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 1
Université des Frères Mentouri Constantine 1

Mémoire

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

Intitulé

Etude Comparative des Méthodes in *Silico* pour la Détection des Ilots CpG

Présenté par :

Amira ZAROUR
Djihad CHELIA

Devant le jury composé de :

Présidant : Dr. Ines Bellil	MCA.UFM Constantine1
Encadreur : Dr. Amira GHERBOUDJ	MCA .UFM Constantine1
Examinateur : Dr. Mohamed Skander DAAS	MCA.UFM Constantine1

Année universitaire : 2020 / 2021


بسم الله الرحمن الرحيم

Résumé

Les îlots CpG sont généralement connus sous le nom de régions de régulation épigénétique, en fonction des modifications de l'histone, de la méthylation et de l'activité du promoteur. La cartographie exacte de la méthylation de l'ADN dans les îles CpG est nécessaire pour comprendre les diverses fonctions biologiques. Cependant, l'identification précise des îlots de CpG à partir du génome entier par des approches expérimentales et computationnelles reste difficile. De nombreuses méthodes de calcul sont développées pour détecter les régions enrichies de CpG, de manière efficace, afin de réduire le temps et le coût des expériences. Ici, nous passons en revue quelques-unes des dernières méthodes de détection de CpG computationnelle qui utilisent le clustering, les modèles et les paramètres de distance physique comme pour la détection d'île de CpG. Les analyses comparatives des méthodes reposant sur différents principes et paramètres permettent de prioriser les algorithmes pour des ensembles de données spécifiques CpG associés afin d'obtenir une plus grande précision et sensibilité. Un certain nombre d'outils de calcul basés sur la fenêtre, les algorithmes basés sur la densité et la distance / longueur sont appliqués sur le génome humains pour la détection précise de CpG. Plusieurs algorithmes ont été développés pour l'identification CGI et ont été appliqués dans de nombreuses études. Ils peuvent être classés en deux groupes : les algorithmes traditionnels qui sont basés sur trois paramètres de séquence (longueur, contenu GC, et rapport de la CpGs observée sur la CpGs attendue (CpG_{Obs}/CpG_{Exp})) et les algorithmes basés sur la propriété statistique dans une séquence sans imposer les trois critères dans les algorithmes traditionnels.

Mots-clés : Bioinformatique ; épigénétique ; île CpG ; Génome humain ; algorithmes de détection.

Abstract

CpG islands are generally known as the epigenetic regulatory regions in accordance with histone modifications, methylation, and promoter activity. There is a significant need for the exact mapping of DNA methylation in CpG islands to understand the diverse biological functions. However, the precise identification of CpG islands from the whole genome through experimental and computational approaches is still challenging. Numerous computational methods are being developed to detect the CpG-enriched regions, effectively, to reduce the time and cost of the experiments. Here, we review some of the latest computational CpG detection methods that utilize clustering, patterns and physical-distance like parameters for CpG island detection. The comparative analyses of the methods relying on different principles and parameters allow prioritizing the algorithms for specific CpG associated datasets to achieve higher accuracy and sensitivity. A number of computational tools based on the window density and distance-/length-based algorithms are being applied on human genomes for accurate CpG detection..Multiple algorithms have been developed for CGI identification and have been applied in numerous studies. They could be categorized into two groups: the traditional algorithms that are based on three sequence parameters (length, GC content, and ratio of the observed over the expected CpGs (Obs_{CpG}/Exp_{CpG})) and the algorithms based on statistical property in sequence without imposing the three criteria in the traditional algorithms.

Keywords: Bioinformatics; epigenetic; CpG Island; human genome; algorithms of detection

ملخص

استنادًا إلى تعديل هيستون والمثيلة ونشاط المروج ، غالبًا ما يشار إلى جزر سيتوزين فوسفات جوانين على انها مناطق تنظيمية جينية' من الضروري جدا تحديد موقع مثيلية الحمض لفهم الوظاف البيولوجية المختلفة. ومع ذلك ' لا يزال التحديد الدقيق لجزر سيتوزين فوسفات جوانين من الجينوم باكمله من خلال لاساليب التجريبية والحسابية يمثل تحديا.

يتم تطوير مجموعة متنوعة من طرق الحساب بشكل لفعال لتقليل وقت وتكلفة التجربة هنا'نراجع بعضا من المناطق المخصبة جزر سيتوزين فوسفات جوانين الحسابية 'والتي تستخدم معلمات مثل التجميع والوضعالذي يسمح بالتحليل المقارن والمسافة المادية كاحدث طرق اكتشاف جزر سيتوزين فوسفات جوانين على مبادئ ومعايير مختلفة بتحدبد اولويات الخوارزميات لمجموعات بيانات محددة لتحقيق دقة وحساسية اعلى متعلقة بجزر سيتوزين فوسفات جوانين.

يتم تطبيق العديد من ادوات الحساب على اساس كثافة النافذة والخوارزميات القائمة الدقيق. تم تطوير على المسافة/الطول على الجينوم البشري لجزر سيتوزين فوسفات جوانين وتم تطبيقها في الكثير من الدراسات .يمكن تقسيمها الى فئتين :

1-الخوارزميات التقليدية بناءا على معلومات التسلسل (الطول'نسبة الملاحظة و المحتور)

2-الخوارزميات القامة على خاصية احصائية متسلسلة دون فرض المعايير الثلاثة في الخوارزميات التقليدية

الكللمات الرئيسية: الحوسبة البيولوجية علم الوراثة CpG الجينوم البشري خوارزميات الكشف

Remerciements

Au nom de dieu le tout puissant et miséricordieux à qui nous exprimons toutes nos reconnaissances et remerciements de nous avoir donné l'inspiration et le pouvoir d'achever notre projet de Master.

« الحـــــــــــــــــــمدلله »

Nous adressons nos remerciements les plus sincères à toutes les personnes qui nous ont permis la rédaction et la réalisation de ce travail. Plus particulièrement, nous tenons à remercier :

- *Docteur Amira Gherboudj, directrice de mémoire, pour nous avoir accordé sa confiance pour la réalisation de ce projet, et pour nous avoir guidées tout au long de cette étude.*
- *Nous souhaitons particulièrement adresser nous vifs remerciements aux membres du jury, pour avoir accepté de participer à l'évaluation de notre mémoire.*
- *Les membres de l'équipe pédagogique de la spécialité bioinformatique, nos enseignants pour leurs soutiens et leurs encouragements*
- *Nos amies, pour nous avoir soutenues et comprises durant cette année, et pour avoir partagé un si belle complicité pendant ces années d'études.*

Dédicaces

Je souhaite personnellement remercier mon binôme et amie Amira, avec laquelle j'ai pris beaucoup de plaisir à travailler. Nous avons formé une belle équipe. Enfin, j'aimerais remercier ma famille pour leur soutien, leur amour et leur encouragement à toute épreuve. Plus particulièrement ma mère, mon père, mes frères, mes sœurs et mon mari « Bilal », qui ont toujours été d'un soutien indéfectible pour moi.

« Djihad »

Je Dédie ce mémoire

A Djihad ma partenaire de mémoire, mon binôme, ma copine avec qui j'ai beaucoup amusée à travailler avec elle. Cette année fut riche en émotions et je tiens à te remercier pour ton soutien et ce lien tout particulier qui s'est créé entre nous …

A mes chers parents ma mère et mon père pour leur patience, leur amour et leur soutien

A mes frères, mes sœurs et mon fiancé pour leurs encouragements.

« Amira »

Table des matières

Liste des tableaux

Liste des figures

Liste des abréviations

ACC : La précision

ADN : Acide désoxyribonucléique

CC : Coefficient de corrélation (corrélation coefficient)

CGI: CpG Island

CpG : Cytosine Phosphate Guanine

CpGo/e : Rapport entre le nombre de CpG observés et le nombre de CpGattendus

CpGIF : Recherche d'île CpG (CpG Island Finder)

CpGIS : Chercheur d'îles (Island Searcher)

GGF : Gardiner-Garden &Frommer

O/E: Observé/attendu **(**observed/expected)

PC : Coefficient de performance **(**performance coefficient)

SN : Sensibilité (sensitivity)

SP : Spécificité **(**specificity)

TSS: Site d'initiation de la transcription (Transcription Start Site)

Introduction Générale

Les nouvelles perspectives du domaine de la bioinformatique et de la médecine personnalisée représentent l'identification des origines et des causes du cancer et les maladies chroniques. Cette identification permet de limiter les conséquences graves des maladies par l'innovation des nouvelles thérapies basées sur les biomarqueurs de diagnostic et de ciblage des médicaments au corps humain. L'une des origines des maladies génétiques liées est le fameux processus moléculaire " Épigénétique ", qui vise une nouvelle approche thérapeutique basée sur l'utilisation d'inhibiteurs spécifiques au niveau de l'expression génétique.

La méthylation de l'ADN est l'un des facteurs épigénétiques les plus étudiés dans le cadre du dépistage du cancer. Détecter la méthylation de certains promoteurs de gènes permettrait d'identifier un cancer en particulier et d'améliorer, ainsi, le diagnostic lors de l'apparition de cellules cancéreuses. La méthylation de l'ADN se produit sur le carbone 5 d'une cytosine dans un dinucléotide CpG (Cytosine-Phosphate-Guanine) chez les mammifères.

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 la 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 en tant 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.

Les méthodes de détection des îlots CpG sont principalement classées en quatre classes en fonction de leurs algorithmes principaux : les méthodes basées sur les fenêtres, les méthodes basées sur la densité et les méthodes basées sur la distance/la longueur et Le modèle de Markov caché (HMM) qui sont appliquées à la détection des ilots CpG par calcul. Ces méthodes sont basées sur les critères proposés par Gardiner-Garden et Frommer (GGF) [1] y compris le contenu GC, le rapport O/E et le seuil de longueur de l'ilot.

Organisation du Mémoire

En plus de l'introduction et la conclusion générales, notre travail est principalement structuré en trois chapitres.

- Le chapitre 1, présente le contexte biologique de notre travail à savoir, les concepts biologiques nécessaires pour la compréhension du problème traité : la génomique, l'épigénétique et méthylation, Dinucléotides CpG méthylés et cancer, ainsi que les îlots CpG.
- Le chapitre 2, décrit les méthodes d'identification des îlots CpG présentées dans la littérature, y compris les : méthodes basées sur les fenêtres coulissantes, méthodes basées sur la densité, méthodes basées sur la distance/longueur, méthodes basées sur le modèle de Markov caché (HMM).
- Le chapitre 3, décrit une étude comparative. Nous utilisons les méthodes CpGclustrer, CpGProt, CpGProd, CpGISetCpGIF pour la détection des ilots CpG dans le génome humain. En outre, nous présentons une étude comparative des méthodes utilisées ainsi que des résultats obtenus.

Chapitre 1 :
Les ilots CpG dans le génome humain

1.1 Introduction

L'épigénétique est un domaine en pleine croissance qui étudie les altérations héréditaires, impliquées dans la régulation d'expression génétique qui ne sont pas dues à des modifications de la séquence de l'ADN.

Les îlots CpG ont été identifiés à l'origine par des propriétés épigénétiques et fonctionnelles, à savoir l'absence de méthylation de l'ADN et l'association fréquente de promoteurs. Cependant, ce concept a rapidement été remplacé par des critères de séquence d'ADN simples, qui ont permis l'annotation des îlots CpG à l'échelle du génome en l'absence d'ensembles de données épigénétiques à grande échelle.

Dans ce chapitre, nous allons aborder en premier lieu la génomique, par la suite nous allons présenter la notion épigénétique et méthylation d'ADN. Dans la troisième partie nous allons décrire la relation entre la méthylation des CpG et le cancer. Enfin, la quatrième partie concerne les îlots CpG.

1.2 Génomique

La génomique est la science qui étudie le génome d'un individu ou d'une tumeur comme la fonction des gènes ou encore le contrôle de leur expression. Cela permet d'en apprendre plus sur le fonctionnement de l'organisme humain et sur certaines maladies, héréditaires ou non[1].

Le génome désigne l'ensemble des gènes contenus dans nos cellules, il représente le matériel génétique d'un organisme. Le génome contient à la fois les séquences codantes, c'est-à-dire celles qui codent pour des protéines, et les séquences non codantes. Chez la plupart des organismes, le génome correspond à l'ADN présent dans les cellules. Le génome humain contient six milliards de bases, soit la quasi-totalité de notre patrimoine génétique.

La séquence complète du génome humain contenu dans le noyau de la cellule sous forme d'ADN a été finalisée en 2006 [1].

1.3 Epigénétique et méthylation

Le terme « épigénétique» est utilisé lorsque l'on veut décrire des altérations de la séquence d'ADN sans aucune modification de la séquence elle-même. En fait, l'épigénétique exprime les modifications transmissibles et réversibles de l'expression des gènes présents dans notre génome sans aucune modification de la séquence des nucléotides. Les deux principales modifications épigénétiques sont la méthylation de l'ADN et une altération de la structure des histones[2].

La méthylation de l'ADN se produit sur le carbone 5 d'une cytosine dans un dinucléotide CpG (Cytosine-Phosphate-Guanine) chez les mammifères[2] (voir figure 1.1).

