cherche à maximiser des lacunes dont on cherche à combler. Partant de ce principe, beaucoup de chercheurs ont envisagé la combinaison des méthodes de résolution des problèmes afin de profit des points forts de chacune et de proposer des alternative plus efficaces et plus performantes.

2 PROBLEME D'OPTIMISATION

Un problème d'optimisation consiste à trouver la meilleure solution parmi toutes les solutions réalisables. Les problèmes d'optimisation peuvent être divisés en deux catégories, selon le type des variables qui peuvent être continues ou discrètes.

Un problème d'optimisation avec des variables discrètes est connu comme une optimisation discrète, dans laquelle un objet tel qu'un entier, une permutation ou un graphique doit être trouvé à partir d'un ensemble dénombrable. Un problème avec des variables continues est connu comme une optimisation continue, dans laquelle une valeur optimale d'une fonction continue doit être trouvée**[17].**

3 LES METHODES DE RESOLUTION DE PROBLEMES D'OPTIMISATION

Les méthodes de résolution des problèmes d'optimisation sont divisées en deux classes principales : la classe de méthodes exactes et la classe de méthodes approchées. L'hybridation des méthodes de ces deux classes a donné naissance à une pseudo classe qui englobe des méthodes dites hybrides. La figure 6 montre toutes les catégories**[18].**

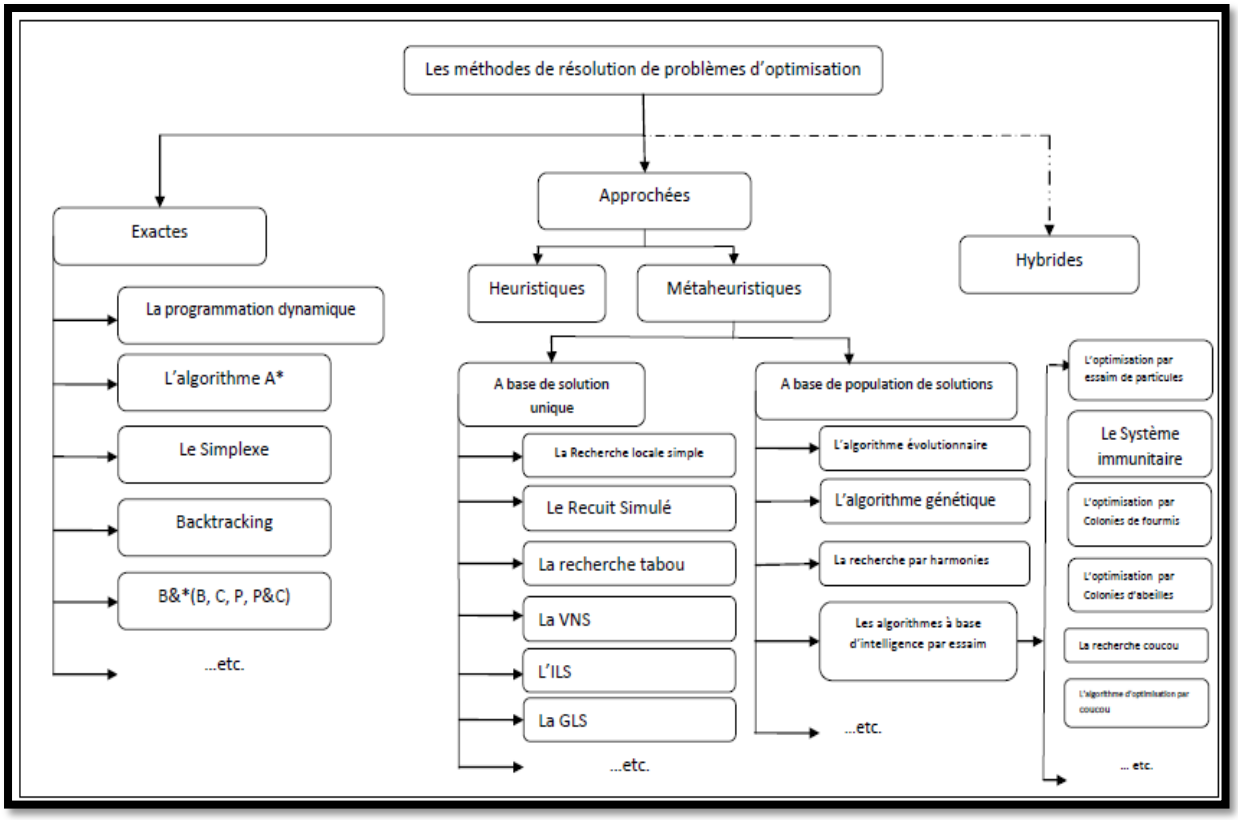


Figure 6: classification de méthodes de résolution de problèmes d'optimisation[18].

4 METHODES EXACTES

Une méthode exacte permet de trouver une solution optimale d'un problème donné. Toutefois, ces méthodes peuvent devenir rapidement couteuses en temps d'exécution, notamment pour les problèmes NP-difficiles**[19].**Il existe de nombreux algorithmes exactes y compris l'algorithme du simplexe, la programmation dynamique, l'algorithme A*, les algorithmes de séparation et évaluation, les algorithmes de retour arrière (Back tracking), sans oublier les algorithmes spécifiques aux problèmes traités comme l'algorithme de Johnson**[20]** pour la résolution de problèmes d'ordonnancement. Notre vocation n'est plus de relater le principe de différentes méthodes exactes mais plutôt d'en citer quelques-unes dans ce qui suit**[21].**

5 METHODES APPROCHEES

Les méthodes approchées permettent de trouver de manière rapide une solution réalisable pour un problème donné. Cependant cette solution n'est pas forcément la solution optimale. Ces méthodes peuvent être classées en deux catégories : méthodes métaheuristiques et méthodes heuristiques. La différence entre les deux est que les métaheuristiques sont

applicables sur de nombreux problèmes. Tandis que, les heuristiques sont spécifiques à un problème donné.

De nombreuses métaheuristiques ont été proposées dans la littérature afin de faire face aux différents problèmes en minimisant ou maximisant le coût de la recherche. Les métaheuristiques peuvent être classées en deux catégories: les méthodes à base de solution unique et les méthodes à base de population de solutions**[22].**

6 METAHEURISTIQUE

Une métaheuristique est un algorithme d'optimisation visant à résoudre des problèmes d'optimisation difficile (souvent issus des domaines de la recherche opérationnelle, de l'ingénierie ou de l'intelligence artificielle) pour lesquels on ne connaît pas de méthode classique plus efficace.

Les métaheuristiques sont généralement des algorithmes stochastiques itératifs, qui progressent vers un optimum global, c'est-à-dire l'extremum global d'une fonction, par échantillonnage d'une fonction objectif. Elles se comportent comme des algorithmes de recherche, tentant d'apprendre les caractéristiques d'un problème afin d'en trouver une approximation de la meilleure solution**[23].**

La figure7 illustre le fonctionnement des métaheuristiques (M) qui sont souvent des algorithmes utilisant un échantillonnage probabiliste. Elles tentent de trouver l'optimum global (G) d'un problème d'optimisation difficile (avec des discontinuités — D —, par exemple), sans être piégé par les optima locaux (L)**[23].**

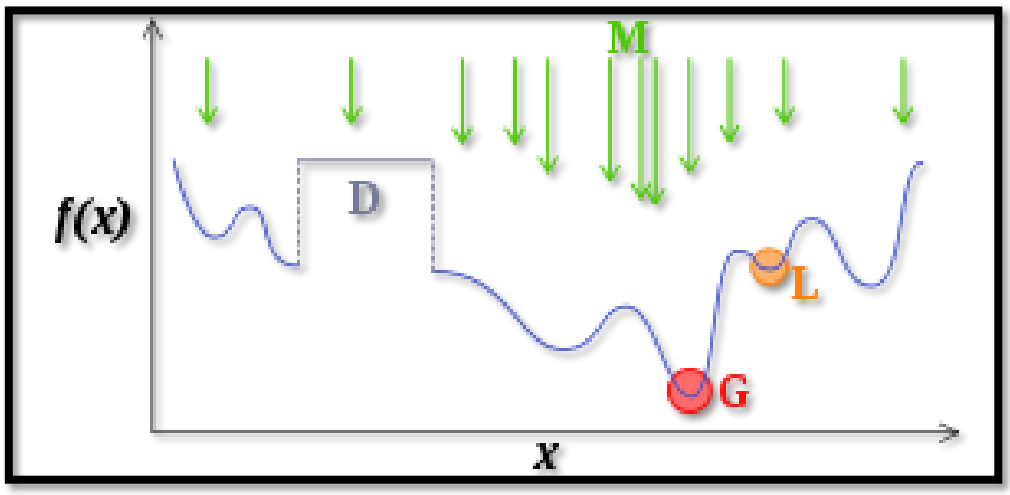


Figure 7 : Echantillonnage probabiliste d'un algorithme métaheuristiques[23].

On distingue deux classes des métaheuristiques que nous présentons ce qui suit de cette section.

6.1 Les métaheuristiques à base de solution unique

Les métaheuristiques à base de solution unique débutent la recherche avec une seule solution initiale. Elles se basent sur la notion du voisinage pour améliorer la qualité de la solution courante. En fait, la solution initiale subit une série de modifications en fonction de son voisinage. Le but de ces modifications locales est d'explorer le voisinage de la solution actuelle afin d'améliorer progressivement sa qualité au cours des différentes itérations. Le voisinage de la solution englobe l'ensemble des modifications qui peuvent être effectuées sur la solution elle-même. La qualité de la solution finale dépend particulièrement des modifications effectuées par les opérateurs de voisinages. En effet, les mauvaises transformations de la solution initiale mènent la recherche vers la vallée de l'optimum local d'un voisinage donné (peut être un mauvais voisinage) ce qui bloque la recherche en fournissant une solution de qualité insuffisante.

