دراسة مقارنة لخصائص جزر الـــCpG المكتشفة باستخدام أداتي newCpGReport وCpGCluster*

م. مجد الصباغ**

أ.م. د. علي صقر**** أ.م. د. رشا مسعود***

الملخص

تؤدي جزر الـــ CpG دوراً مهماً في تنظيم عملية نسخ الجينات (Gene Transcription Regulation). ويعود ذلك لتراكب عدد كبير من هذه الجزر مع محفزات نسخ الجينات (Promoters). إذْ إنَّ مثْيلة (Methylation) الجزر الموجودة في منطقة المحفزات يمكن أن يعطل عملية نسخ الجين المسؤول عنه هذا المحفز. ومن أهم هذه الحالات تعطل الجينات المثبطة لتطور الأورام السرطانية (Cancerous Tumours)، فقد بينت العديد من الدراسات أن مثيلة جزر الـــCpG هو من العلامـــات المهمة على وجود السرطان أو إمكانية تطوره. وانـــطلاقاً من أهمية جزر CpG، فقد طُوِّرَتْ عدة خوارزميات للكشف عنها في سلاسل النكليوتيدات. وتقسم هذه الخوارزميات بشكل عام إلى نوعين رئيسين وهما: الخوارزميات التي تعتمد على المسافة (Distance-based)، ومن أهمها خوارزمية (CpGCluster)، والخوارزميـــات التـــي تعتـــمد على النافذة المنزلقة (Sliding-window) ومنها (newCpGReport). إلا أن هناك بعض الاختلافات في نتائج تطبيق كل من هاتين الخوارزميتين، لذلك فإن هدف هذه الدراسة هو مقارنة أداء كل من الخوارزميتين السابقتين (newCpGReport وCpGCluster).

كُشِفَ عن جزر الـــCpG الموجودة على الصبغي 22 عند الإنسان من خلال تطبيق كلتا الخوارزميتين على هذا الصبغي. وقد أظهرت النتائج وجود اختلاف واضح بين عدد الجزر المكتشفة في كلتا الخوارزميتين، فضلاً عن وجود تباين في طول الجزر المكتشفة. كما أظهرت النتائج وجود تقاطع بين نحو 60% من الجزر المكتشفة في كلتا الطريقتين.دُرِسَ في هذا البحث تأثير قيم بارامترات جزر الـــCpG وهي الطول ومحتوى C+G (C+G content) ونسبة الملاحظ/المتوقع (Observed/Expected ratio) في عدد الجزر المكتشفة. وتبيّن أن لعتبة الطول تأثيراً كبيراً في عدد الجزر المكتشفة باستخدام newCpGReport، على عكس خوارزمية CpGCluster التي لا يتأثر أداؤها بعتبة الطول. كما دُرِسَ تأثير جعل الجزر المكتشفة باستخدام newCpGReport تبدأ وتنتهي بزوج النكليوتيدات CpG، وكانت النتيجة زيادة قيم كل من بارامتري المحتوى ونسبة الملاحظ/المتوقع لنسبة كبيرة من الجزر، مع الأخذ بالحسبان انخفاض طول نحو 25% من الجزر إلى ما دون 200 نكليوتيد.

الكلمات المفتاحية: جزر **CpG**، المثيلة، الجينات، الصبغيات، الأورام السرطانية، محتوى **C+G**، **newCpGReport**، **CpGCluster**.

*أعد هذا البحث في سياق رسالة الماجستير للمهندسة مجد الصباغ بإشراف الدكتورة المهندسة رشا مسعود و مشاركة الدكتور علي صقر، قسم الهندسة الطبية، كلية الهندسة الكهربائية و الميكانيكية، جامعة دمشق.