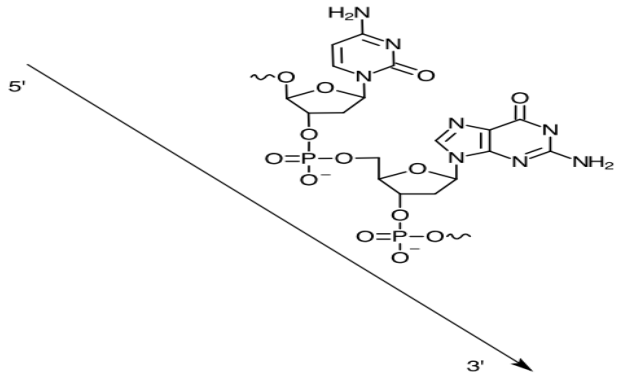

Figure 0.6: Dinucléotides Cytosine_phosphate_Guanine(CpG)[3]

Cette méthylation se fait via un DNMT (DNA méthyle-transférase) et un S-Adénosyle-L-Méthionine (SAM), qui deviendra un S-Adénosyle homocystéine (SAH), selon une substitution électrophile (Figures 2). De manière globale, environ 70% des dinucléotides CpG sont méthylés au sein du génome. La figure 1 schématise la structure d'un dinucléotide CpG soit la présence d'une cytosine suivie immédiatement par une guanine, reliées entre elles par un groupement phosphate, selon le sens 5' vers 3'. Ainsi, le début d'une séquence ADN débute par la section 5' et se termine à la section 3'. Par contre, environ 15% des CpG de l'ADN sont protégés de la méthylation et se retrouvent dans une section de l'ADN appelée îlots CpGs. De manière générale, ces îlots CpGs sont eux-mêmes situés dans les promoteurs.

Ces promoteurs sont situés, généralement, à l'extrémité 5' d'un gène et permettent l'expression du gène. Ils permettent l'initialisation de la transcription et permettent ainsi au gène d'exprimer son caractère. En étant protégés de la méthylation, les facteurs de transcription peuvent avoir lieu. Par contre, lorsqu'il y a un taux anormal de groupe méthyle dans les sections promotrices, il y a une non-reconnaissance du site et les facteurs de transcription ne peuvent pas se lier à la section promotrice et ainsi, il n'y a aucune expression du gène[4].

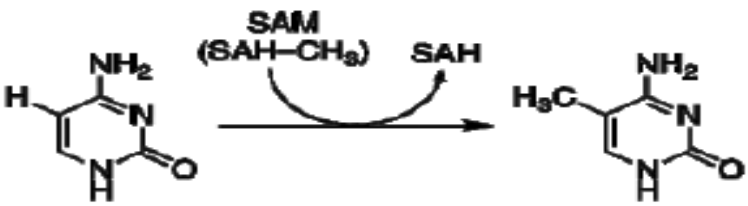

Figure 0.7:Réaction de méthylation d'une cytosine

1.4 Dinucléotides CpG méthyles et cancers

Les promoteurs sont reliés à la transcription d'un gène et permettent à ce dernier de s'exprimer et, par ce fait mènent à l'émission d'une caractéristique spécifique. Ces sections promotrices permettent également un travail de régulation des gènes pour qu'ils puissent s'exprimer adéquatement et sont représentées dans la figure (1.3). Il y aurait environ 21 000 gènes dans la séquence d'un humain et chacun exprime un caractère particulier. La méthylation dans ces sections promotrices a été étudiée et est reliée au néoplasie des cellules. Le néoplasie se définie par la prolifération de cellules dans une formation nouvelle et qui ne présente aucune structure. Cette nouvelle masse de cellules ne possède aucun lien avec son environnement et peut être considérée comme bénigne ou maligne. Dans cette optique, l'hyperméthylation a été étudiée dans les sections promotrices des gènes qui sont considérés suppresseurs de tumeurs et, de manière fréquente, cette aberrante méthylation est reliée à la tumorigénèse [2].

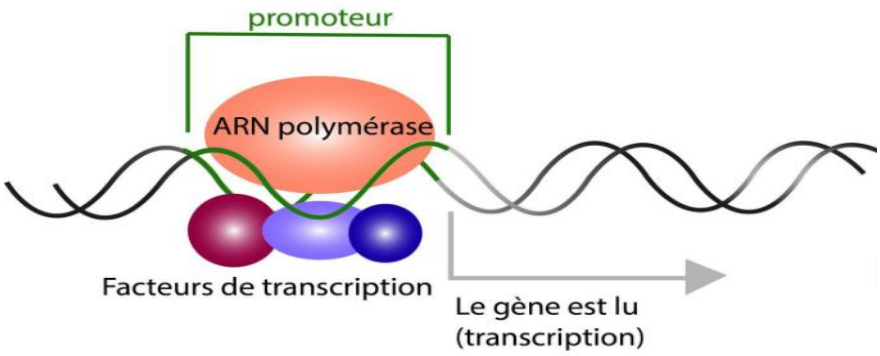

Figure 0.8:Schématisation d'un promoteur et d'un gène[5]

Lorsqu'une ou quelques cellules commencent à se multiplier de manière chaotique et que le gène associé à la suppression de ce type d'aberration ne peut s'exprimer dû à l'hyperméthylation de son promoteur, il peut y avoir l'apparition d'une tumeur maligne. Certaines recherches ont été effectuées pour approfondir les liens entre différents gènes et différents cancers.

L'analyse de ces promoteurs reliés à des gènes de type tumeur-suppresseur a permis d'établir une corrélation entre certaines régions promotrices et certains cancers. En effet, l'hyperméthylation peut être envisagée de deux manières, soit l'analyse d'un gène vers tous les types de cancers ou soit l'analyse des gènes reliés à un cancer en particulier.

Premièrement, lorsqu'un promoteur est méthylé, il peut indiquer la possibilité d'un cancer mais, généralement, il est marqueur pour plus d'un cancer[6].

La perte de méthylation de l'ADN dans les régions du génome est appelée" hypométhylation de l'ADN ", elle est associée à une instabilité génomique et à la progression du cancer[7].En outre, la perte d'intégrité ou la perte de l'expression monoallélique de gènes par des allèles parentaux en raison d'une hypométhylation aberrante de l'ADN peut causer un risque accru de cancer.

Dans les cellules tumorales, en particulier, l'hypométhylation globale s'accompagne d'une hyperméthylation d'îlots CpG associés à un promoteur qui restent généralement non méthylés dans des cellules normales. Ce modèle unique de méthylation de gènes individuels est la caractéristique couramment observée dans les divers gènes suppresseurs de tumeurs dans la plupart des types de cancers humains et sert de substitut aux mutations ou délétions ponctuelles qui provoquent la désactivation de la transcription des gènes suppresseurs de tumeurs [8].

1.5 Îlots CpG

Les îlots CpG ont été très étudiés depuis leur découverte il y a une trentaine d'années, particulièrement chez les mammifères, du fait de leur 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[9].

1.5.1 Découverte des îlots CpG

Cooper, Bird et ses collègue [10] sont mis en évidence au début des années 1980 qu'une petite 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 sur tout en dinucléotides CpG, ont été appelés par la suite îlots CpG[10].

1.5.2 Localisation des îlots CpG

Les sites de dinucléotide cytosine-guanine (CpG) sont des régions de la séquence ADN où la cytosine est suivie par guanine dans la disposition linéaire des nucléotides direction dans 5'

à 3'. Les îlots CpG (CGI) sont définies comme des groupes CpG dans l'ADN en vrac appauvri par le CpG, tel que décrit par Gardiner et Frommer. Les grappes CpG contenant une teneur élevée en GC et Le pourcentage de CpG proche du ratio attendu est reconnu comme Les CGI. La norme seuil a également été proposée pour les GCI, qui sont devenue un critère primaire dans toute prédiction CGI outils. Gardiner et Frommer ont décrit les CGI ayant les caractéristiques suivantes : une fréquence CpG supérieure à 0,6 dans le cas observé/prévu (O/E), plus de 50 % du GC contenu et longueur de la région insulaire supérieure à 200 Pb[11].

Des modifications rigoureuses ont été proposées avec 0,65 de la fréquence O/E, 55 % du contenu GC et une longueur minimale de 500 Pb. La longueur de séquence minimale de l'île a été amplifiée pour empêcher les séquences d'Alu.

Les séquences Alu sont décrites comme de courts éléments dispersés répétitifs d'environ 280 pb de longueur comprennent une haute fréquence O/E et Contenu GC. Les éléments d'Alu sont impliqués dans la régulation des gènes spécifiques aux tissus et parfois altèrent l'expression des gènes. Ces éléments causent également des mutations dans le génome humain. McClellan dtIvaria (1982) [12] a présenté un test de Chi-square pour l'attribution et identification des CGI statistiquement significatifs. Cette technique est considérée comme une approche appropriée selon la définition de l'GCI identifié statistiquement des groupes de PPC importants au sein de l'organisme Régions en voie d'épuisement du PPC.

1.6 Critères de détection des ilots CpG

Les CGIs (les ilots CpG) ont été initialement identifiés par [12] sous forme de petites zones qui contiennent les Hpall d'enzyme de restriction dans le génome et ont donc été appelés Hpall minuscule Fragment (HTF). Ensuite, Gardiner-Garden et Frommer (GGF) [13] ont défini les CGIs comme un segment d'ADN a une longueur supérieure à 200 Pb, une teneur en GC plus de 50% et le rapport CpG observé/ attendu ou CpG OE pas moins de 0,6. Le rapport est calculé de la manière suivante :

$$CpG_{Observ\,é/attendu} = \frac{N_{CpG} * N}{N_c * N_g} \qquad (1.1)$$

OuN_c, $N_g$et N_{CpG}sont respectivement des nombres de C, G et dinucléotidesCpG dans la région de longueur N. Par la suite, Takai et Jones [14] ont réévalué ces trois paramètres et ont proposé un nouvel ensemble de critères (GC_{ratio} >= 55%, $CpG_{O/E}$>= 0, 65, longueur >= 500Pb).

1.7 Conclusion

La méthylation de l'ADN est observée chez de nombreux organismes, elle s'effectue sur la cytosine. Cette base, méthylée, est hyper mutable par rapport à une cytosine non méthylée. Bien que non répliquée en soi, la méthylation de l'ADN est maintenue au cours des divisions cellulaires, mais certains génomes subissent des phases massives deméthylation/reméthylation à des stades précis. Il semble y avoir donc une régulation relativement fine de la répartition de la méthylation.

Ainsi les mammifères présentent un patron de méthylation de l'ADN qui couvre l'ensemble du génome 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.

Dans certains cas, les îlots CpG promoteurs peuvent être méthylés, réprimant ainsi l'expression du gène correspondant. L'hyperméthylation de l'ADN fait principalement référence à un gain de la méthylation sur des sites spécifiques, elle se produit principalement dans les îlots CpG situés aux promoteurs.

L'hyperméthylation des ilots CpG a été étudiée dans les sections promotrices des gènes qui sont considérés suppresseurs de tumeurs et, de manière fréquente, cette aberrante méthylation est reliée à la tumorigènes.

Dans le chapitre suivant nous présentons les méthodes in *silico* de détection des ilots CpG.

Chapitre 2 :

Méthodes de détection des ilots CpG

2.1 Introduction

Les algorithmes de détection des ilots CpG ont considérablement aidé dans la compréhension du profil de méthylation du génome et ont facilité la localisation des régions riches en CpG associées à la régulation des gènes.

Récemment, plusieurs outils et méthodes in *silico* ont été proposés pour détecter les CGIs, majoritairement, ces recherches sont basées sur les critères définis par Gardiner-Garden et Fromme [11].

Ces méthodes sont principalement classées en quatre classes en fonction de leurs algorithmes principaux en tant que méthodes basées sur :

1. Des fenêtres coulissantes.
2. La densité.
3. La distance/longueur.
4. Le modèle de Markov caché.

Dans ce chapitre, nous présentons quelques algorithmes et les approches existantes pour l'identification des CGIs dans le génome.

2.2 Méthodes basées sur les fenêtres coulissantes

Hudson et Kaplan en 1988 ont introduit la technique des fenêtres coulissantes (SWM) (Sliding Window Method) pour examiner les tendances de polymorphisme et de divergence. Il existe plusieurs méthodes de localisation des CGIs dans les séquences d'ADN en se basant sur cette approche. Le point commun entre ces méthodes est l'utilisation des critères de Gardiner-Frommer [13]. Parmi les travaux basés sur les fenêtres coulissantes, nous citons CpGplot, CpGProD, CpGIS, et CpG cluster.

2.2.1 CpGPlot

CpGPlot [15] représente la variante la plus simple de SWM. Le contenu GC et le rapport $CpG_{Obs/Exp}$ sont calculés sur une fenêtre de 100 Pb de longueur se déplaçant sur la longueur de séquence en incrémentant 10 Pb.

CpGPlot identifie les CGIs dans une ou plusieurs séquences de nucléotides. Le rapport entre les motifs observés et le nombre attendu de motifs de dinucléotides GC est calculé sur une fenêtre (région de séquence) qui se déplace sur la séquence. Les ratios calculés sont représentés graphiquement, ainsi que les régions qui correspondent à la définition de cette méthode d'un CGI (une zone riche en dinucléotides CG) [15].