De nombreuses méthodes à base de solution unique ont été proposées dans la littérature. Parmi lesquelles: la descente, le recuit simulé, la recherche tabou, la recherche à voisinage variable (VNS: **V**ariable **N**eighbourhood **S**earch), la recherche locale réitérée, la recherche locale guidée …etc **[22].**

6.2 Les métaheuristiques à base de population de solutions

Les métaheuristiques à base de population de solutions débutent la recherche avec une panoplie de solutions. Elles s'appliquent sur un ensemble de solutions afin d'en extraire la meilleure (l'optimum global) qui représentera la solution du problème traité. L'idée d'utiliser un ensemble de solutions au lieu d'une seule solution renforce la diversité de la recherche et augmente la possibilité d'émergence de solutions de bonne qualité. Une grande variété de méthodes basées sur une population de solutions a été proposée dans la littérature, commençant par les algorithmes évolutionnaires, passant par les algorithmes génétiques et arrivant aux algorithmes à base d'intelligence par essaims comme : l'algorithme d'optimisation par essaim de particules, l'algorithme de colonies de fourmis, l'algorithme de colonies d'abeilles, la recherche coucou, l'algorithme d'optimisation par coucou…qui ont connus une investigation remarquable ces deux dernières décennies**[22].**

6.3 Optimisation par essaim de particule

L'optimisation par essaim de particules (OEP ou PSO en anglais) est une métaheuristique d'optimisation, inventée par Eberhart et Kennedy**[24].**Le mot « essaim » est généralement utilisé pour désigner un ensemble fini de particules ou d'agents interactifs. Les oiseaux évoluant en groupes, les bancs de poissons, les colonies de fourmis, les colonies d'abeilles et même les systèmes immunitaires sont des exemples d'essaim.

L'optimisation par essaim de particules s'inspire du comportement social des oiseaux évoluant en groupe et des bancs de poissons. L'algorithme d'optimisation par essaim de particule lance la recherche avec une population de solutions, où chacune est appelée « particule ».

Ainsi, grâce à des règles de déplacement très simples (dans l'espace des solutions), les particules peuvent converger progressivement vers un minimum global. Au départ de l'algorithme chaque particule est donc positionnée (aléatoirement ou non) dans l'espace de recherche du problème **[16].** Chaque itération fait bouger les particules en fonction de 3 composantes(comme le montre la **figure 8**) :

1. Sa vitesse actuelle.
2. Sa meilleure solution.
3. La meilleure solution obtenue par l'essaim.

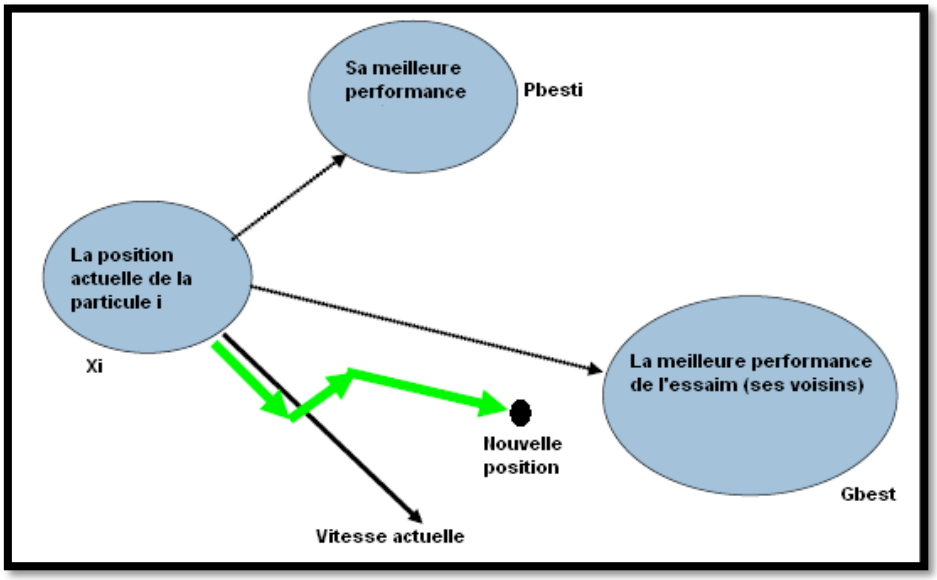


Figure 8 : Déplacement d'une particule.

Cela donne les équations de mouvement suivantes :

$vid(t) = vid(t-1) + c1\ r1\ (pbestid(t-1) - xid(t-1)) + c2\ r2\ (gbestd(t-1) - xid(t-1))$ (2.1)

$xid(t) = xid(t-1) + vid(t)$ (2.2)

Avec :

- xid(t), xid(t-1) : la position de la particule i dans la dimension d aux temps t et t-1, respectivement.
- vid(t), vid(t-1) : la vitesse de la particule i dans la dimension d aux temps t et t-1, respectivement.
- pbestid(t-1) : la meilleure position obtenue par la particule i dans la dimension d au temps t-1.
- gbestd(t-1) : la meilleure position obtenue par l'essaim dans la dimension d au temps t-1.
- c1, c2 : deux constantes qui représentent les coefficients d'accélération, elles peuvent être non constantes dans certains cas [16],[26].
- r1, r2 : nombres aléatoires tirés de l'intervalle [0,1].

v_{id} (t-1), c1 r1 ($pbest_{id}$ (t-1) - x_{id} (t-1)), c2 r2 (gbestd(t-1) - xid (t-1)), représentent respectivement, les trois composantes citées au-dessus.

Algorithme 2.1. L'optimisation par essaim de particules[27].

Début

- Initialiser les paramètres et la taille S de l'essaim;
- Initialiser les vitesses et les positions aléatoires des particules dans chaque dimension de l'espace de recherche;
- Pour chaque particule, $p_{bestid} = x_{id}$;
- Calculer f (x_{id}) de chaque particule;
- Calculer g_{bestd}; // la meilleure p_{bestid}
- **Tant que** (la condition d'arrêt n'est pas vérifiée) **faire**
- **Pour** (i allant de 1 à S) **faire**
- Calculer la nouvelle vitesse à l'aide de l'équation (3.1) ;
- Trouver la nouvelle position à l'aide de l'équation (3.2) ;
- Calculer f (x_{id}) de chaque particule;
- **Si** (f (x_{id}) est meilleur que f (p_{bestid})) **alors**
- pbestid = xid;
- **Si** (f (p_{bestid}) est meilleur que f (g_{bestd})) **alors**
- $g_{bestd} = p_{bestid}$;
- **Fin pour**
- **Fin tant que**
- Afficher la meilleure solution trouvée g_{bestd} ;

Fin

6.3.1 Les variantes de l'algorithme PSO

L'idée des pionniers de l'optimisation par essaim de particules : Kennedy et Eberhart a sollicité l'attention de plusieurs chercheurs qui ont mené des études dans l'objectif d'améliorer la performance de la méthode proposée. En fait, malgré l'efficacité et la simplicité de la mise en œuvre de l'algorithme d'optimisation par essaim de particules, ce dernier souffre du problème de la convergence prématurée**[28].**

En 1996 Eberhart et ses collègues ont proposé de limiter la vitesse de la particule par l'intervalle [- v_{max}, v_{max}]**[29].**Leur objectif était d'échapper au problème de déviation des particules de l'espace de recherche lors de leur déplacement. Le nouveau paramètre v_{max}

permet de mieux contrôler le mouvement de particules. D'autres études sur l'introduction de v_{max} sont disponibles dans **[30],[31],[32],[33],[34].**

En 1998, Shi et Eberhart **[35]** ont proposé dans une variante de l'équation (2.1).La modification apportée consiste à appliquer un facteur d'inertie pour contrôler la vitesse des particules de la manière suivante:

(2.3) Où ω est un coefficient d'inertie, généralement une constante qui sert à contrôler l'influence de la vitesse de la particule sur son prochain déplacement afin de garder un équilibre entre l'exploitation et l'exploration de l'espace de recherche. Elle peut être variable dans certains cas Eberhart et Shi**[30]**voient que la valeur raisonnable de ω doit diminuer linéairement au cours du processus de l'optimisation. De même Kusum Deep et Jagdish Chand Bansal**[33]**proposent de réduire la valeur de ω linéairement de 0.8 à 0.4 et d'autres études disponibles dans **:[36],[37],[38].**

De sa part, Clerc a proposé**[39],[27]** une autre variante de l'équation (2.1). Sa variante consiste à ajouter un facteur de constriction K dont l'objectif est de contrôler la vitesse des particules afin d'échapper au problème de la divergence de l'essaim qui cause la convergence prématurée de l'algorithme. L'équation permettant la mise à jour de la vitesse est la suivante :

$$v_{id}(t) = K[v_{id}(t-1) + c_1 r_1(P_{bestd}(t-1) - x_{id}(t-1))] \quad (2.4)$$

Où

$$k = \frac{2}{|2-\varphi-\sqrt{\varphi(\varphi-4)}|} \quad (2.5)$$

Avec = c_1+ c_2 et ; $c_1 = c_2 = 2.05$; ce qui donne : K=0.729844. 4 **[22].**

7 METHODES DE DETECTION DES ILOTS CPG

Les méthodes *in silico* pour la détection des ilots CpG sont appliqués sur les génomes humains ou mammifère pour une détection précise de CpG, elles sont principalement classées en quatre classes en fonction de leurs algorithmes principaux en tant que :

Méthodes basées sur des fenêtres coulissantes : Hudson et Kaplan en 1988 ont introduit la technique de fenêtre de coulissantes (SWM) elle consiste à examiner le génome en faisant défiler les fenêtres et en identifiait les CpG par normes statistique, le but est de déplacer la fenêtre réglable pour chaque nucléotide pour calculer le contenu GC et le CpG O/E. CpG Plot, CpG Report, CpG Prod, CpG IS, CpG IE sont des exemples des méthodes de cette classe. Les méthodes de cette catégorie sont largement appliquées en raison de leur efficacité statistique

des normes, mais une limitation majeure de ces algorithmes sont la taille de fenêtre limitée, qui décide le succès de la prédiction. La petite taille de la fenêtre augmente les chances d'échecs pour identifier le potentiel CGI et réduit le calcul de complexité, tandis que la plus grande taille de fenêtre réduit le calcul de la vitesse et augmente la granularité prédictive. La faible sensibilité est également considérée comme un inconvénient de ces méthodes pour la prédiction des ilots CGI entières **[40].**

Méthode basée sur la densité : détermine instinctivement la densité des sites CpG, comme les méthodes basées sur les fenêtres, qui utilisent les statistiques normes. Le pourcentage de sites CpG dans CGI et la totalité de la longueur de CGI est évalué pour calculer la densité de CGI. Le principe de cette méthode est de définir les graines initiales pour la régulation de manières répétitive les variables de densité et ainsi augmenter la couverture des régions riches en CpG**[41],[42],**CpG IF**[43]** est un exemple des méthodes de cette catégorie.