**قسم الهندسة الطبية، كلية الهندسة الكهربائية و الميكانيكية، جامعة دمشق.

****قسم التشريح والنسج والجنين، كلية الطب البشري، جامعة دمشق.

*** قسم الهندسة الطبية، كلية الهندسة الكهربائية و الميكانيكية، جامعة دمشق.

المقدمة:

نتيجة لتراكب جزر الـــCpG مع عدد كبير من محفزات عملية نسخ سلسلة الـــDNA في الجينات؛ فقد اِسْتُخْدِمَتْ في تحديد مواقع هذه المحفزات. يؤدي حدوث المثيلة لجزر الـــCpG الموجودة في منطقة المحفزات إلى تعطيل عملية نسخ الجين المسؤول عنه هذا المحفز. ومن ثَمَّ توقف إنتاج البروتينات المرتبطة بهذا الجين، كما في حالات تعطل الجينات المثبطة لتطور الأورام السرطانية. وقد بينت العديد من الدراسات أن مثيلة جزر الـــCpG هو من العلامات المهمة على وجود السرطان وإمكانية تطوره. وسُجِّلَ وجود علاقة بين المثيلة للجزر وتعطيل عمل أكثر من 100 جين في العديد من أنواع السرطانات [1][2][3].

يوجد حالياً العديد من الخوارزميات للكشف عن جزر الـــCpG في سلاسل النكليوتيدات. تعاني هذه الخوارزميات من بعض المشكلات التي تؤدي إلى اختلاف النتائج فيما بينها وعدم قدرتها على الكشف الدقيق في الحالات جميعها. كما أن الحاجة إلى التدخل البشري الكبير في تحديد بارامترات البحث يعدُّ من المشكلات التي أدت إلى التوجه نحو استخدام تقنيات الذكاء الصنعي (Artificial Intelligence) وتعليم الآلة (Machine Learning) لتطوير خوارزميات كشف أدق وأسرع من الخوارزميات السابقة [5] [6] [7] [8] [9] [10] [14] [15].

تنقسم الخوارزميات الحالية إلى نوعين:1) خوارزميات النافذة المنزلقة: التي تستخدم نافذة بطول محدد، تتحرك على سلسلة الـــنكليوتيدات وتقيس مجموعة بارامترات ضمن النافذة لتقييم كونها جزيرة CpG أو لا. بناءً على نتيجة التقييم تُحَرَّكُ النافذة بخطوة محددة وتُكَرَّرُ العملية حتى نهاية السلسلة المدروسة[5][6][14][15].2) خوارزميات تعتمد على المسافة الفيزيائية بين أزواج الـــCpG على السلسلة، ومن ثَمَّ لا تعتمد على العتبات التقليدية المحددة للجزر[7].

مما سبق كلّه؛ تأتي أهمية بنية نظام الكشف عن جزر الـــCpG في تحسين عملية الكشف مقارنة بالخوارزميات السابقة. إِذْ يؤدي التحديد الدقيق لهذه الجزر دوراً مهماً في العديد من الدراسات الجينية والسرطانية، من حيث تحديد مواقع محفزات النسخ ودراسة عمليات المَثيلة وارتباطها بالتغيرات الجينية المختلفة.

هَدَفَ هذا البحث إلى دراسة مجموعة خصائص لجزر الـــCpGالمكتشفة باستخدام خوارزميتين، إحداهما تعتمد مبدأ النافذة المنزلقة (newCpGReport)، والأخرى تعتمد على المسافة (CpGCluster). يأتي ذلك كمحاولة لبناء صورة واضحة عن الاختلاف والتقارب بين هذه الخوارزميات من حيث خصائص الجزر التي تكشفها، مما قد يسهم في بناء نظام كشف ذي أداء أفضل، ويحاول جمع محاسن الخوارزميات الحالية وتجاوز مشكلاتها. تفيد هذه الدراسة من الدراسات المنشورة عن المقارنة بين خوارزميات الكشف المختلفة [11][12] لتضيف إليها نتائج جديدة تدعم هدف تحسين عملية الكشف. هذه النتائج تبيّن التشابه بين الخوارزميات المعتمدة مبدأ النافذة المنزلقة من حيث الأداء، وذلك من خلال دراسة نتائج إحدى هذه الأدوات وهي newCpGReport ومقارنتها بنتائج CpGCluster. كما تبيّن الدراسة أن جعل الجزر المكتشفة باستخدام

newCpGReport تبدأ وتنتهي بـــCG يؤدي إلى زيادة قيم بارامتري المحتوى ونسبة الملاحظ/المتوقع لغالبية هذه الجزر.