2.2.2 CpGIS

Généralement appelé algorithme Takai-Jones[14] qui utilise une fenêtre de 200 Pb se déplaçant le long de la séquence avec des étapes de 200 pb. Il a un seuil supplémentaire pour les dinucléotides CpG minimaux dans la CGI prédite, égal à l'espérance mathématique de dinucléotides CpG dans la séquence de Bernoulli des probabilités de longueur et de nucléotides données. Cette espérance est multipliée par le seuil CpGObs/Exp. CpG Islande Searcher permet aux utilisateurs d'exclure des «CGI mathématiques » comme une séquence de 300 Pb avec 150 cytosines et une guanine dans un contexte CpG qui correspond aux critères standards de CGI. Cet algorithme fusionne également deux ou plusieurs CGI s'ils sont espacés de moins de 100 Pb. Takai et Jones suggèrent également d'utiliser des seuils plus stricts : de 500 pb pour la longueur des CGI, 0,55 pour le contenu en GC et 0,65 pour les CpGObs/Expafin de déterminer les CGI associés aux promoteurs de gènes connus codant pour les protéines et d'éviter les CGI associés aux répétitions Alu [16].

L'algorithme CpG Island Searcher (CpGIS) [16] a été conçu à l'origine selon les critères des îles CpG décrits par Gardiner-Garden et Frommer pour éviter de manquer les îles CpG qui répondent à ces critères. L'algorithme original a fusionné deux îles CpG alors qu'elles étaient séparées de moins de 100pb et que l'île fusionnée CpG répondait toujours aux critères. Cela diffère de l'approche utilisée par un autre site de recherche de l'île CpG, CpG plot. La dernière version du CpG Island Searcher permet à l'utilisateur de choisir la taille de l'écart entre deux îles CpG putatives. Des observations antérieures ont montré que l'utilisation de critères plus rigoureux d'augmentation des limites inférieures pour le %GC, CpGObs/CpGExp et la longueur conduisait à une meilleure association entre les îles CpG et les gènes.

Le programme permet d'utiliser des critères définis par l'utilisateur (c'est-à-dire la limite inférieure variable de %GC, ObsCpG/ExpCpG et la longueur) pour l'extraction de l'île CpG. Auparavant, une fenêtre coulissante de 200 Pb était utilisée pour la localisation initiale d'une île CpG. Cependant, l'algorithme révisé utilise les limites inférieures définies par l'utilisateur pour le balayage initial d'une séquence soumise afin d'éviter de rater les îlots CpG en utilisant les critères définis par l'utilisateur. L'utilisation d'une plus grande taille de fenêtre permet potentiellement l'extraction des îles CpG qui ne pourraient pas être extraites avec une taille de fenêtre plus petite. Puisque notre première priorité pour cet algorithme était de ne pas manquer de séquences répondant aux critères, le résultat initial pourrait différer de la perception de l'utilisateur. Ainsi, la lecture du site Web inclut une carte graphique des sites CpG afin que l'utilisateur puisse réduire la région de l'île CpG en utilisant un plus grand

%GC, CpGObs/CpGExpet une plus petite longueur de limite inférieure après la recherche initiale[16].

2.2.3 CpGProD

CpGProD [17] est un programme dédié à la prédiction des promoteurs associés aux CGIs dans la séquence génomique des mammifères. Dans chaque séquence trouvée par la fenêtre glissante et répondant aux critères de CGI, la probabilité de trouver un promoteur est estimée à :

$$P = \exp(Z) / (1 + \exp(Z)) \qquad (2.1)$$

Où Z est une combinaison linéaire de longueur CGI, du contenu GC et de CpGObs/Exp.

De plus, la probabilité qu'un brin soit un modèle pour la transcription est estimée selon l'équation précédente, où Z est une combinaison linéaire des biais AT et GC, propriétés connues de la séquence nucléotidique autour du TSS (sites de début de la transcription). Les coefficients pour Z sont estimés à partir de deux régressions linéaires généralisées formées avec deux jeux de données composés de CGI obtenant et non de TSS pour les gènes codant pour les protéines ou de deux jeux de données avec différents modèles de transcription chez l'homme [17].

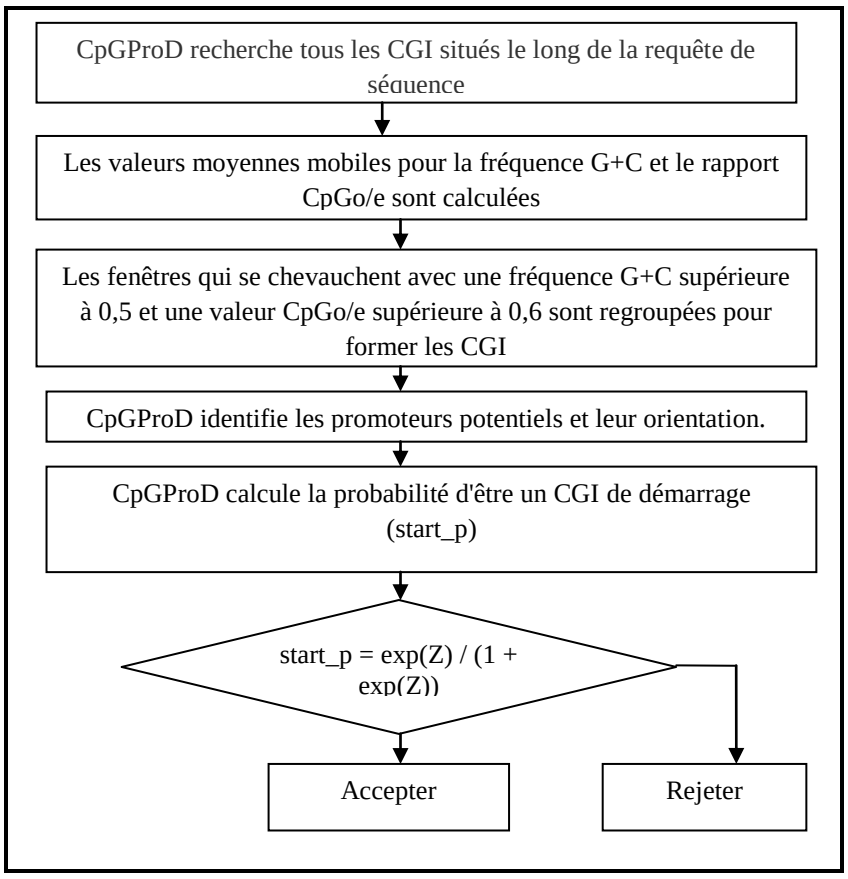

Figure 2.9 :ProD détecte les conditions CpG dans la séquence d'ADN

2.3 Méthodes basées sur La distance/longueur

Les méthodes basées sur la distance/longueur fournissent une approche rapide pour la prédiction des CGI qui assemble des données dans le contexte de la distance entre les sites CpG. Cette méthode examine la propriété de séquence entre deux sites CpG en ligne, ce qui amène également des critiques sur cette technique. La même CGI dans des situations différentes donne des résultats variés, la faible sensibilité prédictive avec des résultats non pertinents en raison de la composition de la séquence est également considérée comme l'inconvénient de cette technique. Toutes les méthodes existantes ont été basées sur le vaste espace de paramètres fait par le contenu du CG, la fraction CpG et le seuil de longueur. La distribution de la distance diffère dans les CGI et l'ADN en vrac entre les CpG adjacents en raison du nombre élevé de dinucléotides CpG dans les CGI (Hackenberg et al) a développé une nouvelle approche (CpG cluster) capable de déterminer directement les clusters CpG en fonction des distances physiques. Les grappes statistiquement significatives sont déclarées comme CGI après l'attribution de la valeur p à chaque groupe. La séquence de test a été extraite de la bibliothèque CGI expérimentale en comparant les prédictions de CpG cluster avec les méthodes CGI existantes. Il fournit le plus haut degré de chevauchement avec les éléments phylogénétiques conservés vertébrés et, en même temps, le plus faible chevauchement avec le rétro-transpose d'Alu.

Il a la capacité de différencier les CGI fonctionnels dans le génome en vrac parce que les CGI qui se chevauchent avec le site de début de la transcription présentent une importance statistique maximale par rapport aux autres îlots génomiques. La seule application d'entiers dans les opérations arithmétiques permet un outil de calcul rapide et efficace pour prédire des clusters précis et statistiquement significatifs. Le début et la fin avec le dinucléotide CpG de tous les CGI prédits est une autre caractéristique supérieure. CpG cluster ne fonctionne que sur la distance entre les CpGs adjacents qui conduisent à un faible chevauchement et à une grande spécificité avec les éléments Alu, tandis que d'autres paramètres de recherche dans d'autres méthodes de prédiction existantes ne sont pas requis comme principaux paramètres statistiques et de recherche[18].Dans cette classe, nous allons présenter l'algorithme CpG cluster.

2.3.1 CpG Cluster

CpG cluster [19] comporte deux étapes distinctes : une recherche de groupe CpG et une estimation de la probabilité de trouver un tel groupe par hasard. La distance entre les dinucléotides CpG voisins dans une séquence aléatoire est simulée par une loi géométrique avec la fréquence CpG en tant que paramètre. Hackenberg et al. Supposent que dans un cluster fonctionnel de CpG, la distance entre les CpG voisins est plus petite que prévu dans

une séquence aléatoire. Les auteurs montrent que les distances inférieures à la médiane de la distribution théorique sont surreprésentées dans le génome humain. La distance médiane entre les CpG voisins (23-53 Pb, en fonction du chromosome) est utilisée comme seuil. Chaque cluster est donc composé de CpG situés non au-delà du seuil. Tous les CGI résultants commencent et finissent par un dinucléotide CpG. Chaque cluster a une p-value calculée en fonction de la distribution binomiale négative. Seules les grappes dont la valeur p est inférieure à 1,0e-5 (1,0e-20 dans) sont considérées comme des CGIs. Les auteurs trouvent environ 200 000 dans le génome humain (25 000 îles CpG avec un seuil de p-value égal à 1,0e-20). Un grand nombre de ces îles CpG ont moins de 200 pb. Cependant, les auteurs montrent les fonctionnalités de quelques CGI et les appellent des *isletsCpG* [20].

2.3.1.1 Algorithme CpG Cluster

La méthode de recherche par grappe est basée sur les propriétés statistiques des distances physiques entre les dinucléotides CpG voisins sur la séquence d'ADN. En principe, si les CpG sont distribués totalement au hasard le long de la séquence chromosomique, les distances entre les dinucléotides CpG voisins devraient suivre la distribution géométrique :

$P(d) = (1 - p)\, d\text{-}1p$

où P(d) représente la probabilité de trouver une distance d entre les CpGs voisins et p correspond à la probabilité de CpGs dans la séquence, calculée comme le rapport entre les CpGs et le nombre total de dinucléotides dans la séquence d'ADN [20].

L'algorithme effectue les opérations suivantes :

1. La séquence chromosomique de l'ADN est scannée pour les dinucléotides CpG, puis l'enregistrement des positions occupées par le 'C' : x1, x2 ... xN, N étant le nombre total de CpGs dans la séquence. La séquence était habituellement scannée dans la direction 5' → 3'. Trivialement, le balayage inverse (3' → 5') produit les mêmes résultats.

2. Par convention, la distance physique séparant deux CpGs voisins est définie comme suit: $d\ i = xi+1- x\ i - 1$, de sorte que la distance minimale entre deux CpGs voisins (CGCG) est égale à 1.

3. Au cours de l'analyse, la première distance au-dessous d'un seuil donné (d t) identifie le premier groupe CpG. Le seuil d t peut être aisément dérivé de la distribution des distances entre les CpGs voisins dans la séquence chromosomique. La distance médiane donne souvent les meilleurs résultats parce que la distance médiane de la distribution observée est

approximativement au point de transition des petites distances surreprésentées (intra-grappe) et des distances intermédiaires sous-représentées. Ce n'est pas une propriété exclusive du chromosome 1, car il est partagé par tous les chromosomes, ce qui indique que la distance médiane peut être choisie en général comme un bon seuil (d_t).

4. Nous essayons ensuite d'étendre ce premier cluster en aval (→ 3') en ajoutant le prochain CpG alors que les distances sont inférieures à d t. Lorsqu'une distance supérieure à d t est trouvée, le cluster est terminé et la recherche d'un nouveau cluster se poursuit en aval

5. Les étapes 3 et 4 sont répétées jusqu'à ce que tous les groupes CpG de la séquence soient identifiés.

Notez que cet algorithme acquiert deux caractéristiques importantes et distinctives par construction. Premièrement, tous les CGI prédits commencent et se terminent par un dinucléotide CpG, ce qui semble approprié. Deuxièmement, l'algorithme n'utilise que l'arithmétique des entiers, ce qui est efficace sur le plan du calcul. Aucun autre algorithme de recherche CGI ne partage ces deux propriétés importantes. La figure 2.1 montre un algorigramme qui résume l'algorithme CpG cluster.