Méthode basée sur la distance Word cluster, CpG-MI /longueur : CpG Cluster

Elle fournit une approche rapide pour la prédiction des CGI qui assemble les données dans le contexte de la distance entre CpG sites. Cette méthode examine la propriété de séquence entre deux sites CpG en ligne, ce qui amène également des critiques à ce sujet technique. Le même CGI dans de différentes situations entraine de divers résultats, la faible sensibilité prédictive avec les résultats dus à la composition de la séquence sont également pris en compte comme l'inconvénient de cette technique. Toutes les méthodes existantes étaient basées sur le vaste paramètre fait par le contenu GC, la fraction CpG et la longueur seuil. La distribution de la distance diffère dans les CGI et ADN entre les CpG adjacents en raison du nombre élevé des nucléotides CpG aux CG**[44].**Hachenberg et ses collègues**[45]**ont développé une nouvelle approche nommée CpG Cluster capable de déterminer directement les clusters CpG en fonction des distance physiques. Les grappes statistiquement significatives sont déclarées comme CGI après l'attribution de la valeur p à chaque groupe.

Le modèle de Markov Caché (HMM) :cette méthode permet l'approche extensible pour la détection CGI, la détermination du statif des CGI et le calcule de probabilité transitive entre et dans le CGI en appliquant le modèle de transition statistique**[46],[47],[48],** cette dernière est calculée entre deux nucléotides pendant l'état d'apprentissage pour CGI et non les régions CGI, sa valeur est beaucoup plus faible que la région riche en CpG. Par conséquent les variations entre les régions CpG et non CpG régions sont déterminées par le rapport de vraisemblance du log de probabilité pour chaque séquence possible**[46].**L'efficacité informatique de la méthode HMM n'est pas productif, mais son application a été mise en

œuvre pour la séquence d'analyse, et plus tard avec succès pour la répartition des génomes**[49].**

8 CONCLUSION

Dans ce chapitre nous avons présenté un état de l'art sur les méthodes de résolution de différents problèmes d'optimisation présentés dans la littérature commençant par les méthodes exactes aux méthodes approchées. Nous avons essayé de présenter le principe des méthodes métaheuristiques y compris les méthodes basées sur l'intelligence par essaim qui ont construit une tendance très active ces dernières décennies.

Dans le chapitre suivant nous présentons notre contribution pour la détection des ilots CpG qui consiste à la proposition d'une méthode hybride entre les métaheuristiques PSO et les méthodes CpG Cluster.

Chapitre 3 :

Etude Expérimentale

1 INTRODUCTION

Lesméthodesde détection des îlots CpGactuelles sont principalement basées sur lescritères proposés par Gardiner-Garden et Former (GGF)**[50],**y compris le contenu GC, le rapport O/E et le seuil de longueur de l'ilot. Gardiner-Garden et Former ont défini l'ilot CpG comme une séquence ADN dont : la plage de longueur de l'ilot≥ 200 pb (équation 3.1), le contenu GC ≥ 50 % (équation 3.2), ratio O/E ≥ 0.6 (équation 3.3) et l'écart entre les îlots adjacents est défini à 100 Pb.

$$CpG_{len}(P_i) = \frac{\#A+ \#T+ \#C+ \#G}{P_{max} - P_{min}} \quad (3.1)$$

$$GC(P_i) = \frac{\#C+ \#G}{\#A+ \#T+ \#C+ \#G} \quad (3.2)$$

$$Obs_{CpG}/Exp_{CpG}(P_i) = \frac{\frac{\#CpG}{CpG_{length}}}{\frac{\#C}{CpG_{length}} \times \frac{\#G}{CpG_{length}}} \quad (3.3)$$

Où #A, #T, #C et #G sont respectivement le nombre de nucléotides Adénine (A), Thymine (T), Cytosine (C) et Guanine (G) dans la région prédite de l'îlot CpG à Pi. Pmin est la position debutdu cluster moins 200, et Pmax est la position de départ du cluster plus200. #CpG représente le nombre de CpG dans la région prévue de l'îlot CpG à Pi.

Dans cette étude, nous proposons une approche combinant deux méthodes : CpG Cluster et les méthodes PSO pour une prédiction efficace des ilots CpG. Afin de comparer nos résultats avec celles d'autres méthodes, nous nous sommes basés sur les critères GGF pour la prédiction des ilots CpG et nous avons mesuré la performance de notre méthode en utilisant cinq mesures y compris :

$$ACC = \frac{TP+ TN}{TP+ FP+ TN+ FN} \quad (3.4)$$

$$SN = \frac{TP}{TP+ FN} \quad (3.5)$$

$$SP = \frac{TN}{TN+FP} \quad (3.6)$$

$$PC = \frac{TP}{TP+FP+FN} \quad (3.7)$$

$$CC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FN) \times (TP+FP) \times (TN+FN) \times (TN+FP)}} \qquad (3.8)$$

Où TP est un vrai positif indique que le nombre de base détectées dans la région de l'îlot CpG sont de vrai îlots CpG, FN est un faux négatif indique que le nombre de bases détectées dans la région de l'îlot CpG est incorrect, TN est un vrai négatif indique que le nombre de bases détectées dans la région de l'îlot non-CpG est correct et FP est un faux positif indique que certaines des bases détectées dans la région de l'îlot CpG sont des îlots non CpG. Nous avons prédit les îlots CpG selon les critères du GGF. Par la suite, nous avons utilisé cinq critères d'évaluation pour évaluer la performance de toutes les méthodes de prédiction des îlots CpG.

2 METHODE PROPOSEE

Notre méthode (Cluster KPSO) est composée de deux phases :

- La première consiste à utiliser la méthode de clustring (CpG Cluster), pour la détection des îlots CpG. CpGCluster est introduit en tant que stratégie de prétraitement pour la détection de tous les ilots CpG candidats.
- La deuxième étape consiste à utiliser l'algorithme PSO, pour prédire les ilots CpG à partir de tous les candidats obtenus à partir de la première étape.

2.1 Identification des Ilots CpG en utilisant CpGCluster

CpG Cluster a été proposé pour la première fois par Hachenberg et al **[11].** La théorie de base de CpG Cluster est basée sur la distance physique entre les CpGs voisins afin de détecter directement les clusters de CpG sans tenir compte des critères subjectifs de CpG. La procédure CpG Cluster comporte deux parties :

1- Un algorithme basé sur la distance des clusters CpG recherchés dans le génome.

2- le critère de la valeur de sélectionne les clusters CpG statistiquement significatifs.

Les étapes de CpG Cluster sont décrites ci-dessous :

Etape 1 : Toutes les positions CpG sont scannées de 5' à 3' dans une séquence d'ADN, et les positions CpG sont collectées dans un ensemble C.

Etape 2 : les distances de tous les CpG adjacents sont calculées, dans lesquelles une distance physique entre deux CpG adjacents est calculée par $d_i = x_{i+1} - x_i - 1$. La plus courte distance des CpG adjacents est 1, c'est-à-dire CGCG.

Etape 3 : Une valeur seuil (dt) est définie en fonction de la distribution des distances de tous les CpG dans une séquence d'ADN, et utilisée pour déterminer si les CpG adjacents appartiennent à un même Cluster ou pas.

Etape4 : CpG Cluster utilise une valeur de seuil pour collecter les positions des ilots CpG et générer les clusters. Lorsque la distance adjacente entre CpG est inférieure à la valeur seuil, les deux CpG adjacents sont classés au même cluster, sinon, un nouveau cluster sera créé. Ainsi l'étape 4 continue de chercher de nouveaux clusters jusqu'à ce qu'elle rencontre le dernier CpG.

Etape 5 : Après la détermination de tous les clusters, le p-value de chaque cluster est calculée pour estimer la probabilité de découvrir un cluster CpG dans une séquence aléatoire. La distribution binomiale négative est calculée par la fonction de densité cumulée au point n_f du cluster CpG et est considérée comme p-value :

$$\mathrm{p}_{\mathrm{Np}}^{\mathrm{aim}}(x <= n_f) = \sum_{x=0}^{n_f} \binom{\mathrm{x}-(\mathrm{N}+1)-1}{(\mathrm{N}-1)-1} \times \mathrm{p}^{\mathrm{N}-1} \times (1-\mathrm{p})^{\mathrm{x}} \quad (3.9)$$

$$n_f = L - 2 \times N \quad (3.10)$$

$$\mathrm{p} = \frac{N_s}{N_{is}} \quad (3.11)$$

3 Raffinement des ilots CpG avec la métaheuristique PSO

La métaheuristique PSO (Particule Swarm Optimization) a été proposée en 1995 par Kennedy et Eberhart. En 2002, CLERC ET KENNEDY ont proposé**[39][51]** une nouvelle varianteau PSO qui consiste à ajouter un facteur de constriction K dont l'objectif est de contrôler la vitesse des particules afin d'échapper au problème de la divergence de l'essaim qui cause la convergence prématurée de l'algorithme. PSO est méthode d'optimisation qui a montré son efficacité et sa performance dans la résolution de différents problèmes. Dans cette phase de la méthode Cluster KPSO, nous utilisons la méthode PSO avec sa variante proposée par CLERC ET KENNEDY pour optimiser les candidats d'ilots CpG obtenus en utilisant la méthode CpG Cluster. Les étapes de cette phase sont les :

Etape 1 : Toutes les positions CpG sont scannées de 5' à 3' dans une séquence d'ADN, et les positions CpG sont collectés dans un ensemble C=$\{c1, c2, \ldots\ldots, cn\}$.