ستجري الدراسة وفق ثلاثة محاور:

1) دراسة طول جزر الـــCpG المكتشفة وعددها على الصبغي 22 عند الإنسان باستخدام الأداتين المذكورتين، وذلك عند قيم مختلفة لعتبات البارمترات التقليدية المعرفة للجزر.

2) دراسة التقاطع بين الجزر المكتشفة باستخدام الأداتين المذكورتين.

3) دراسة تأثير جعل الجزر المكتشفة باستخدام newCpGReport جميعها تبدأ وتنتهي بـــCpG على البارامترات المعرفة لهذه الجزر.

جزر الـــCpG:

جزر الـــ CpG هي عبارة عن أجزاء من سلاسل النكليوتيدات تحتوي على تركيز عالٍ من أزواج الـــCpG التي لم تحدث عليها عملية المَثيلة. إذْ يدل حرف الـــp على رابطة فوسفاتية phosphodiester (bond) بين نكليوتيدي السيتوزين (C) والغوانين (G) الموجودين إلى جانب بعضهما بعضاً على السلسلة. أمَّا المثيلة فهي تغيّر كيميائي يتمثل بارتباط زمرة الميثيل CH3 (methyl) بحلقة نكليوتيد السيتوزين [1]. ويبيّن الشكل (1) توضع أزواج الـــCpG في سلسلة النكليوتيدات.

tttaacttaa acctCGgcCG gcCGccCGcc
gcaattggtc ccCGCGcCGa cctcCGccCG
CGtcoctccc cctCGgcccc gCGCGtCGcc
gCGcCGCGag cttctcctct cctcaCGacc
CGCGggCGgg ggcaggggag cCGCGggggc
gCGcaagggt gccCGgcCGg gCGgggtCGg

الشكل (1) – توضع أزواج الـــCpG في سلسلة النكليوتيدات

عُرِّفَتْ جزر الـــCpG من قبل Gardiner-Garden وFrommer في العام 1987على أنها سلاسل بطول لا يقل عن 200 نكليوتيد تحقق محتوى من السيتوزين والغوانين أعلى أو يســـاوي 50% ≤ (C+G content) (50% ونسبة الملاحظ/المتوقع أكبر أو يساوي 0. [1]، إذْ:

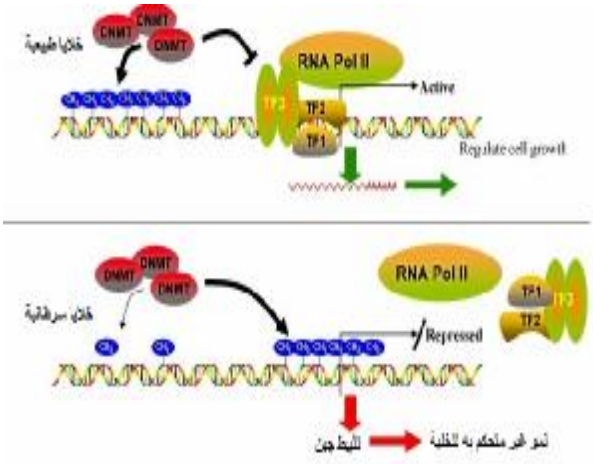
$$\text{محتوى C+G} = \frac{\text{عدد الـ (C) في الجزيرة} + \text{عدد الـ (G) في الجزيرة}}{\text{عدد النكليوتيدات في الجزيرة}}$$

$$\text{O/E} = \frac{\text{عدد الـ (CG) في الجزيرة}}{\frac{\text{عدد الـ (C) في الجزيرة} * \text{عدد الـ (G) في الجزيرة}}{\text{عدد النكليوتيدات في الجزيرة}}}$$

ولا يزال هذا التعريف مستخدماً في أغلب خوارزميات الكشف عن جزر الـــCpG.