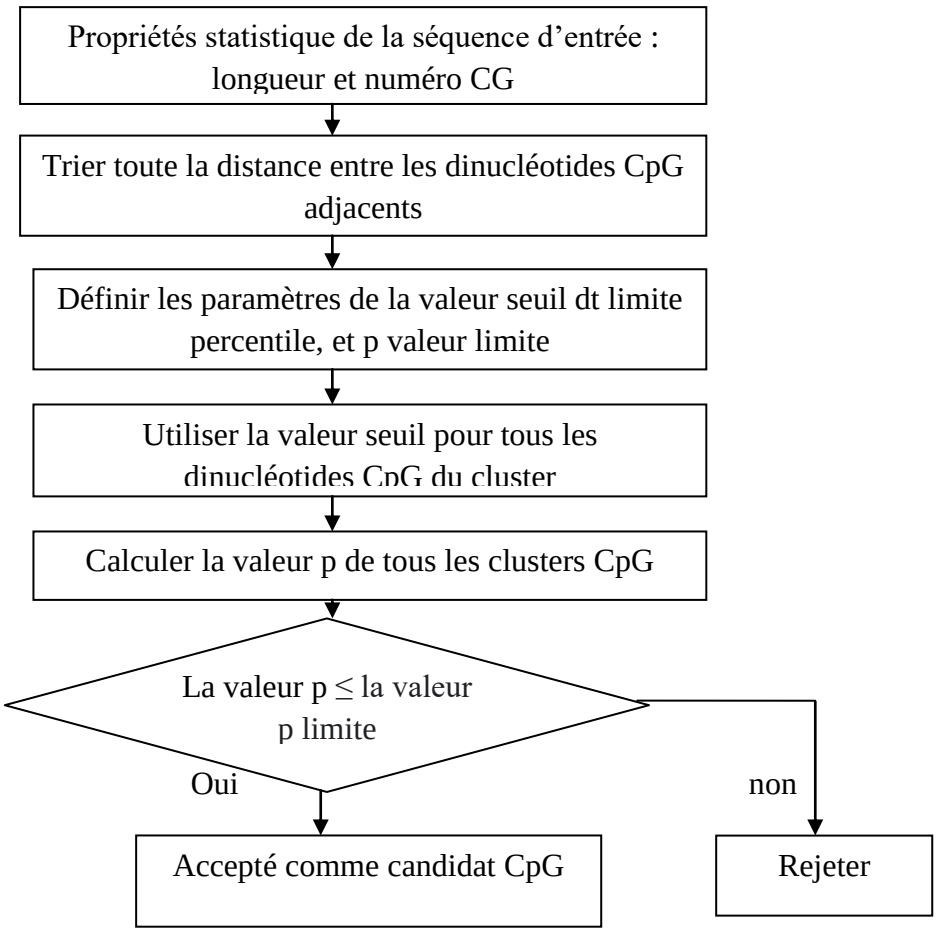

Figure 0.10: Organigramme du point de vente du cluster.

2.4 Méthodes basées sur La densité

Les méthodes basées sur la densité déterminent instinctivement la densité des sites CpG comme les méthodes basées sur les fenêtres, qui utilisent les normes statistiques. Le pourcentage de sites CpG dans CGI et la longueur totale de CGI sont calculés pour calculer la densité CGI. Le principe de base de cette méthode est de définir les semences initiales, réguler de façon répétitive les variables de densité et augmenter ainsi la couverture des régions riches en CpG. Initialement, les limites estimées des CGI sont analysées en ajustant une valeur seuil lâche/faible de la densité. Par la suite, la valeur seuil stricte/élevée est utilisée pour déterminer l'étendue des bordures de la GCI, où la séquence répond aux critères de densité. Cependant, un modèle linéaire ne pourrait probablement pas définir la distribution de CpG dans CGI, alors que cette méthode dépend fortement du seuil de densité qui indique l'association linéaire des sites de CpG et la longueur totale de CpG, ce qui est considéré comme un inconvénient de cette méthode[18].

2.4.1 CpGIF

Dans CpGIF [21] un seuil de densité est appliqué pour exclure les "îles CpG mathématiques" causées par un rapport G/C (ou C/G) élevé. La même coupure a également été utilisée dans certains outils précédents, comme CpGIS.

CpGIF se compose de quatre étapes principales. D'abord, nous scannons la séquence d'ADN de 5’ à 3’ fin pour trouver tous les dinucléotides CpG et enregistrer leurs positions. Ensuite, nous essayons d'identifier toutes les graines initiales avec une densité par défaut de 0,10. Dans cette étape, un tableau est construit pour enregistrer les nombres de Gs et de Cs dans chaque graine initiale et dans la région située entre deux graines adjacentes. Au cours des étapes suivantes, nous continuerons de mettre à jour la matrice pour calculer le contenu du CG et le ratio CpG/e. Ensuite, les graines initiales sont étendues itérativement en diminuant la limite de densité de 0,09 à 0,05. La limite est réduite de 0,01 à chaque itération. Enfin, deux graines étendues voisines sont regroupées si la distance entre elles est inférieure à la longueur maximale de deux graines étendues adjacentes ou de 100 nt, la plus petite de ces deux valeurs étant retenue.

Pour évaluer la performance de prédiction de CpGIF et la comparer à d'autres programmes, nous avons créé un ensemble de séquences de test en utilisant la même méthode décrite par La longueur de chaque CGI connue dans notre séquence d'essai est d'au moins 200 nt puisque le même critère de longueur a été utilisé dans tous les programmes testés, à l'exception de CpG Cluster[7].

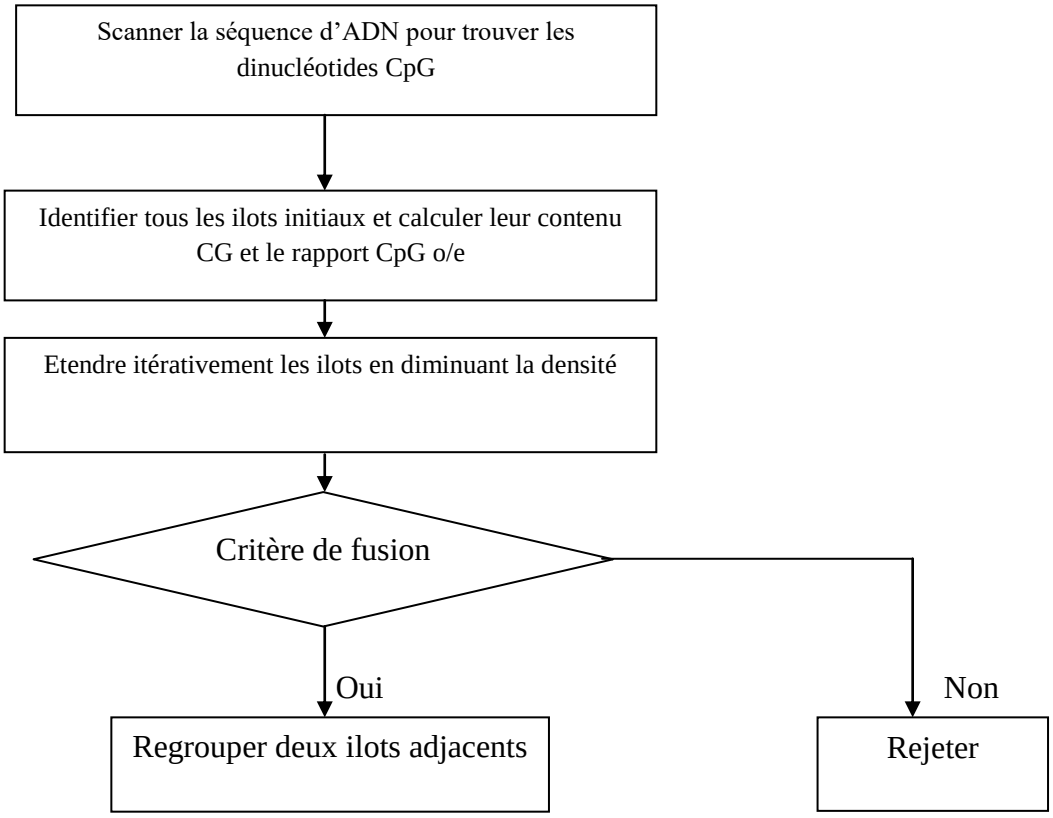

Figure 2.3 : IF détecte les conditions CpG dans la séquence d'ADN

2.5 Méthodes basées sur Le modèle de Markov caché (HMM)

La publication classique sur les HMM de Rabiner en 1989[22] a conduit une grande partie de la recherche dans le domaine des applications HMM en général. L'application initiale à la reconnaissance vocale a ensuite été adaptée aux applications biologiques dans les données de séquence. L'article de Rabiner décrit les algorithmes appropriés à utiliser et les astuces mathématiques pour contourner les problèmes de calcul avec la multiplication de petites valeurs fractionnelles. En 2006, un article de Mann a mis au point les moyens par lesquels les HMM pouvaient être étendus aux modèles testés à partir de plus grandes quantités de données, en appliquant des échelles et des logarithmes. Ces techniques sont maintenant appliquées comme méthodologie standard dans les algorithmes HMM.

L'utilisation de chaînes Markov cachées pour modéliser des séquences d'ADN a été initiée par Churchill en 1989, et depuis lors, leur utilisation à cette fin a augmenté. Diverses approches alternatives ont été introduites pour améliorer la précision de ces prédictions, l'une étant les modèles de Markov cachés. Plusieurs algorithmes basés sur l'utilisation d'un HMM ont gagné en popularité.

Le processus d'exécution d'un HMM pour identifier les îlots CpG dans les données de séquences d'ADN induit la séquence d'état caché la plus probable parmi toutes les séquences possibles, sous réserve de l'observation de la séquence de tous les nucléotides de la séquence. Cette séquence d'états cachés est alors acceptée comme étant "correcte sur le plan déterministe" (c.-à-d. la séquence d'états cachés la plus susceptible d'expliquer la séquence observée en fonction des paramètres du modèle) et des motifs tels que les îlots CpG sont trouvés en examinant la séquence. Aston et al. Se sont concentrés sur une analyse plus approfondie de la séquence de l'état caché et ont développé une méthode de calcul pour trouver de telles distributions de patrons qui ont identifié les îles CpG. Bien qu'ils aient reconnu les avantages de l'utilisation de MHM de commande plus élevée, ils ont fait remarquer que l'une des raisons pour lesquelles les MHM de commande plus élevée ne sont pas utilisées plus fréquemment est la complexité et le nombre croissant de paramètres du modèle à mesure que l'ordre augmente.

Différents algorithmes CpG, basés sur le même ensemble de séquences de données, produisent très communément des prédictions très différentes de l'île CpG. Hsieh et al. Ont souligné que ces identifications incohérentes avec des non-chevauchements significatifs indiquent que chacun de ces algorithmes peut manquer une fraction élevée des îles CpG véritables. D'autre part, ils ont fait remarquer que les chercheurs d'îles CpG qui ont une assez bonne sensibilité ne peuvent être calculés que pour des séquences de génome relativement courtes. Étant donné qu'il n'existe pas de définition opérationnelle unique d'une île CpG, les définitions de l'île CpG sont mathématiquement incomplètes, ce qui nuit à la faisabilité de toute recherche exhaustive fondée sur des critères de filtrage. Pour remédier à cette situation, ils ont proposé une approche très technique pour diagnostiquer le résultat d'un processus HMM, ce qui a permis d'identifier des "cœurs" de dinucléotides CpG agrégés. Aucune analyse comparative avec d'autres méthodes concurrentes n'a été fournie.

2.6 Conclusion

Nous sommes vus dans ce chapitre les différentes catégories des méthodes d’identification des îlots CpG et nous avons détaillé le principe des algorithmes les plus évidentes et les plus largement utilisées comme modèle de chaque catégorie.

Dans le chapitre suivant, nous présentons l’étude expérimentale que nous avons réalisée sur ces méthodes appliquées pour la détection des ilots CpG dans le génome humain.

Chapitre 3 :
Etude expérimentale

3.1 INTRODUCTION

Les méthodes actuelles de détection des îlots CpG sont principalement basées sur les critères proposés par Gardiner-Garden et Frommer (GGF)[11] y compris le contenu GC (équation 3.1), le rapport O/E (équation 3.3) et de longueur de l'ilot (équation 3.2). Gardiner-Garden et Frommer ont défini l'ilot CpG comme une séquence ADN dont :la plage de longueur de l'ilot≥ 200 pb, le contenu GC ≥ 50 %, ratio O/E ≥ 0.6etl'écart entre les îlots adjacents est défini à 100 pb.

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

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

$$Obs_{CpG}/Exp_{Cpg}(P_i) = \frac{\frac{\#CpG}{CpG\ lengt\ h}}{\frac{\#C}{CpG\ lengt\ h} * \frac{\#G}{CpG\ lengt\ h}} \qquad (3.3)$$

Où #A, #T, #C et #G sont respectivement les nombres d'adénine (A), de thymine (T), de cytosine (C) et de guanine (G) dans la région prédite de l'île CpG à P_i. P_{min} est la position de départ du cluster moins 200, et P_{max} est la position de départ du cluster plus 200. #CpG représente le nombre de CpG dans la région prédite de l'île CpG àP_i [23].

Dans ce chapitre, nous allons présenter une étude comparative des méthodes de détection des ilots CpG que nous avons abordées dans le deuxième chapitre.