Etape 2 : les distances de tous les CpG adjacents sont calculées, dans laquelle une distance physique entre deux CpG adjacents est calculé par di = Xi -1. La plus courte distance de CpG adjacent est 1, c'est-à-dire CGCG.

Etape 3 : Une valeur seuil (dt) est définie en fonction de la distribution des distances de tous les CpG dans une séquence d'ADN, et utilisé pour déterminer si les CpG adjacents appartiennent à un même Cluster ou pas. L'ensemble C est trié en fonction de la distance entre les CpG adjacents, et le seuil dt est défini au centile de l'ensemble C.

Mettre à jour le pb est et gb est pour toutes les particules. Chaque particule trouve sa meilleure position.

$$v_{ij}^{new} = k \times \left[v_{ij}^{old} + c_1 \times r_1 \times \left(pbest_{ij} - x_{ij}^{old}\right) + c_2 \times r_2 \times (gbest - x_{ij}^{old})\right] \quad (3.12)$$

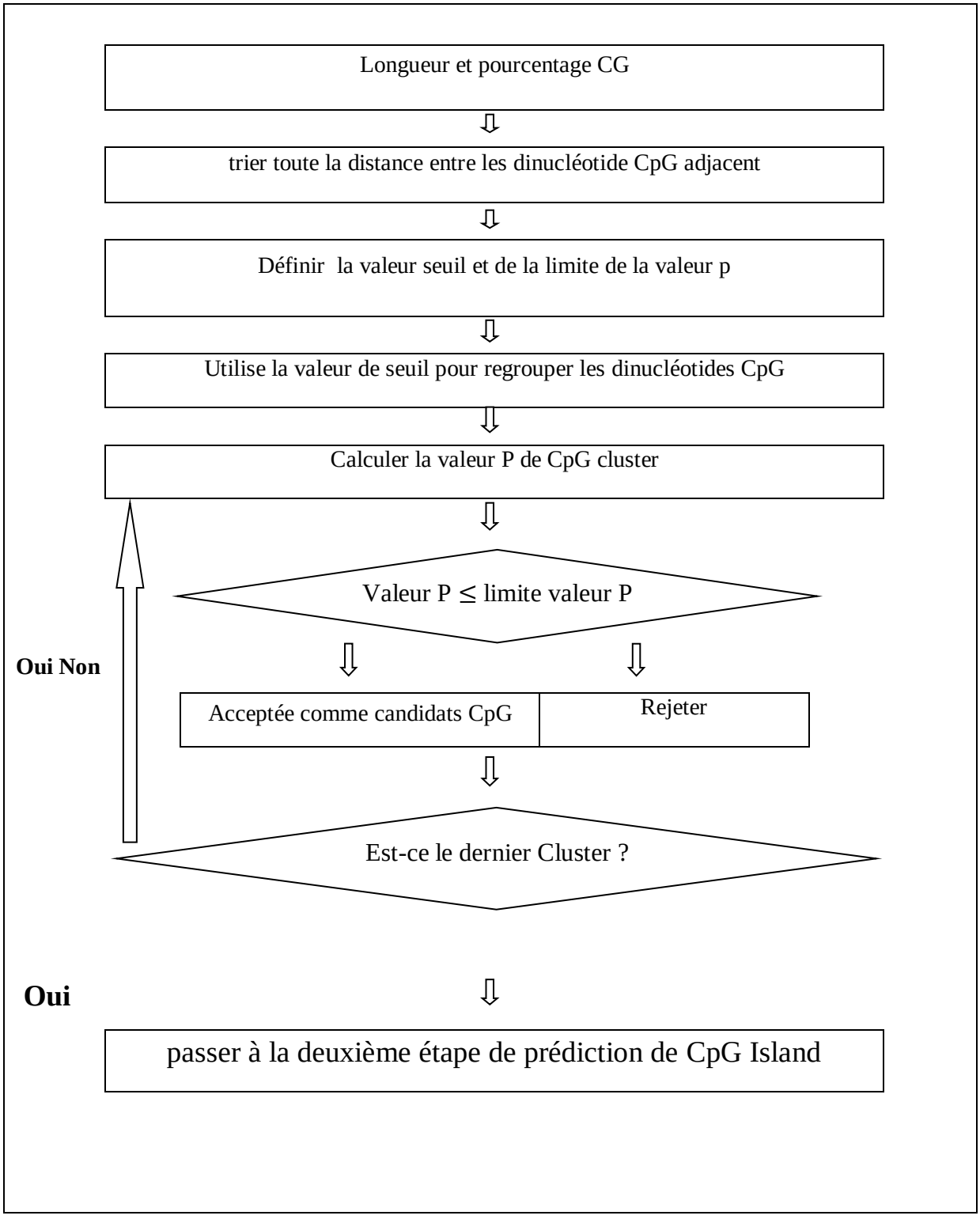
Etape4 : CpG Cluster utilise une valeur de seuil pour commencer à étendre en aval (! 3') à partir du première CpG de l'ensemble C. Lorsque la distance adjacente entre CpG est inférieure au seuil valeur, les deux CpG adjacents sont classées en un seul groupe, autrement, la position de la grappe de fermeture. Ainsi l'étape 4 continue de chercher de nouveaux regroupements jusqu'à ce qu'elle rencontre le dernier CpG.

Étape 5 : Confirmez si le critère d'arrêt (génération maximale = 100) est respecté ou non. Les étapes 2 à 5 sont répétées jusqu'à ce que la génération maximale soit atteinte.

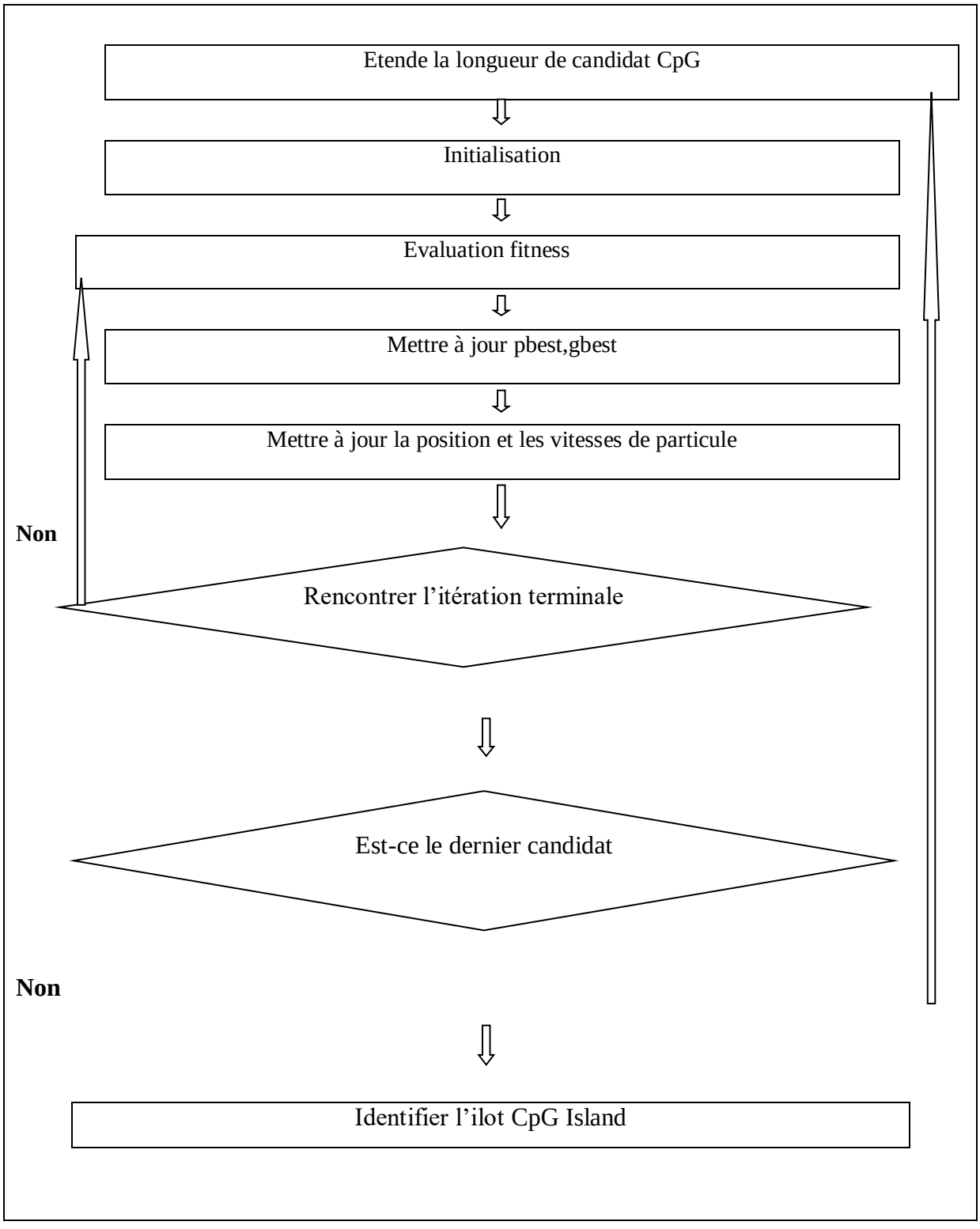
5 sont répétées jusqu'à ce que la génération maximale soit atteinte.

Lorsque tous les candidats d'îlots CpG sont prédits par PSO, tous les îlots CpG identifiés représentent les résultats de la détection des îlots CpG.

La figure 9 montre le diagramme de ClusterKPSO divisé en deux étapes.



Etape 1 : Détection de cluster CpG candidates dans l'ADN séquence



Etape 2 : prédiction des CpG Island dans les CpG candidats

Figure 9 : Les étapes de la méthode proposée (Cluster KPSO).