تؤدي عملية المثيلة دوراً مهماً في كثير من النشاطات البيولوجية. ومعظم أجسام الجينات وسلاسل الـــDNA بين الجينات تخضع لهذه العملية باستثناء جزر الـــCpG، وهذا ما يجعل لهذه الجزر دوراً محورياً كعلامات جينية وسرطانية في سلاسل الـــDNA. حيث تتراكب جزر الـــCpG مع منطقة محفز عملية النسخ لسلسلة الـــDNA في نحو (60-70)% من جينات الانسان، لذلك فهي تؤدي دوراً مهماً في عملية التحديد والبحث عن موقع محفز النسخ للجينات. كما أن بقاءها دون مثيلة يسهم في تفعيل عملية النسخ للجينات. أسهمت بعض جزر الـــCpG قد في الكشف عن محفزات نسخ جينات لم تكن معروفة سابقاً [1]. في بعض الحالات تحدث مثيلة لجزر الـــCpG الأمر الذي يؤدي إلى تعطيل تفعيل الجين الذي تتراكب الجزيرة مع محفز نسخه. كمثال على ذلك البصمة الوراثية (Gene imprinting) وعملية تعطيل الصبغي X (X inactivation) [1].

إحدى أهم تأثيرات مثيلة جزر الـــCpG ما يحدث في السرطانات كما يوضح الشكل (2). إذْ أثبتت الدراسات أن مثيلة جزر الـــCpG يعدُّ من أوائل التغيرات وأكثرها حدوثاً في السرطانات، إذْ تسبب هذه العملية تعطيل الجينات المثبطة لحدوث التسرطن[1][2][3].



الشكل (2) – عملية المثيلة وتأثيرها في الخلايا السرطانية. يبيّن الشكل أنه في الخلايا الطبيعية تتراكب عوامل النسخ TF (Transcription Factor) المسؤولة عن تفعيل بدء عملية نسخ الجينات مع منطقة المحفز الذي لم يخضع للمثيلة (لا وجود لجزيئات CH3)، وعليه يتفعل الجين. في حين تحدث في الخلايا السرطانية المثيلة في منطقة المحفز ومن ثَمَّ يُثَبَّطُ الجين.

خوارزميات الكشف عن جزر الـــCpG:

1- خوارزميات تعتمد مبدأ النافذة المنزلقة:

تستخدم هذه الخوارزميات نافذة بطول محدد تتحرك على سلسلة الـــنكليوتيدات وتقيس محتوى C+G ونسبة الملاحظ/المتوقع ضمن النافذة لتقييم كونها جزيرة CpG أو لا. وبناءً على نتيجة التقييم تُحَرَّكُ النافذة بخطوة محددة وتكرر العملية حتى نهاية السلسلة المدروسة. وتختلف الخوارزميات التي تعتمد على هذه الطريقة عن بعضها بحجم النافذة، وحجم الخطوة، وتغيير اتجاه تحريك النافذة على السلسلة، والعتبات المستخدمة والمسافة التي تفصل بين جزيرتين متجاروتين.

العديد من هذه الخوارزميات موجودة على الإنترنت كأدوات مفتوحة المصدر (open source) يمكن استخدامها في البحث عن جزر الـــCpG. منها CpGProd، newCpGReport، CpGIS، CpGIE. يسمح بعض هذه الأدوات للمستخدم بتحديد حجم النافذة أو حجم الخطوة أو العتبات أو المسافة بين الجزيرتين المتجاورتين وتُدْخَلُ بعدها السلسلة المطلوب البحث ضمنها عن جزر الـــCpG لتقوم الأداة بإعطاء نتيجة الكشف [5][6][14][15] .

تعاني الخوارزميات المعتمدة على مبدأ النافذة المنزلقة المذكورة سابقاً من بعض المشكلات وهي:

- يعتمد عدد الجزر المكتشفة وطولها على حجم النافذة وحجم الخطوة، فإذا كان حجم النافذة كبيراً فإنه قد يحدث دمج للجزر الصغيرة في جزيرة واحدة.
- الجزر المكتشفة بهذه الخوارزميات لا تبدأ وتنتهي بـــCpG.
- زمن تنفيذ طويل.[8]

2- خوارزميات تعتمد على المسافة:

بسبب ارتفاع نسبة أزواج الـــCpG ضمن جزر الـــCpG مقارنة بباقي سلسلة الـــDNA فإن توزعات المسافة بين هذه الأزواج ضمن الجزر تختلف عنها بباقي سلسلة الـــDNA. لذلك طُرِحَتْ خوارزمية جديدة هي CpGCluster تعتمد المسافة الفيزيائية بين أزواج الـــCpG على السلسلة، ومن ثَمَّ لا تعتمد على العتبات السابقة المحددة للجزر. أظهرت هذه الخوارزمية دقة كشف عالية مقارنة بالأدوات السابقة، واستطاعت