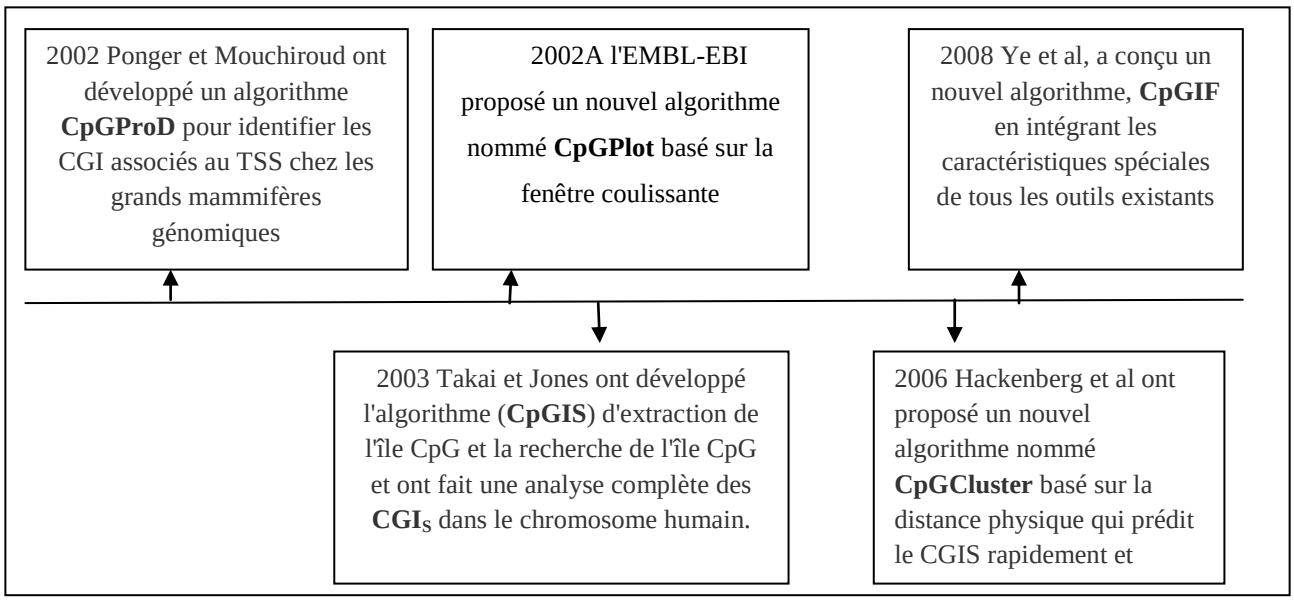

Figure 3.1 : La chronologie des outils de détection des CGI.

3.2 Méthodes étudiées

Notre étude est basée sur les méthodes : CpGProd, CpGPlot, CpGCluster, CpGIS et CpGIF. La figure 3.1 présentent plus de détail sur ces méthodes ainsi que la chronologie de leur ordre de présentation dans littérature.

3.3 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 Equations. (3.1-3.5) Ces cinq critères d'évaluation permettent de déterminer la supériorité d'un algorithme.

$$SN = \frac{TP}{TP+FN} \qquad (3.1)$$

$$SP = \frac{TN}{TN+EP} \qquad (3.2)$$

$$ACC = \frac{TP+TN}{TP+FP+TN+FN} \qquad (3.3)$$

$$PC = \frac{TP}{TP+FN+FP} \qquad (3.4)$$

$$CC = \frac{TP*TN-FP*FN}{\sqrt{(TP+FN)*(TP+FP)*(TN+FN)*(TN+FP)}} \qquad (3.5)$$

Où TP est un vrai positif indique que le nombre de base détectées dans la région de l'île CpG sont devrai îles CpG, FN est un faux négatif indique que le nombre de bases détectées dans la région de l'île CpG est incorrect, TN est un vrai négatif indique que le nombre de bases détectées dans la région de l'île non-CpG est correct et FP est un faux positif indique que certaines des bases détectées dans la région de l'île CpG sont des îles non CpG. Nous avons prédit les îles 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 îles CpG.

3.4 Données utilisées

Notre étude expérimentale a été menée sur des contigs des chromosomes humains. Nous avons utilisé les contigs : NT_113953.1, NT_113954.1, NT_113955.2, NT_113958.2 et NT_113952.1 du chromosomes 21 et NT_028395.3 du chromosome 22. Les séquences des contigs utilisés ont été téléchargés depuis la banque de données NCBI accessible via le lien : http://www.ncbi.nlm.nih.gov.

3.5 Résultats expérimentaux

La figure 3.1 présentent des informations sur es six contigs testés avec CpGProD : référence de la séquence, sa longueur, nombre d'îlots CpG trouvés dans la séquence, fréquence G+C de la séquence, rapport CpG o/e de la séquence et l'orientation (flèche noire) et pour chaque îlot CpG.

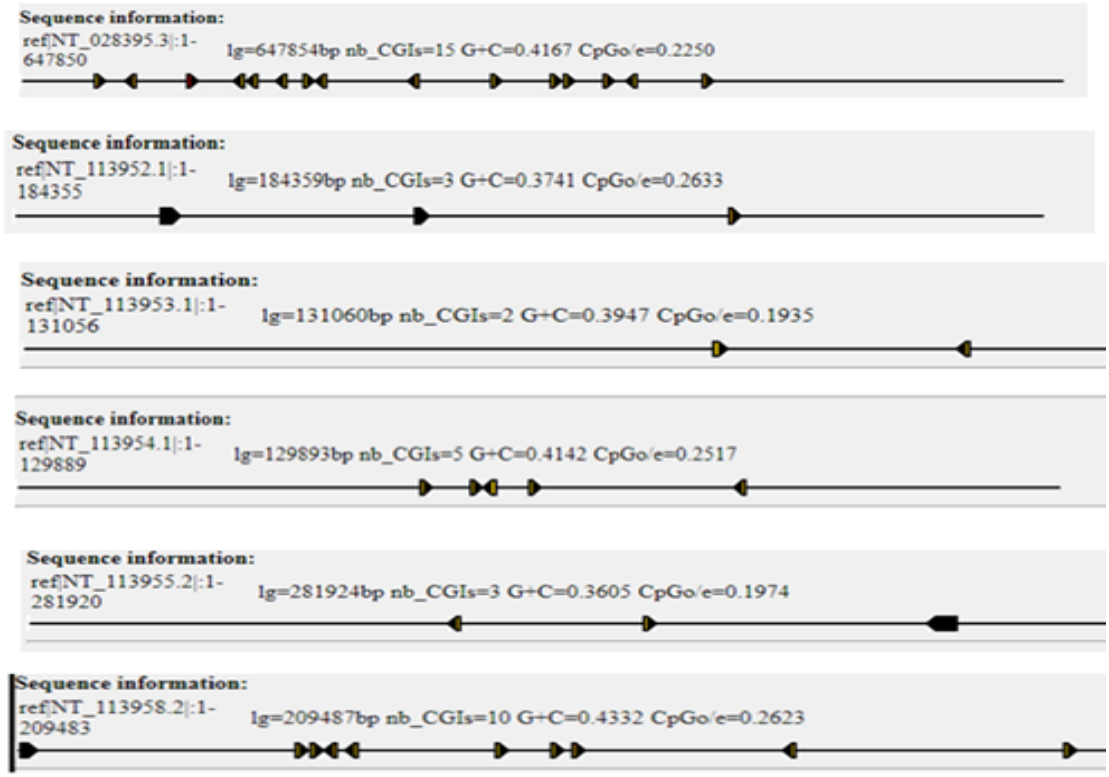

Figure 0.2: Des métriques sur les ilots CpG trouvés avec CpGProd pour les six contigs

Les tableaux de 3.1 à 3.6 détaillent les résultats obtenus avec CpGProd pour les six contigs utilisés. La première colonne représente l'indice de la séquence (une ligne par île CpG), la deuxième présente le nombre de l'ilot, la troisième et quatrième représentent la position début et fin de l'ilot, la cinquième, sixième et septième colonne représentent la longueur de l'ilot (à l'exclusion des nucléotides indéterminés), la fréquence CG et le ration o/e, respectivement.

Tableau 3.1 : Résultats obtenus avec le contig NT_028395.3

Nom	N°	Début	Fin	Longueur(pb)	G+C	CpG o/e
ref|NT_028395.3|:1-647850	1/15	47231	48258	1028	0.6041	0.6086
ref|NT_028395.3|:1-647850	2/15	72325	73293	969	0.5573	0.7587
ref|NT_028395.3|:1-647850	3/15	105278	108284	3007	0.6462	0.7240
ref|NT_028395.3|:1-647850	4/15	142303	143285	983	0.5972	0.5955
ref|NT_028395.3|:1-647850	5/15	150016	150520	505	0.6020	0.6130
ref|NT_028395.3|:1-647850	6/15	166943	167791	849	0.6620	0.5576
ref|NT_028395.3|:1-647850	7/15	177878	178646	769	0.5488	0.5879
ref|NT_028395.3|:1-647850	8/15	179113	179709	597	0.5578	0.5608
ref|NT_028395.3|:1-647850	9/15	237887	239164	1278	0.6017	0.8126
ref|NT_028395.3|:1-647850	10/15	285771	287161	1391	0.5162	0.6448
ref|NT_028395.3|:1-647850	11/15	324100	325043	944	0.5837	0.5376
ref|NT_028395.3|:1-647850	12/15	329540	330097	558	0.5072	0.5966
ref|NT_028395.3|:1-647850	13/15	353980	355107	1128	0.5426	0.7085
ref|NT_028395.3|:1-647850	14/15	373574	374475	902	0.5089	0.6033
ref|NT_028395.3|:1-647850	15/15	416398	418439	2042	0.6224	0.6376

Tableau 0.2 : Résultats obtenus avec le contig NT_113952.*1*

Nom	N°	Début	Fin	Longueur(pb)	G+C	CpG o/e
ref|NT_113952.1|:1184355	1/3	25884	28634	2751	0.6401	1.1457
ref|NT_113952.1|:1184355	2/3	71976	73754	1779	0.5829	1.2197
ref|NT_113952.1|:1184355	33	128971	130014	1044	0.6801	0.6267

Tableau 3.3 : Résultats obtenus avec le contig NT_113953.1

Nom	N°	Début	Fin	Longueur (pb)	G+C	CpG o/e
ref|NT_113953.1|:1131056	1/2	78564	79689	1126	0.5373	0.5939
ref|NT_113953.1|:1131056	2/2	108140	108711	572	0.5087	0.5693

Tableau 0.4: Résultats obtenus avec le contig NT_113954.1

Nom	N°	Début	Fin	Longueur(pb)	G+C	CpG o/e
ref|NT_113954.1|:1129889	1/5	49910	50847	938	0.5149	0.5335
ref|NT_113954.1|:1129889	2/5	56613	57353	741	0.5290	0.6049
ref|NT_113954.1|:1129889	3/5	58259	59325	1067	0.6148	0.6194
ref|NT_113954.1|:1129889	4/5	63206	64023	818	0.5831	0.5611
ref|NT_113954.1|:1129889	5/5	90319	90854	536	0.5037	0.5898

Tableau 0.1: Résultats obtenus avec le contig NT_113955.2

Nom	N°	Début	Fin	Longueur(pb)	G+C	CpG o/e
ref|NT_113955.2|:1-281920	1/3	109141	109744	604	0.4983	0.6141
ref|NT_113955.2|:1-281920	2/3	155994	156685	692	0.4971	0.7522
ref|NT_113955.2|:1-281920	3/3	230462	237184	6723	0.6192	0.6255

Tableau 0.2: Résultats obtenus avec le contig NT_113958.2

Nom	N°	Début	Fin	Longueur(pb)	G+C	CpG o/e
ref|NT_113958.2|:1-209483	1/10	1	2403	2399	0.7228	1.0901
ref|NT_113958.2|:1-209483	2/10	52032	52659	628	0.6274	0.5701
ref|NT_113958.2|:1-209483	3/10	52995	54109	1115	0.6135	0.5743
ref|NT_113958.2|:1-209483	4/10	56038	57317	1280	0.5938	0.6779
ref|NT_113958.2|:1-209483	5/10	60460	61061	602	0.5465	0.5858
ref|NT_113958.2|:1-209483	6/10	86475	87183	709	0.6417	0.5617
ref|NT_113958.2|:1-209483	7/10	96855	97587	733	0.5593	0.6641
ref|NT_113958.2|:1-209483	8/10	100891	102176	1286	0.5933	0.6462
ref|NT_113958.2|:1-209483	9/10	142500	143649	1150	0.5643	0.5947
ref|NT_113958.2|:1-209483	10/10	193168	193933	766	0.4765	0.5759

Les figures de 3.2 à 3.7shématisent les résultats (Ration o/e, ilots détectés et fréquence CG) d'utilisation de CpGPlot avec les six contigs.