4 MESURES DE PERFORMANCE

Nous avons utilisé cinq critères communs pour déterminer la prédiction de la précision des méthodes comparées, à savoir la sensibilité (SN), la spécificité (SP), la précision (ACC), le coefficient de performance (PC) et le coefficient de corrélation (CC). Les cinq critères sont définis dans les Eqs. (3.3 à 3.8) Ces cinq critères d'évaluation permettent de déterminer la supériorité d'un algorithme.

4.1 Paramètres utilisés

Dans ce tableau (1), nous retrouvons les différents paramètres utilisés**.**

Tableau 1 : Paramètres des méthodes utilisées.

Paramètres CpG Cluster	Paramètres PSO	Paramètres GGF
➢ Distance Seuil : 65 ➢ P-value : 0.01	➢ Taille de population : 300 ➢ Itération :100 ➢ c_1, c_2 : 2	➢ Longueur : 200pb. ➢ Contenu GC : 0,5 ➢ O/E : 0,6 ➢ Ecart entre les îlots adjacents : 100 Pb.

4.2 Données utilisées

du chromosome 21 et 22 qu'on a téléchargé depuis la banque de données NCBI (https://www.ncbi.nlm.nih.gov/), on les retrouve en dessous :

Contigs1:ref|NT_028395.3|:1-647850 Homo sapiens chromosome 22 genomic contigs, GRCh37 reference primary assembly.

Contigs2:ref|NT_113952.1|:1-184355 Homo sapiens chromosome 21 genomic contigs, GRCh37 reference primary assembly.

Contigs3:ref|NT_113953.1|:1-131056 Homo sapiens chromosome 21 genomic contigs, GRCh37referenceprimaryassembly.

Contigs 4:ref|NT_113954.1|:1-129889 Homo sapiens chromosome 21 genomic contig, GRCh37 reference primary assembly.

Contigs5:ref|NT_113955.2|:1-281920 Homo sapiens chromosome 21 genomic contig,

GRCh37 reference primary assembly.

5 RESULTATS

Nous avons comparé notre méthode avec cinq autres méthodes tirées de la littérature : CpG Plot**[52] ,** CpG Cluster **[53],**CpG ProD**[54],**CpG IS**[55]** et PSO **[56].**

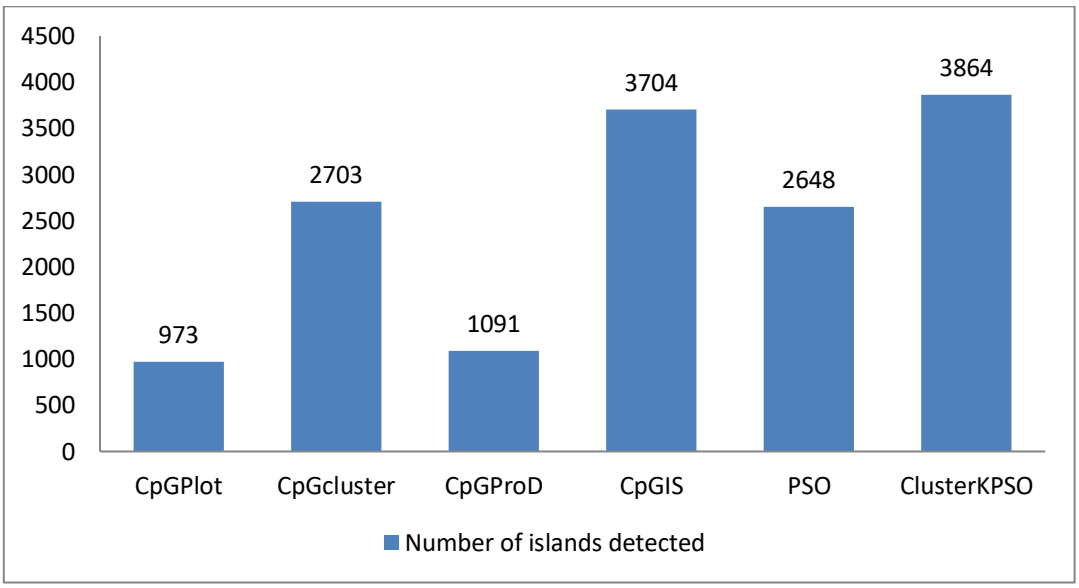


Figure 10 : Nombre des ilots CpG détectés (Chromosome 21).

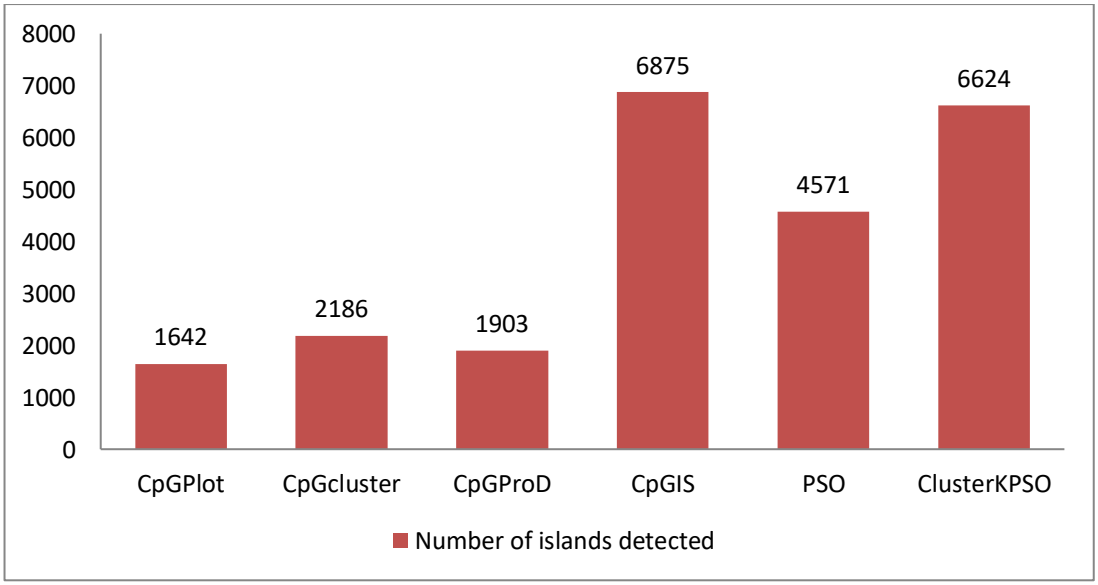


Figure 11 : Nombre des ilots CpG détectés (Chromosome 22).

Les figures 10 et 11 représentent le nombre de détection des ilots du chromosome 21 et 22, en effectuant le test avec notre méthode proposée Cluster KPSO et les autres méthodes CpG Plot, CpG Cluster, CpG ProD, CpG IS, et PSO, on remarque que le meilleur résultat a été trouvé par la nôtre dans le premier chromosome 21, qui est de 3864 suivi de près de la méthode CpG IS, après viennent les autres méthodes. Par contre dans le deuxième chromosome 22, on remarque que le meilleur résultat a été trouvé par la méthode CpG IS dont sa valeur est de 6875 suivi de près de notre méthode Cluster KPSO, qui sa valeur est de 6624, la différence n'est pas importante et est très petite, après viennent les autres méthodes.

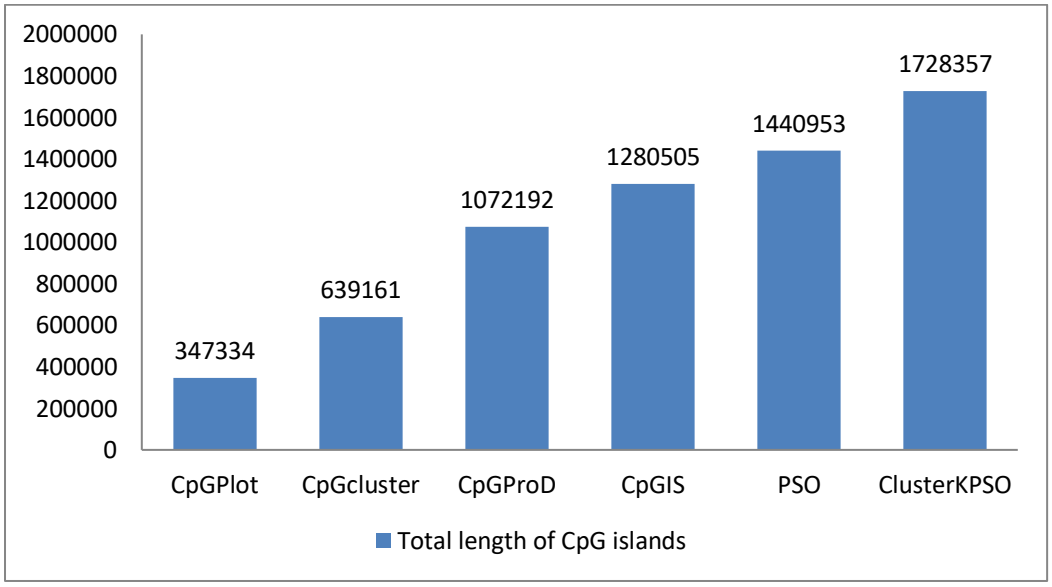


Figure 12 : La longueur totale des îlots CpG (Chromosome21).