الكشف عن جزر صغيرة لم تكن مكتشفة بالأدوات السابقة [7]. لهذه الخوارزمية مشكلاتها وأهمها أن نتائج البحث تعتمد على تركيبة السلسلة المدروسة، أي إنَّ الجزيرة المكتشفة في سلسلة ما قد يجري تجاهلها في سلسلة أخرى لها تركيبة مختلفة [8].

3- خوارزميات الكشف الذكية:

جرى في السنوات الأخيرة التوجه نحو بناء خوارزميات كشف عن جزر الـــCpG تعتمد على تقنيات الذكاء الصنعي وتعليم الآلة، وذلك في محاولة لتحسين عملية الكشف والتخلص من مشكلات الخوارزميات السابقة، ومن هذه الخوارزميات:

1) خوارزمية تعليم الآلة للكشف عن جزر الـــCpG في سلاسل الـــDNA للإنسان CpG-Discover:

يُنَفَّذُ في هذه الخوارزمية نظام تعليم الآلة للكشف عن جزر الـــCpG يعتمد على نموذج ماركوف المخفي (Hidden Markov model HMM)، يتألف هذا النظام من مرحلتين:

الأولى الكشف اعتماداً على نظام ماركوف المخفي: إذْ بُنِيَ النظام ودُرِّبَ على إحدى قواعد البيانات لجزر الـــCpG وهي قاعدة البيانات الموجودة على موقع المعهد الأوروبي للمعلوماتية الحيوية (European Bioinformatics Institute). والثانية معالجة لاحقة لتحسين الكشف وتصحيح الخطأ: إذْ اِسْتُخْدِمَتْ ثلاثة نماذج معالجة لاحقة لتحسين دقة الكشف وتصحيح الخطأ اعتماداً على الخصائص الجينية للجزر من حيث الطول ومحتوى C+G ونسبة observed/expected.

أظهرت نتائج تطبيق هذا النظام أنه نظام واعد في الكشف عن جزر الـــ CpG بدقة. وهو يختلف عن باقي الأدوات بأنه يتجاوز مشكلات مبدأ النافذة المنزلقة، ويقلل التدخل البشري بالبحث من حيث تحديد العتبات والمسافة بين الجزيرتين المتجاورتين التي تعاني منها باقي الأدوات. دُرِسَ أداء النظام بمقارنة نتائجه بالقيم الحقيقية للكشف، وتبين قدرته على كشف أدق لجزر الـــCpG بحسب مُطوريه [9].

2) تصميم خوارزمية جينية (Genetic Algorithm) تعتمد على النظام العائم (Fuzzy System) لتوقع جزر الـــCpG في جينوم الإنسان:

اِقْتُرِحَتْ في هذا البحث خوارزمية جينية تعتمد على النظام العائم (Fuzzy Genetic Algorithm-CGI) للكشف عن جزر الـــCpG في جينوم الإنسان. إذْ بُنِيَتْ خوارزمية جينية تعتمد التعليم المعزز واِسْتُخْدِمَ النظام العائم لتحسين الخوارزمية. وقد أظهر هذا النظام قدرة كشف للجزر بدقة وحساسية عالية.

استخدم في هذا النظام سلاسل الـــDNA المأخوذة من موقع الـــNCBI الذي يحوي على المعلومات المطلوبة جميعها. بُنِيَ تابع الكفاءة fitness function بالاعتماد على محددات الطول ومحتوى الـــC+G ونسبة الملاحظ/المتوقع لجزر الـــCpG. قُورِنَتْ نتائج النظام بثلاث أدوات معروفة للكشف عن جزر CpG وأظهر تفوقه عليها بحسب مطوريه [10].