Séquence1 : NT_028395.3

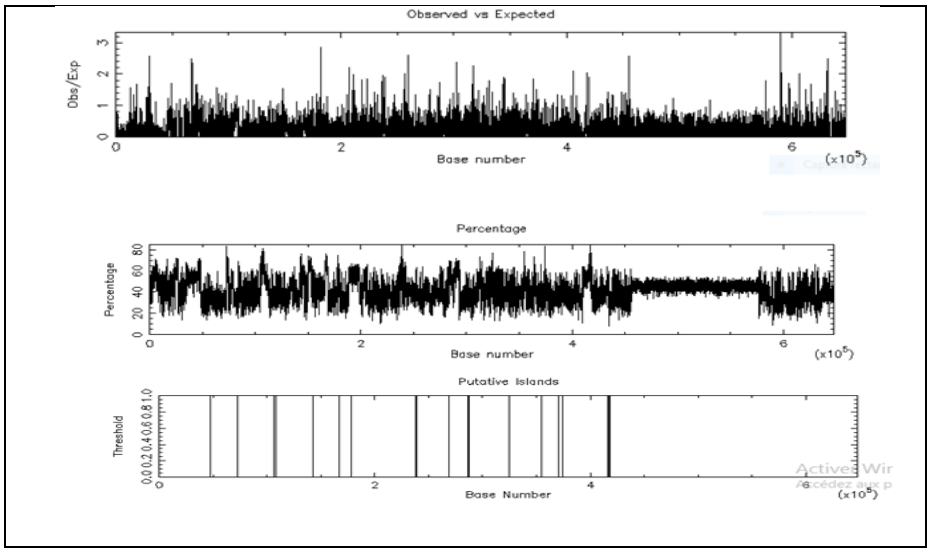

Figure 0.3: Obs/Exp pourcentage, putative ilots séquence 1

Séquence2 :NT_113952.1

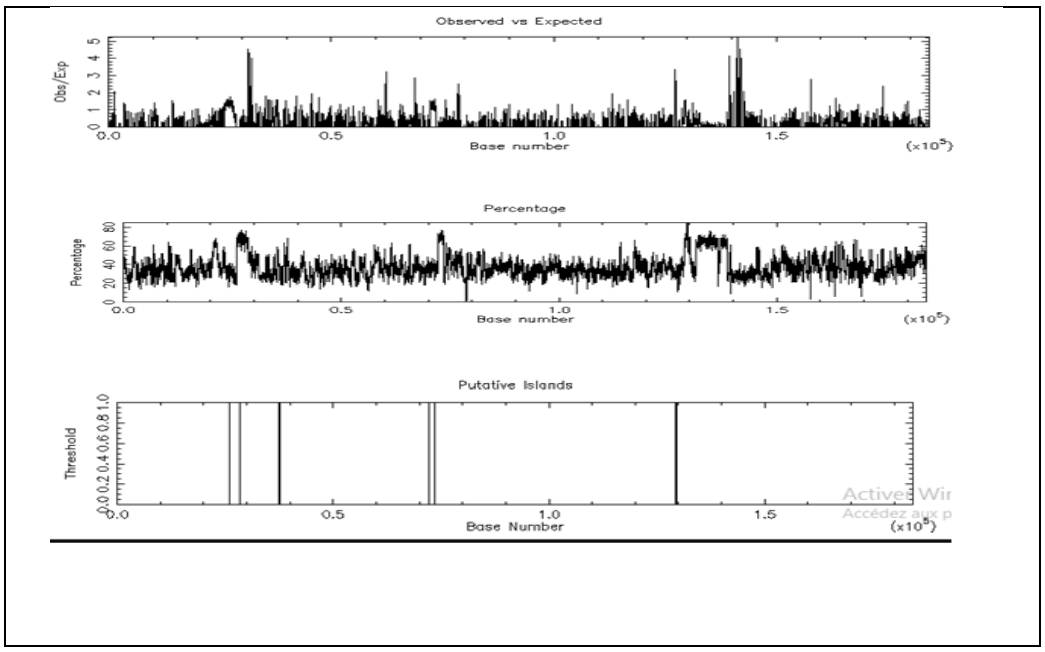

Figure 0.4: putative ilots, Obs/Exp, pourcentage séquence 2

Séquence3 :NT_113953.1

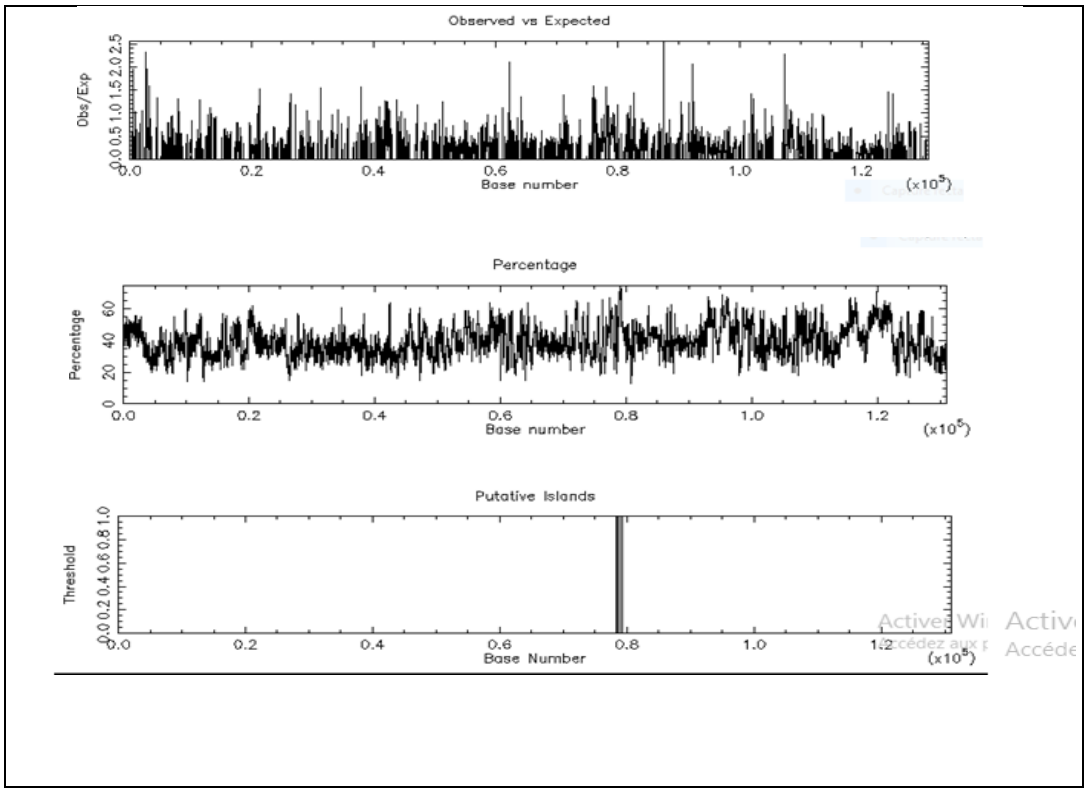

Figure 0.5: putative ilots, Obs/Exp, pourcentage séquence 3

Séquence4 : NT_113954.1

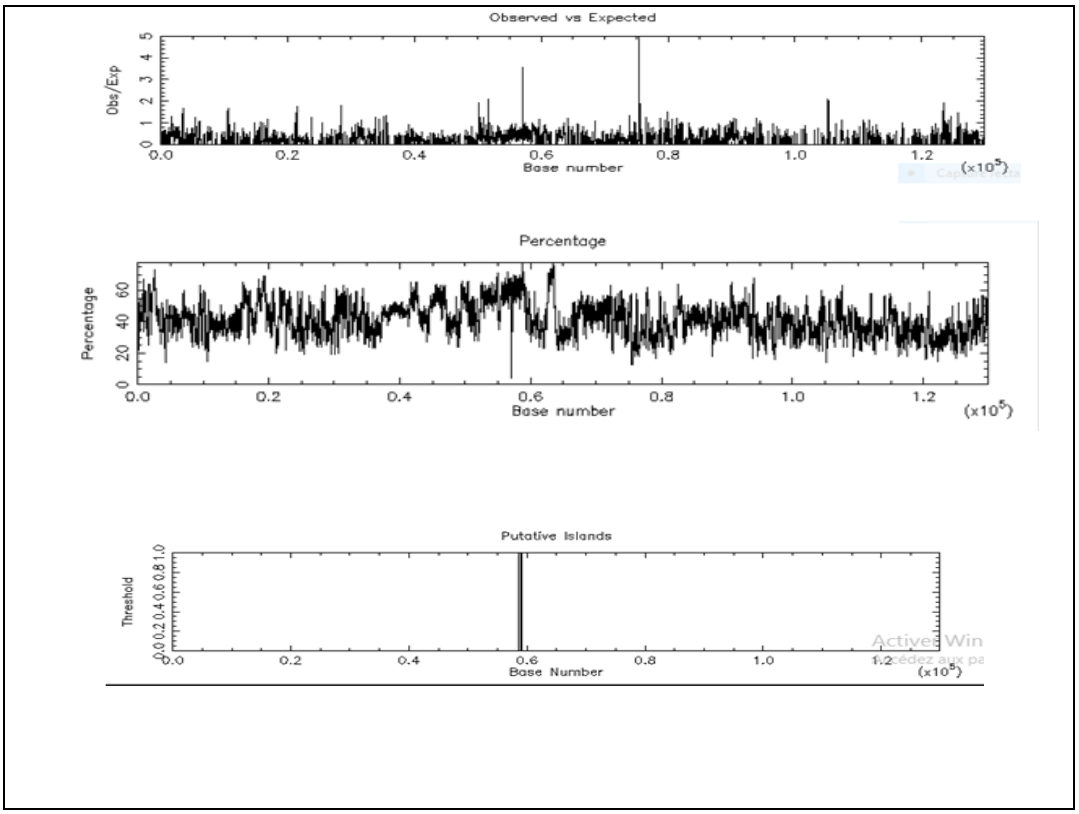

Figure 0.6: putative ilots, Obs/Exp, pourcentage séquence 4

Séquence5 : NT_113955.2

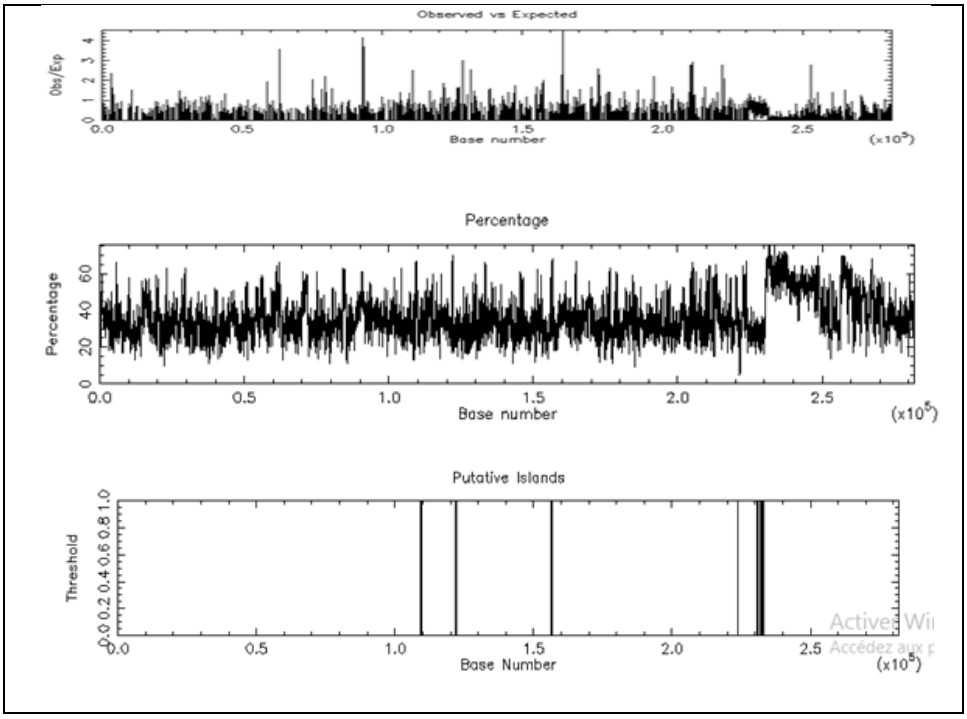

Figure 0.7: putative ilots, Obs/Exp, pourcentage séquence 5

Séquence6 : NT_113958.2

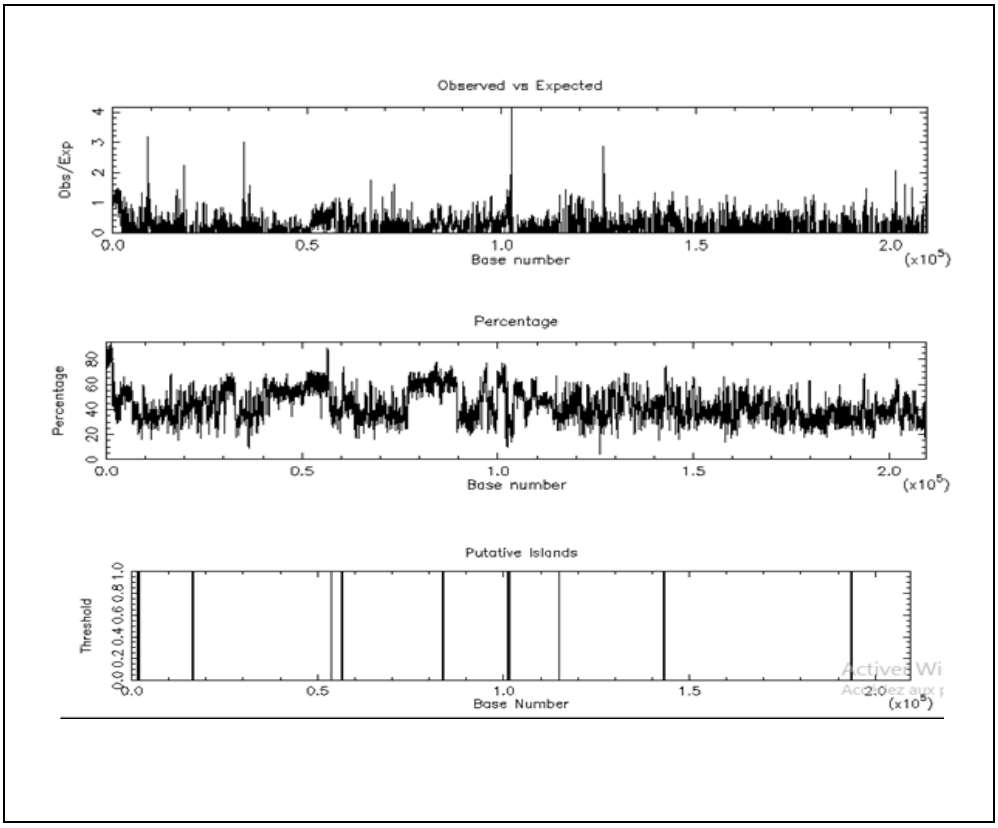

Figure 0.8: putative ilots, Obs/Exp, pourcentage séquence 6

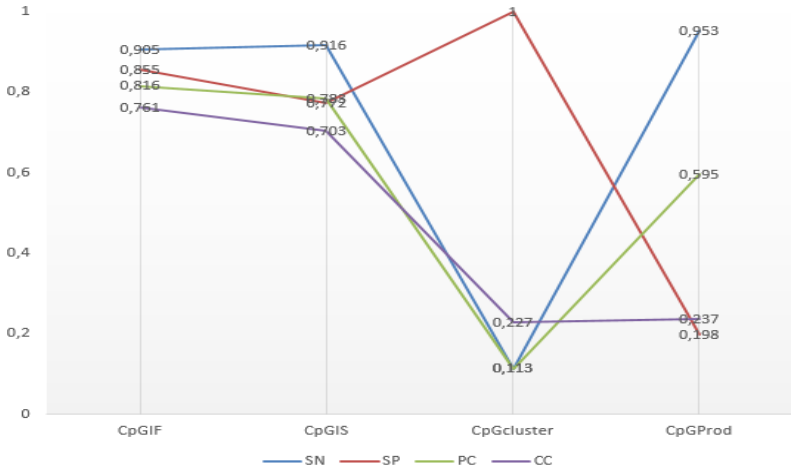

Figure 3.9 : Résultats de performance des méthodes utilisées

La figure 3.9 présente un graphe qui montre les résultats de la comparaison des mesures (SN,SP,PC,CC) obtenues par les méthodes CpGIF, CpGcluster, CpGProD et CpGIS appliquées sur les séquences présentées dans[20].La figure confirme la lecture des résultats trouvés dans le tableau 3.8. Elle montre que :

• CpGIS présente toujours de bons résultats en SN, PC et CC, Et de bons résultats en SP.