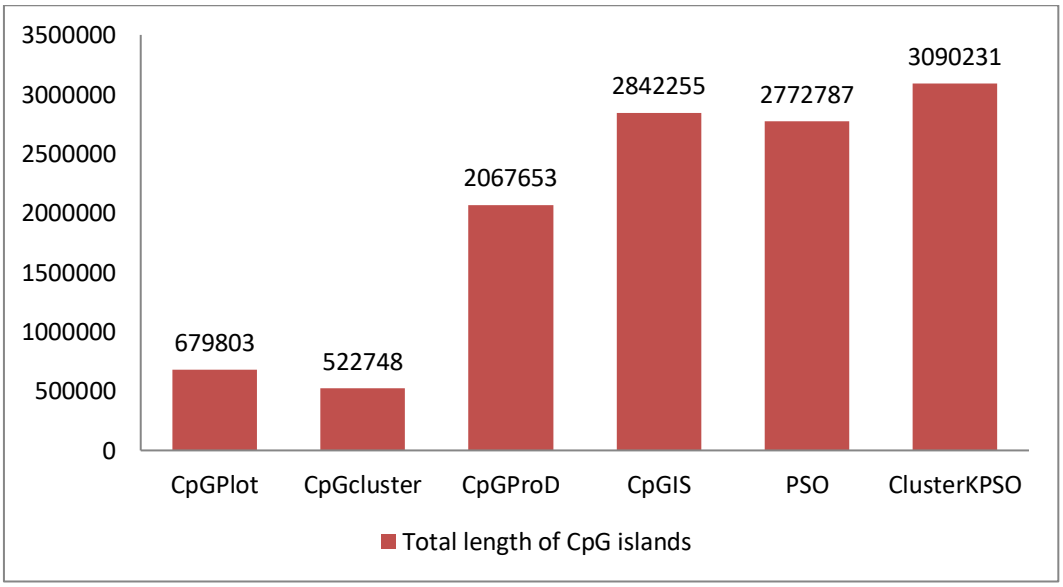


Figure 13: La longueur totale des îlots CpG (Chromosome22).

Les deux figures 12 et 13représentent la longueur totale de CpG islands du chromosome 21 et 22, en comparaison de notre méthode proposée Cluster KPSO avec les autres méthodes CpG Plot, CpG Cluster, CpG ProD, CpG IS, et PSO. On remarque que la meilleure valeur a été trouvé par la nôtre qui est Cluster KPSO, qui est de 1728357 dans le chromosome 21, et 3090231 dans le chromosome 22, suivi par les autres méthodes.

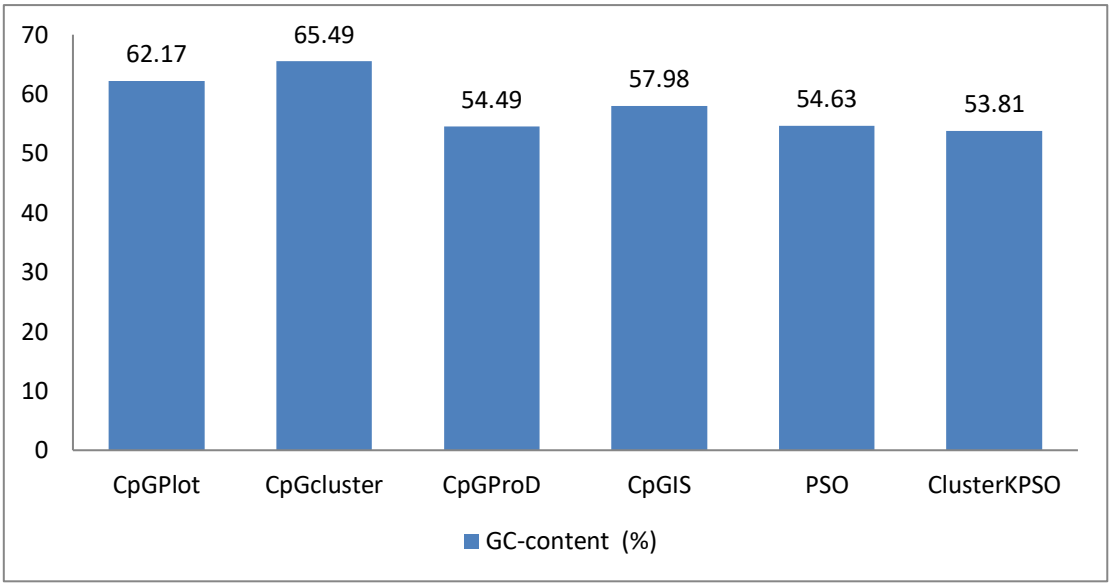


Figure 6 : Teneur en GC(%) (Chromosome21).

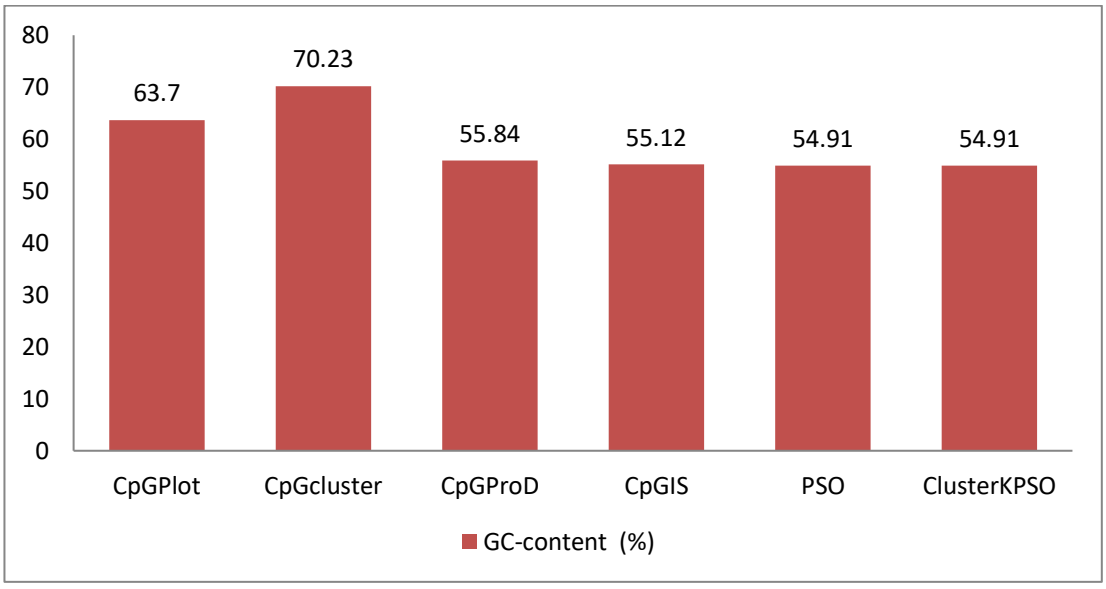


Figure 15 : Teneur en GC (%) (Chromosome 22).

Les figure 14 et 15 représentent la teneur en GC(%) des ilots du chromosome 21 et 22, en effectuant avec plusieurs méthodes, parmis Cluster KPSO qui a été proposé par nous, et les autres méthodes CpG Plot, CpG Cluster, CpG ProD, CpG IS, et PSO , on remarque que le

meilleur résultat a été retrouvé par CpG Cluster dans les deux chromosomes 21 et 22 qui leurs valeurs sont de 65,49 et 70,23 respectivement, suivi de près de CpG Plot, suivi de Cp GIS, PSO, CpG ProD, et en dernière position vient Cluster KPSO, on remarque que le plus faible résultat a été trouvé par notre méthode proposée.

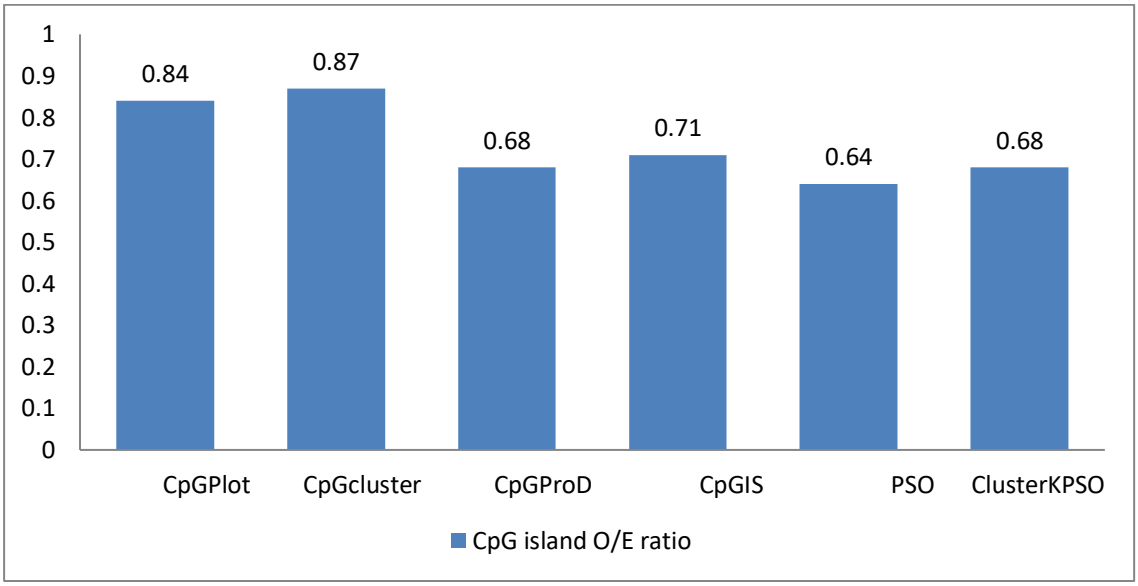


Figure16 : Rapport O/E de l'ilot CpG (Chromosome 21).

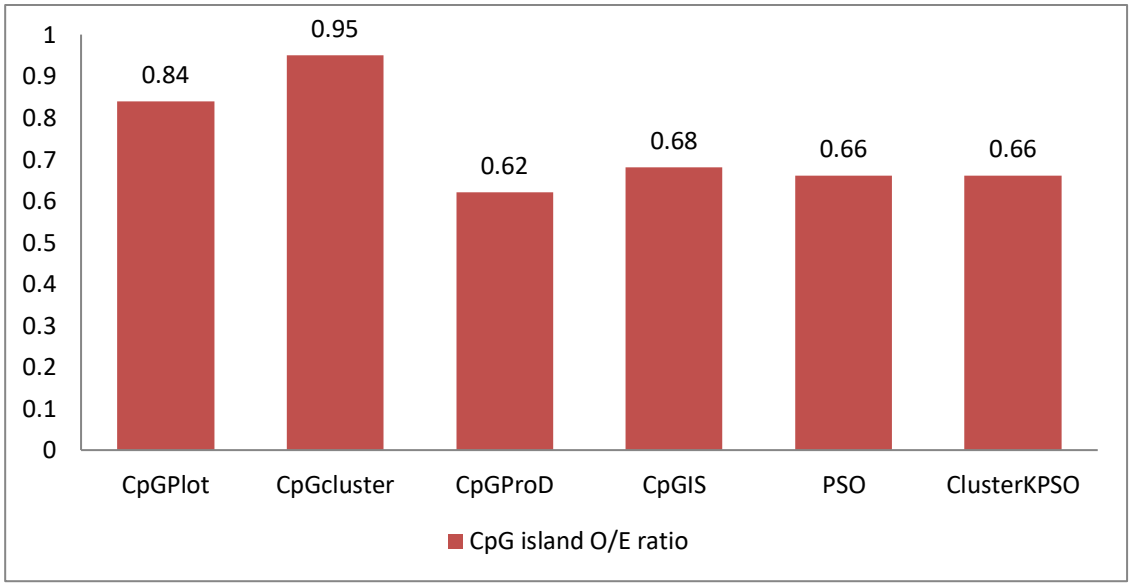


Figure17 : Rapport O/E de l'ilot CpG (Chromosome 22).

Les figures16 et17 représentent le rapport O/E d'îlot CpG du chromosome 21. En comparaison de notre méthode Cluster KPSO avec les autres méthodes CpG Plot, CpG

Cluster, CpG ProD, CpG IS et PSO, on remarque que la meilleure valeur a été trouvée par la nôtre, qui est de 1728357 suivis par les autres méthodes. Le rapport O\E d'ilots CpG du chromosome 21 et 22, on remarque que CpG Cluster a trouvé la meilleure valeur dans les deux chromosomes qui est de 0,95 dans le chromosome 21, et la valeur de 0,87 dans le chromosome 22. Et CpG Plot 0,84, CpG IS 0,68, PSO et Cluster KPSO égaux à 0,66, et en dernier CpG proD dans le chromosome 21. Et CpG Plot 0,84, PSO 0,71, Cluster KPSO et CpG IS égaux à 0,68, et en dernier CpG proD dans le chromosome 22.

Tableau 2 : Comparaison de différentes méthodes de détection des ilots CpG.

Contig	Mesure de performance	CpG plot	CpG Cluster	CpG Prod	CpG IS	PSO	ClusterKPSO
NT_113952.1	SN	56.43	50.46	58.07	83.98	69.22	95.98
	SP	100.0	99.95	99.50	99.05	99.61	99.47
	ACC	98.09	97.78	97.69	98.39	98.28	99.32
	PC	56.42	49.92	52.36	69.59	63.77	86.16
	CC	74.38	69.41	68.83	81.25	77.66	92.28
NT_113955.2	SN	47.19	67.15	68.51	85.12	54.47	94.67
	SP	100.0	99.72	99.63	99.30	99.96	99.51
	ACC	98.08	98.54	98.50	98.79	98.31	99.33
	PC	47.14	62.47	62.35	71.78	53.87	83.81
	CC	67.94	77.03	76.65	82.96	72.41	90.92
NT_113958.2	SN	51.29	27.16	46.41	82.13	79.27	88.56
	SP	99.99	99.94	98.93	98.26	98.13	99.10
	ACC	96.90	95.32	95.60	97.24	96.93	98.43
	PC	51.24	26.92	40.10	65.36	62.10	78.20
	CC	70.38	49.96	56.80	77.63	75.03	86.93
NT_113953.1	SN	22.80	57.32	29.79	74.05	60.20	82.74
	SP	100.0	99.74	99.56	98.83	99.27	99.47
	ACC	97.76	98.51	97.53	98.11	98.13	98.99
	PC	22.80	52.74	25.96	53.23	48.39	70.39
	CC	47.21	69.89	43.61	68.64	64.50	82.09
NT_113954.1	SN	31.24	29.86	52.01	76.31	56.92	78.02
	SP	100.0	99.46	98.72	97.62	98.40	98.23
	ACC	97.47	96.90	97.00	96.83	96.87	97.48
	PC	31.24	26.19	38.94	47.05	40.12	53.34
	CC	55.17	43.81	54.68	63.29	55.65	68.72
NT_028395.3	SN	27.11	44.89	54.18	76.68	68.97	81.52
	SP	100.0	99.47	99.45	98.93	99.27	99.24
	ACC	97.98	97.53	98.19	98.14	98.19	98.60
	PC	27.10	39.26	45.36	59.36	57.49	67.53
	CC	51.51	57.21	62.26	73.57	72.21	79.90

Pour le premier contig, terme de SN, notre algorithme qui est Cluster PSO est plus performant que le reste des algorithmes simple sa valeur de 95.98, et est la meilleure par

rapport aux autres méthodes, en terme de SP, l'algorithme CpG plot a trouvé le meilleur résultat qui est de 100, suivi de près de Cluster PSO qui sa valeur est de 99, 47, suivi des autres méthodes. En terme de ACC, cluster PSO trouvé le meilleur résultat qui est de 99.32 en comparant aux autres algorithmes, en terme de PC et CC. Cluster PSO est toujours le meilleur en trouve la meilleure solution qui valeur sont de 86.16 et 92.28 respectivement.

En la deuxième contigüe, on analyse que Cluster PSO a trouvé la meilleure valeur en terme de SN, ACC, PC et CC on qui est 94.67, 99.33, 83.81, 90,92 sauf pour SP l'algorithme CpG plot a trouvé la meilleure valeur qui est de. 100 et la Cluster KPSO 99.51 donc en remarque que n'y a pas une grande différence entre eux.

Pour le troisième, quatrième, cinquième et sixième contig, on remarque la même close par rapport aux deux premiers contigs ce qui veut dire, qu'en terme de SN, ACC, PC, CC, Cluster KPSO a trouvé le meilleur résultat par rapport aux autres algorithmes, tant dit que pour SP la méthode CpG plot a trouvé la meilleure valeur suivie de tout près de Cluster KPSO.

6 CONCLUSION

D'après les résultats de la table et des histogrammes, on conclue que la précision de prédiction de Cluster KPSO était assez élevée par rapport aux autres méthodes, ce qui indique que Cluster KPSO a plusieurs avantages sur les autres algorithmes, mais cela. N'empêche que la détermination précise et rapide de CpG îles pour les séquences d'ADN entier reste expérimentalement et computationnellement difficile.

Conclusion Générale et Perspectives

CONCLUSION GENERALE ET PERSPECTIVES

Dans ce mémoire, nous avons abordé le problème de prédiction es îlots CpG dans le génome humain. Les îlots CpG sont des régions de densité accrue de séquence dinucléotidiques Cytosine-phosphate-Guanine. Elles forment des étendues de plusieurs centaines à plusieurs milliers de paires de bases. Ils sont associés aux régions transcrites, particulièrement aux promoteurs géniques. Sachant qu'une partie de ces îlots CpG sera methylés, les promoteurs qui contiennent un niveau intermédiaire de dinucléotides CpG étaient significativement plus méthylés que ceux riches en dinucléotides CpG. Ces derniers sont méthylés différentiellement dans huit tissus somatiques, Mais cela n'est pas forcément lié à un patron d'expression particulier des gènes correspondants. Les recherches scientifiques ont montré que l'hyperméthylation des îlots CpG au niveau des promoteurs de gènes suppresseur du cancer entraine une inhibition de leur transcription. Cela créé des cellules cancéreuses.

Les cancers sont aujourd'hui des maladies autant génétiques qu'épigénétiques. En altérant l'expression de gènes impliqués dans la régulation cellulaire, les modifications épigénétiques jouent un rôle fondamental dans l'initiation et la progression des tumeurs ; contrairement aux mutations génétiques, elles sont potentiellement réversibles. Des inhibiteurs épigénétiques sont ainsi évalués comme agents anti tumoraux. Par ailleurs, l'étude de la méthylation de l'ADN se profile comme un marqueur biologique pouvant contribuer à la classification tumorale, au diagnostic et au pronostic en pratique clinique. Le traitement du cancer par chimiothérapie repose sur la destruction, en général par apoptose, des cellules tumorales. La thérapie épigénétique fonctionne en modifiant le profil d'expression génique au sein des cellules tumorales.

A la fin, on propose comme perspective :

-L'utilisation d'autres hybridations avec d'autres méthodes d'îlots CpG à part la méthode CpG Cluster.

- l'application d'autres métaheuristiques comme : l'algorithme génétique, algorithme de la recherche basée sur les coucou, les colonies d'abeilles, les colonies de fourmis … etc.

Référence

[1] « Génomique — Wikipédia ». https://fr.wikipedia.org/wiki/G%C3%A9nomique (consulté le juin 10, 2021).

[2] « Génome — Wikipédia ». https://fr.wikipedia.org/wiki/G%C3%A9nome (consulté le juin 10, 2021).

[3] Claire Guillet-Renard, « Evolution des Ilots CpG chez les Primates ». L'Université Claude Bernard - Lyon 1, oct. 07, 2009.

[4] Bird A, « DNA methylation patterns and epigenetic memory ». Genes Dev.

[5] Ponger L, Duret L, Mouchiroud D, « Determinants of CpG islands: expression in early embryo and isochore structure. Genome Res ». 2001.

[6] Ponger L, Duret L, Mouchiroud D, « Determinants of CpG islands: expression in early embryo and isochore structure. Genome Res. » 2001.

[7] Antequera F, Bird A, . « Number of CpG islands and genes in human and mouse ». Proc Natl Acad Sci USA, 1993.

[8] Cadoret J-C, Meisch F, Hassan-Zadeh V, Luyten I, Guillet C, et Duret L, Quesneville H, Prioleau M-N, « Genome-wide studies highlight indirect links between human replication origins and gene regulation ». Proc Natl Acad Sci USA, 2008.

[9] Cadoret J-C, Meisch F, Hassan-Zadeh V, Luyten I, Guillet C, Duret L, Quesneville H, Prioleau M-N, Cadoret J-C, Meisch F, Hassan-Zadeh V, Luyten I, Guillet C, et Duret L, « Genome-wide studies highlight indirect links between human replication origins and gene regulation ». Proc Natl Acad Sci USA, 2008.