• CpGProd et CpG cluster présentent de bons résultats en SN et SP, respectivement.

• CpGIF présentent les meilleurs résultats en PC, CC et de bons résultats en SP.

3.6 Comparaison basée sur les caractéristiques

Tableau 0.3:Comparaison basée sur les caractéristiques des méthodes étudiées

Nom dela Méthode	**CpGIF**	**CpG Cluster**	**CpGProD**	**CpGIS**	**CpGPlot**
Année de publication	2008	2006	2001	2003	2002
Paramètre données	densité par défaut 0,10 limite de densité plus tardive réduite de 0,9 à 0,5	paramètre par défaut	TSS	GGF	GGF
Performance	Corrélation et PC élevés	AC, SP élevés FP inférieures	sensibilité supérieure et spécificité supérieure	organismes supérieurs dans certaines situations	d'excellents résultats SP
Classe d'algorithme	Méthode basée sur la densité	Méthode basée sur la distance	Méthode basées sur les fenêtres coulissantes	Méthode basées sur les fenêtres coulissantes	Méthode basées sur les fenêtres coulissantes
Génome de test	humain	humain et souris	humain et souris	humain	humain
Paramètre de performance	SN , SP ,Ppv ,pc	ACC,SP,FP, SN	SN,SP, TP	%GC, ObsCpG/ExpCpG	ACC, SP, FP,SN

3.7 Comparaison basée sur les performances

Le tableau 3.8 montre une comparaison de CpGIF, CpGProD, CpG Cluster et CpGIS pour l'identification des CGI. Cette comparaison est basée sur les 3 critères : SN, SP, ACC. La première colonne du tableau représente d'identifiant du contig, la deuxième représente les mesures de performance, les autres colonnes représentent les résultats obtenus par les méthodes CpGPlot, CpG Cluster, CpGProd et CpGIS avec chaque critère.

Tableau 0.4:Comparaison des différentes méthodes de prédiction des îlots CpG.

Contig	Mesure	CpGPlot	CpGCluster	CpGProd	CpGIS
NT_113952.1	SN	56.43	50.46	58.07	83.98
	SP	100.0	99.75	99.50	99.05
	ACC	98.09	97.78	97.69	98.39
	PC	56.42	49.92	52.36	69.59
	CC	74.38	69.41	68.83	81.25
NT_113955.2	SN	47.19	67.15	68.51	85.12
	SP	100.0	99.72	99.63	99.30
	ACC	98.08	98.54	98.50	98.79
	PC	47.14	62.47	62.35	71.78
	CC	67.94	77.03	76.65	82.96
NT_113958.2	SN	51.29	27.16	46.41	82.19
	SP	99.99	99.94	98.93	98.26
	ACC	96.90	95.32	95.60	97.24
	PC	51.24	26.92	40.10	65.36
	CC	70.38	49.96	56.80	77.63
NT_113953.1	SN	22.80	57.32	29.79	74.05
	SP	100.0	99.74	99.56	98.11
	ACC	97.76	98.51	97.53	53.23
	PC	22.80	52.74	25.96	53.23
	CC	47.21	69.89	43.61	68.64
NT_113954.1	SN	31.24	29.86	52.01	76.31
	SP	100.0	99.46	98.72	97.62
	ACC	97.47	96.90	97.00	96.83
	PC	31.24	26.19	38.94	47.05
	CC	55.17	43.81	54.68	63.29
NT_028395.3	SN	27.11	44.89	54.18	76.68
	SP	100.0	99.47	99.45	98.93
	ACC	97.98	97.53	98.19	98.14
	PC	27.10	39.26	45.36	59.36
	CC	51.51	57.21	62.26	73.57

Comme il est illustré dans le tableau 3.8:

- CpGPlot présente d'excellents résultats SP avec les six séquences et de bons résultats ACC et CC avec lesquels il se classe dans la 2ème position avec 3 contigs.
- CpGIS fonctionne mieux que les méthodes CpGPlot, CpGcluster et CpGProD. Il présente les meilleurs résultats en termes de : SN (avec les 6 contigs), ACC (avec 3 contigs), Pc (avec les 6 contigs) et CC (avec 5 contigs).
- CpGProd présente de bons résultats avec SN. Il occupe la 2ème position avec 3 contigs.
- CpG cluster présente de bons résultats avec SP. Il occupe la 2ème position avec 5 contigs.

Les figures de 3.10 à 3.14 présentent les résultats de détection des îlots CpG pour les séquences chromosomiques 21 et 22enutilisantCpGIS, CpG cluster ,CpGProD, CpGPlot, CpGIF.

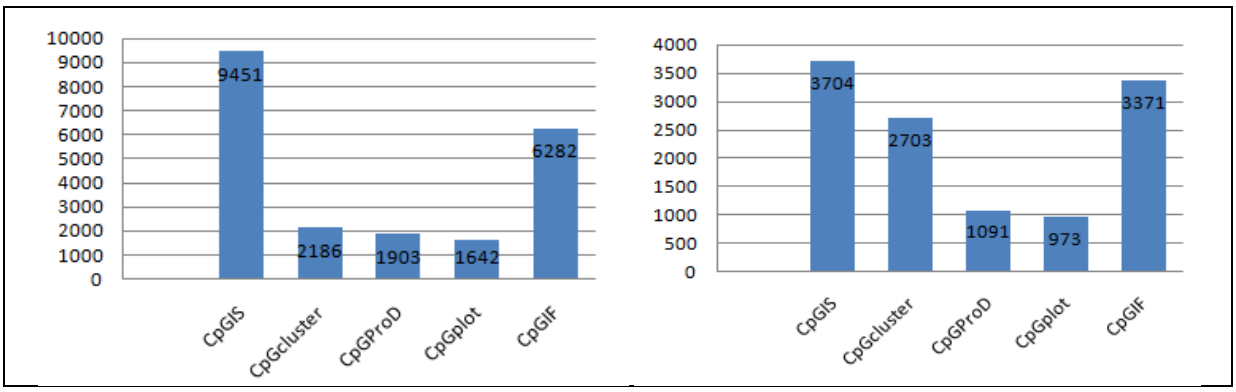

Figure 0.10: Nombre des ilots CpG détectés avec chromosome 21et chromosome 22, respectivement

La figure 3.10 montre qu'en termes de nombre de CGI:

- CpGIS performe mieux que les autres méthodes avec les deux chromosomes.
- Les résultats de CpGIS et CpGIF sont très proches avec les deux chromosomes.

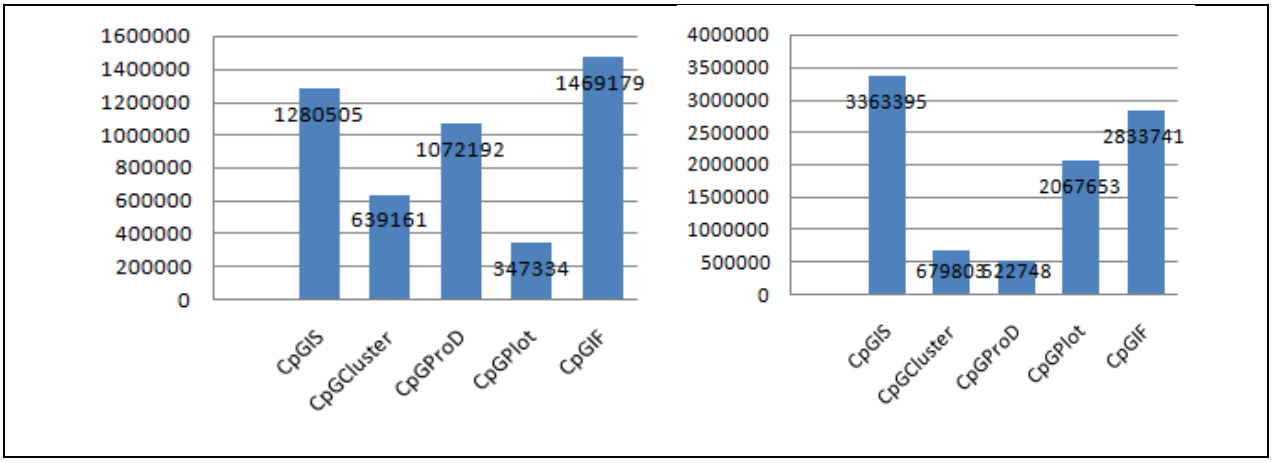

Figure 0.11: Longueur totale de s'ilots CpG détectés avec chromosome 21et chromosome 22, respectivement

La figure 3.11 montre qu'en termes de longueur totale des CGI, également CpGIS et CpGIF trouvent les meilleurs résultats avec les deux chromosomes.

La figure 3.11 montre qu'avec la longueur moyenne des CGI détectés :

- CpGProd trouve d'excellents résultats.
- Les résultats de CpGPlot et CpGIF sont proches.

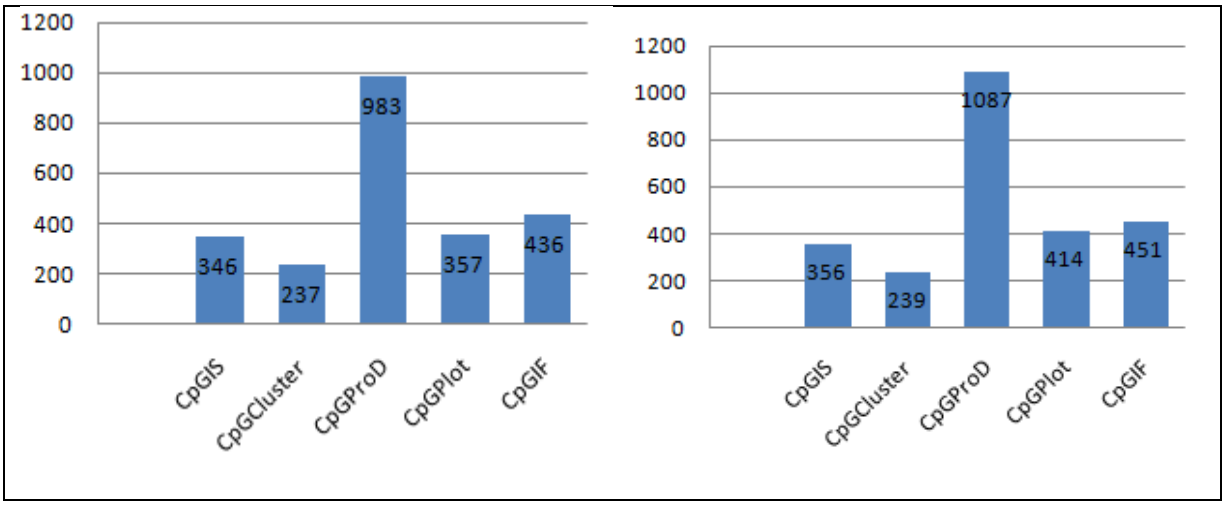

Figure 0.12: Longueur moyenne des ilots CpG détectés avec chromosome 21et chromosome 22, respectivement

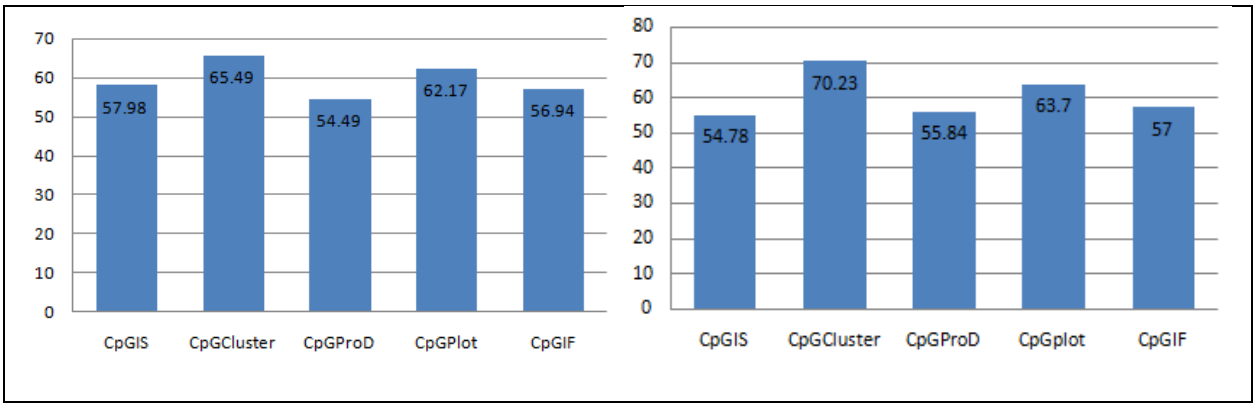

Figure 0.13: Contenu CG des ilots détectés avec chromosome 21et chromosome 22, respectivement

La figure 3.13 montre que :

- Les meilleurs résultats sont obtenus avec CpG cluster.
- Le pourcentage CG détecté par les cinq méthodes est presque similaire chez des deux chromosomes 21 et 22.