[10] « These_GuilletRenard_vfinale.pdf ».

[11] Han Q , Lu J , Su D, Duan J, Hou X, Li F , Wang X , Huang B, «) L'acétylation des histones H3 induite par Gcn5 et Elp3 régule la transcription du gène hsp70 chez la levure », 2008.

[12] Takai D, Jones PA, *) Comprehensive analysis of CpG islands in human chromosomes 21 and 22*. Proceedings of the National Academy of Sciences of the United States of America, 2002.

[13] Guenther MG, Levine SS, Boyer LA, Jaenisch R, Young RA, « A chromatin landmark and transcription initiation at most promoters in human cells ». Cell, 2007.

[14] Antequera F, Bird A., « Number of CpG islands and genes in human and mouse ». Proc Natl Acad Sci USA., 1993.

[15] F.Fuks, « DNA methyltransferases : from chromatine remondling to cancer ». Med Sci, 2003.

[16] « L'hyperméthylation des gènes suppresseurs de tumeur comme marqueur en cancérologie- EM consulte ». https://www.em-consulte.com/article/201591/lhypermethylation-des-genes-suppresseurs-de-tumeur (consulté le juin 12, 2021).

[17] « Optimization problem », *Wikipedia*. avr. 06, 2021. Consulté le: juin 13, 2021. [En ligne]. Disponible sur: https://en.wikipedia.org/w/index.php?title=Optimization_problem&oldid=1016282288

[18] F. Glover and M. Laguna., « Tabu Search ». Kluwer Academic Publishers, Boston, 1997.

[19] B. Mille, « Méthodes approchées pour la résolution d'un problème d'ordonnancement avec travaux interférants », p. 54.

[20] S. M. Johnson., « Optimal two-and three-stage production schedules with setup time included ». Naval Research Logistic Quarterly, 1954.

[21] A. Gherboudj, « Methode-optimisation (1) », Abd El Hamid Mehdi.

[22] A. Gherboudj, « Méthodes de résolution de problèmes difficiles académiques », Abd El Hamid Mehdi, 2013.

[23] « Métaheuristique », *Wikipédia*. mars 16, 2021. Consulté le: juin 13, 2021. [En ligne]. Disponible sur: https://fr.wikipedia.org/w/index.php?title=Métaheuristique&oldid=180927297

[24] J. Kennedy, R. C. Eberhart, « Particle swarm optimization ». Proceedings of the IEEE International Conference Neural Networks, 1995.

[25] A. Ratnaweera, SK. Halgamuge, « Watson. Self-organising hierarchical particle swarm optimizer with time-varying acceleration coefficients ». IEEE Trans Evol Comput, 2004.

[26] I-H. Kuo, S-J. Horng, T-W. Kao, T-L. Lin, P. Fan, « An Efficient Flow-Shop Scheduling Algorithm Based on a Hybrid Particle Swarm Optimization Model ». New Trends in Applied Artificial Intelligence, 2007.

[27] A. Gherboudj, « Méthodes de résolution de problèmes difficiles académiques », p. 49.

[28]] F. Van den Bergh, « Analysis of Particle Swarm Optimizers ». PhD Thesis. University of Pretoria, 2001.

[29] R.C. Eberhart, P. Simpson, R. Dobbins, « Computational PC Tools ». chapter 6, AP Professional., 1996.

[30] R.C. Eberhart, Y. Shi, « Comparing inertia weights and constriction factors in particle swarm optimization ». Proceedings of the 6th IEEE Congress on Evolutionary Computation, IEEE Press, 2000.

[31] H.Y. Fan, Y. Shi, « Study on Vmax of particle swarm optimization. Proceedings of the 2001 Workshop on Particle Swarm Optimization ». Indiana University-Purdue University Indianapolis Press, 2001.

[32] X. Cai, Y.Tan, « A study on the effect of vmax in particle swarm optimization with high dimension ». International Journal of Bio-Inspired Computation (IJBIC), 2009.

[33] K. Deep, J. C. Bansal, « Hybridization of particle swarm optimization with quadratic approximation ». OPSEARCH. Vol, 2009.

[34] J. Barrera, C. A.C. Coello, « Limiting the velocity in particle swarm optimization using a geometric series ». Genetic And Evolutionary Computation Conference, Proceedings of the 11th Annual conference on Genetic and evolutionary computation,pp, 2009.

[35] R.C. Eberhart, Y. Shi, « Comparing inertia weights and constriction factors in particle swarm optimization ». Proceedings of the 6th IEEE Congress on Evolutionary Computation, IEEE Press, 2000.

[36] A. Chatterjee, P. Siarry, « Nonlinear inertia weight variation for dynamic adaptation in particle swarm optimization ». Computers & Operations Research, 2006.

[37] S-K S. Fan, J-M Chang., « A Modified Particle Swarm Optimizer Using an Adaptive Dynamic Weight Scheme ». Digital Human Modeling, 2007.

[38] X. Zhang, H. Qiu, « Hybrid particle swarm optimisation with k-centres method and dynamic velocity range setting for travelling salesman problems ». International Journal of Bio-Inspired Computation, 2010.

[39] M. Clerc, « The swarm and the queen: towards a deterministic and adaptive particle swarm optimization ». Proceedings, 1999 ICEC, Washington, DC, 1999.

[40] Yu N, Guo X, Zelikovsky A and Pan Y 2017, « GaussianCpG: A Gaussian model for detection of CpG island in human genome sequences ». BMC Genomics, 2017.

[41] Sujuan Y, Asaithambi A and Liu Y, « CpGIF: An algorithm for the identification of CpG islands ». Bioinformation, 2008.

[42] Elango N and Soojin VY, « Functional relevance of CpG island length for regulation of gene expression ». Genetics, 2011.

[43] Sujuan Y, Asaithambi A and Liu Y, « CpGIF: An algorithm for the identification of CpG islands ». Bioinformation, 2008.

[44] Hackenberg M, Previti C, Luque-Escamilla PL, Carpena P, et Martı´nez-Aroza J and Oliver JL, « CpGcluster: A distancebased algorithm for CpG-island detection ». BMC Bioinform, 2006.

[45] Hackenberg M, Carpena P, Bernaola-Galva´n P, Barturen G, et Alganza A´ M and Oliver JL, « WordCluster: Detecting clusters of DNA words and genomic elements. » Algorithms Mol. Biol. 6, 2011.

[46] Yoon B-J and Vaidyanathan P, « Identification of CpG islands using a bank of IIR lowpass filters [DNA sequence detection]; in Digital Signal Processing Workshop ». and the 3rd IEEE Signal Processing Education Workshop IEEE, 2004.

[47] Wu H, Caffo B, Jaffee HA, Irizarry RA and Feinberg AP, « Redefining CpG islands using hidden Markov models ». Biostatistics, 2010.

[48] Chuang L-Y, Yang C-H, Lin M-C, Yang C-H, « CpGPAP: CpG island predictor analysis platform ». BMC Genet, 2012.

[49] Churchill GA, « Stochastic models for heterogeneous DNA sequences. B ». Math. Biol, 1989.

[50] Gardiner-Garden M, Frommer M, « CpG islands in vertebrate genomes ». J Mol Biol, 1987.

[51] M. Clerc, J. Kennedy, « The particle swarm: explosion, stability, and convergence in multi-dimensional complex space ». IEEE Transactions on Evolutionary Computation, 2002.

[52] Olson SA, « EMBOSS opens up sequence analysis ». Brief Bioinform, 2002.

[53] Hackenberg M, Previti C, Luque-Escamilla PL, Carpena P, Martinez-Aroza J, Oliver J, « CpGcluster: a distance-based algorithm for CpG-island detection. » BMC Bioinformatics, 2002.

[54] Ponger L, Mouchiroud D, « CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences ». Bioinformatics, 2002.

[55] Takai D, Jones PA, « Comprehensive analysis of CpG islands in human chromosomes 21 and 22. » Proceedings of the National Academy of Sciences of the United States of America, 2002.

[56] Kennedy J, Eberhart R, « Particle swarm optimization ». pp. 1942–1948, 1995.

Année universitaire : 2020-2021	**Présenté par :** Bourghoud Yasmine Moumene Ferial

Titre

Nouvelle approche basée sur l'optimisation par essaim de particule pour la détection *in Silico* des ilots CpG dans le génome humain.

Mémoire pour l'obtention du diplôme de Master en Bioinformatique

Résumé

Dans ce mémoire nous avons abordé le sujet du cancer Le terme « cancer » englobe un groupe de maladies se caractérisant par la multiplication et la propagation anarchiques de cellules anormales. Si les cellules cancéreuses ne sont pas éliminées, l'évolution de la maladie va mener plus ou moins rapidement au décès de la personne touchée.

Un cancer peut être dû à des facteurs externes (mode de vie, facteurs environnementaux ou professionnels, infections), ou internes (mutations héréditaires, hormones, dérèglement du système immunitaire, etc.). Ces facteurs de risque peuvent agir ensemble ou de façon successive, et enclencher ou favoriser le développement du cancer. Souvent, plusieurs dizaines d'années séparent l'exposition à des facteurs externes et le déclenchement de la maladie.

Un cancer peut être soigné par un ou une combinaison de plusieurs traitements (chirurgie, radiothérapie, chimiothérapie, hormonothérapie, immunothérapie ou traitement ciblé).

Nous avons également établi quelques techniques pour la prédication du cancer avec quelques méthodes de résolutions que nous avons abordé.

Mots clés : cancer, mutations héréditaires, traitement ciblé.

Devant le jury :

Président du jury : Dr. Kamel KELLOU ; Université Frères Mentouri Constantine 1.

Encadreur : Dr. Amira GHERBOUDJ; Université Frères Mentouri Constantine 1.

Examinateur : Dr. Hamza CHEHILI; Université Frères Mentouri Constantine 1.

Date de soutenance : 30/09/2021