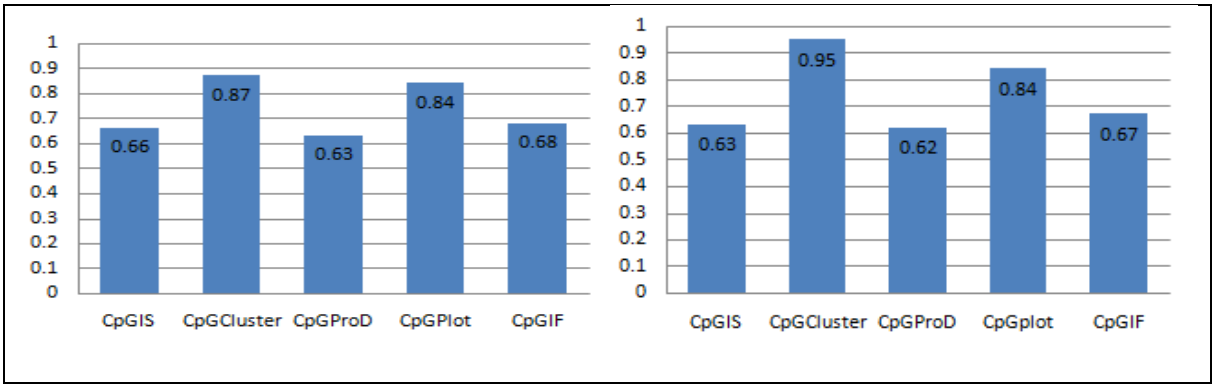

Figure 0.14: Ratio CpG o/e détecté avecchromosome 21 et 22, respectivement

L'observation de la figure 3.14 montre que :

- CpGcluster performe mieux que les autres méthodes.
- Les résultats de CpGPlot et sont proches de celle de CpGcluster
- Les performances de CpGIS, CpGIF et CpGProd sont très proches.

3.8 Conclusion

Les méthodes actuelles de détection des îlots CpG sont principalement basées sur les critères proposés par Gardiner-Garden et Frommer (GGF). Dans ce chapitre nous nous sommes basés sur un ensemble de critères définis par des équations tirées de la littérature afin d'établir une étude comparative basées sur les caractéristiques et sur les performances de chaque méthode. Nous avons appliqué ces méthodes sur six contigs du chromosome 21 et 22

de l’être humain. Cette étude a montré que chaque méthode est caractérisée par un algorithme spécifique, des points forts et a des lacunes.

Conclusion Générale
Et Perspectives

Conclusion et Perspectives

La méthylation de l'ADN est un événement héréditaire important dans les marques épigénétiques du génome associé aux événements de développement et à la régulation des gènes. L'état de méthylation des CGIs fournit des indications fonctionnelles pour déterminer les rôles régulateurs dans la transcription, car les CpGs méthylés sont exceptionnellement associés à la transcription. La prédiction computationnelle de la méthylation peut stimuler le profilage de la méthylation à l'échelle du génome et identifier les principales caractéristiques sous-jacentes de divers patrons de méthylation. Il existe de nombreux outils de calcul disponibles pour la détection de CGI basés sur les distances physiques, les propriétés physiochimiques, le regroupement, les fenêtres coulissantes et les motifs. Ces algorithmes de calcul fonctionnent selon différents principes et fournissent une grande précision, sensibilité ou spécificité, mais uniquement pour des ensembles de données et des paramètres spécifiques. Les analyses comparatives des outils permettent d'identifier les caractéristiques et les limites permettant ainsi de prioriser les algorithmes de détection des CGI pour les épigénomes cibles. Il a été suggéré que de nouveaux algorithmes pourraient être développés en incorporant les principes précédents, en améliorant les algorithmes traditionnels, en modifiant les paramètres et les avancées créatives.

Il faut également disposer d'outils de calcul avancés et plus efficaces pour la détection rapide et précise des CGI dans les génomes entiers. En général, la capacité de prédire par calcul les CGI et l'état de la méthylation est très prometteuse pour accélérer les découvertes en épigénétique et en réglementation des gènes. L'identification précise des CGI à partir de données de séquençage à haut débit au moyen d'approches computationnelles pourrait mener à des percées significatives dans les mécanismes épigénétiques et les troubles neurologiques associés.

Comme perspectives pour ce travail, plusieurs pistes méritent d’être explorées, nous pouvons :

- Etablir un état de l’art sur l’utilisation des méthodes d’intelligence artificielle pour la détection des ilots CpG dans le génome de différentes espèces.
- Utiliser des algorithmes d’apprentissage profond pour identifier les îlots CpG dans le génome humain.
- Envisager des hybridations entre plusieurs méthodes et les tester sur plusieurs génomes

Références

Références

[1] I.N.Cancer « La génomique, une révolution dans la recherche et le traitement des cancers - Comprendre la recherche ». https://www.e-cancer.fr/Comprendre-prevenir-depister/Comprendre-la-recherche/La-revolution-de-la-genomique (consulté le mai 15, 2021).

[2] K.Ouellet« 28688.pdf ». Consulté le: juin 07, 2021. [En ligne]. Disponible sur: https://corpus.ulaval.ca/jspui/bitstream/20.500.11794/23492/1/28688.pdf

[3] R. Holliday et J. E. Pugh, « DNA modification mechanisms and gene activity during development », *Science*, vol. 187, n° 4173, p. 226-232, janv. 1975.

[4] K.D.Robertson, P.A.Jones« DNA methylation: past, present and future directions | Carcinogenesis | Oxford Academic ». https://academic.oup.com/carcin/article/21/3/461/2365668 (consulté le juin 08, 2021).

[5] F.R.G« Régulation de l'expression des gènes ». https://sciencesnaturelles.ch/fr/id/5irpQ (consulté le août 01, 2021).

[6] M. Esteller« Epigenetic gene silencing in cancer: the DNA hypermethylome | Human Molecular Genetics | Oxford Academic ». https://academic.oup.com/hmg/article/16/R1/R50/2355975?login=true (consulté le juin 08, 2021).

[7] A.P.Feinberg « The Key Role of Epigenetics in Human Disease Prevention and Mitigation », *N. Engl. J. Med.*, vol. 378, n° 14, p. 1323-1334, avr. 2018, doi: 10.1056/NEJMra1402513.

[8] L.S.Kristensen, M.B.Treppendahl, et K.Grønbæk, « Analysis of epigenetic modifications of DNA in human cells », *Curr. Protoc. Hum. Genet.*, vol. Chapter 20, p. Unit20.2, avr. 2013, doi: 10.1002/0471142905.hg2002s77.

[9] C.G.Renard, « Évolution des îlots CpG chez les primates », These de doctorat, Lyon 1, 2009. Consulté le: juin 07, 2021. [En ligne]. Disponible sur: http://www.theses.fr/2009LYO10162

[10] D.N.Cooper et S. Gerber-Huber, « DNA methylation and CpG suppression », *Cell Differ.*, vol. 17, n° 3, p. 199-205, sept. 1985, doi: 10.1016/0045-6039(85)90488-9.

[11] M.G.Garden et M.Frommer, « CpG islands in vertebrate genomes », *J. Mol. Biol.*, vol. 196, n° 2, p. 261-282, juill. 1987, doi: 10.1016/0022-2836(87)90689-9.

[12] M.L.Tykocinski et E.E.Max, « CG dinucleotide clusters in MHC genes and in 5' demethylated genes. », *Nucleic Acids Res.*, vol. 12, n° 10, p. 4385-4396, mai 1984.

[13] G.G.M ET F.M « CpG islands in vertebrate genomes », *J. Mol. Biol.*, vol. 196, n° 2, juill. 1987, doi: 10.1016/0022-2836(87)90689-9.

[14]D.Takai et P.A.Jones « Comprehensive analysis of CpG islands in human chromosomes 21 and 22 - PubMed ». https://pubmed.ncbi.nlm.nih.gov/11891299/ (consulté le juin 30, 2021).

[15] EMBOSS« EMBOSS Cpgplot< Sequence Statistics < EMBL-EBI ». https://www.ebi.ac.uk/Tools/seqstats/emboss_cpgplot/ (consulté le juin 13, 2021).

[16] D. Takai et P. A. Jones, « The CpG island searcher: a new WWW resource », *In Silico Biol.*, vol. 3, n° 3, p. 235-240, 2003.

[17] L. Ponger et D. Mouchiroud « CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences », *Bioinforma. Oxf. Engl.*, vol. 18, n° 4, p. 631-633, avr. 2002, doi: 10.1093/bioinformatics/18.4.631.

[18]R.A.Tahir , D.Zhzng, A.Nazir et H.Qing « A review of computational algorithms for CpG islands detection.pdf ».

[19] M. Hackenberg, C. Previti, P. L. Luque-Escamilla, P. Carpena, J. Martínez-Aroza, et J. L.Oliver « CpGcluster: a distance-based algorithm for CpG-island detection », *BMC Bioinformatics*, vol. 7, n° 1, p. 446, oct. 2006, doi: 10.1186/1471-2105-7-446.

[20] M. Hackenberg, C. Previti, P. L. Luque-Escamilla, P. Carpena, J. Martínez-Aroza, et J. L. Oliver, « CpGcluster: a distance-based algorithm for CpG-island detection », *BMC Bioinformatics*, vol. 7, n° 1, p. 446, oct. 2006, doi: 10.1186/1471-2105-7-446.

[21] Y. Sujuan, A. Asaithambi, et Y. Liu, « CpGIF: an algorithm for the identification of CpG islands », *Bioinformation*, vol. 2, n° 8, p. 335-338, mai 2008.

[22] I.E.E.E.Xplore « A tutorial on hidden Markov models and selected applications in speech recognition ». https://ieeexplore.ieee.org/document/18626 (consulté le juin 30, 2021).

[23] C.H. Yang, Y.D. Lin, Y.C. Chiang, et L.Y. Chuang, « A Hybrid Approach for CpG Island Detection in the Human Genome », *PLOS ONE*, vol. 11, n° 1, p. e0144748, janv. 2016, doi: 10.1371/journal.pone.0144748.

Présenté par : ZAROUR Amira
CHELIA Djihad

Date de soutenance : 22/09/2021

Résumé

Les îlots CpG sont généralement connues sous le nom de régions de régulation épigénétique, en fonction des modifications de l'histone, de la méthylation et de l'activité du promoteur. La cartographie exacte de la méthylation de l'ADN dans les îles CpG est nécessaire pour comprendre les diverses fonctions biologiques. Cependant, l'identification précise des îlots de CpG à partir du génome entier par des approches expérimentales et computationnelles reste difficile. De nombreuses méthodes de calcul sont développées pour détecter les régions enrichies de CpG, de manière efficace, afin de réduire le temps et le coût des expériences. Ici, nous passons en revue quelques-unes des dernières méthodes de détection de CpG computationnelle qui utilisent le clustering, les modèles et les paramètres de distance physique comme pour la détection d'île de CpG. Les analyses comparatives des méthodes reposant sur différents principes et paramètres permettent de prioriser les algorithmes pour des ensembles de données spécifiques CpG associés afin d'obtenir une plus grande précision et sensibilité. Un certain nombre d'outils de calcul basés sur la fenêtre, les algorithmes basés sur la densité et la distance / longueur sont appliqués sur le génome humains pour la détection précise de CpG.Plusieurs algorithmes ont été développés pour l'identification CGI et ont été appliqués dans de nombreuses études. Ils peuvent être classés en deux groupes : les algorithmes traditionnels qui sont basés sur trois paramètres de séquence (longueur, contenu GC, et rapport de la CpGs observée sur la CpGs attendue (CpG_{Obs}/CpG_{Exp})) et les algorithmes basés sur la propriété statistique dans une séquence sans imposer les trois critères dans les algorithmes traditionnels.

Mots-clés : Bioinformatique ; épigénétique ; île CpG ; Génome humain ; algorithmes de détection.

Membre du jury :

Présidant : Dr. Ines Bellil

Encadreur : Dr. Amira GHERBOUDJ

Examinateur : Dr. Mohamed Skander DAAS