4- دراسات مقارنة بين CpGCluster وخوارزمية Takai وJones المعتمدة على مبدأ النافذة المنزلقة:

بيّنت دراسة قام بها الباحثان Takai وJones عام 2002 أن العتبات التقليدية المستخدمة للكشف عن جزر الـــCpG ((200-%50-0.6) للطول والمحتوى ونسبة

الملاحظ/المتوقع على الترتيب) المعتمدة في أغلب خوارزميات الكشف التي تعتمد مبدأ النافذة المنزلقة تؤدي إلى عَدِّ عدد من السلاسل على أنها جزر في حين هي ليست جزراً. لذلك قام الباحثان برفع هذه العتبات إلى (500-55%-0.65) مما أدى إلى انخفاض عدد الجزر المكتشفة بنحو 90%. إذْ قام الباحثان بتطوير خوارزمية تعتمد مبدأ النافذة المنزلقة وتطبيقها على سلاسل النكليوتيدات للصبغيين 21 و22 عند الإنسان عند العتبات (200-50%-0.6)، وكان عدد الجزر المكتشفة 14062 جزيرة، وعند رفع عتبتي المحتوى ونسبة الملاحظ/المتوقع إلى 55% و 0.65 على الترتيب انخفض عدد الجزر المكتشفة بنسبة نحو 49%. وبزيادة عتبة الطول إلى 500 مع الإبقاء على رفع العتبتين السابقتين انخفض عدد الجزر المكتشفة بنسبة نحو 92%. فسَّر الباحثان هذا الانخفاض باستبعاد السلاسل التي صُنِّفَتْ خطأً على أنها جزر CpG، ولكن اِسْتُبْعِدَ في الوقت نفسه عدد من الجزر الصحيحة. وبين الباحثان أن عدد الجزر المكتشفة التي تتراكب مع محفزات نسخ الجينات لم يتأثر كثيراً بزيادة العتبات وإنما التأثير الأكبر كان في عدد الجزر الواقعة خارج هذه المنطقة [1][13].

قامت بعض الدراسات بالمقارنة بين خوارزمية الباحثين Takai وJones [13] المعتمدة على مبدأ النافذة المنزلقة بعتبات (500-55%-0.65) وبَيَّنَ CpGCluster [11][12]. إذْ تبيَّن إحدى الدراستين أن نحو 35% من الجزر التي تكشفها خوارزمية Takai وJones ضمن جينوم الإنسان تتراكب مع محفزات نسخ الجينات، في حين 14.7% فقط من الجزر المكتشفة باستخدام CpGCluster تتراكب مع هذه المنطقة. هذا ما يجعل خوارزمية Takai وJones أفضل في الكشف عن الجزر المتراكبة مع محفزات نسخ الجينات. ولكن تبيّن الدراسة الثانية أن استخدام CpGCluster بصرامة أكبر وذلك بتغيير بعض القيم الإحصائية المستخدمة ضمن الخوارزمية (Strict-set)، فإن 52% من الجزر المكتشفة تتراكب مع منطقة المحفزات. وتبيَّن الدراسة الأولى أيضاً أن توزع الجزر المكتشفة باستخدام CpGCluster التي يتجاوز طولها 500 نكليوتيد ضمن الجينوم لا يختلف كثيراً عن الجزر المكتشفة باستخدام خوارزمية Takai وJones. كما أن نحو 58% من الجزر المكتشفة باستخدام CpGCluster والمتراكبة مع محفزات النسخ للجينات تفصل بينها مسافات أقل من 100 نكليوتيد، علماً أن خوارزمية Takai وJones تدمج الجزر التي تفصل بينها مسافة أقل من 100 ضمن جزيرة واحدة، وهذا يعدُّ أحد الأسباب في أنَّ الجزر المكتشفة باستخدام الخوارزمية الأخيرة أطول من الثانية. كما أن العديد من الجزر المكتشفة باستخدام خوارزمية Takai وJones المعتمدة على مبدأ النافذة بعتبات (500-55%-0.65) تغطي الواحدة منها عدة جزر مكتشفة باستخدام CpGCluster [11][12].

وتبيَّن الدراسات المذكورة أعلاه أيضاً أن 37.8% من الجينات عند الإنسان تحوي في منطقة محفز نسخها على أكثر من جزيرة مكتشفة باستخدام CpGCluster، في حين 3.2% من الجينات تحتوي على أكثر من جزيرة مكتشفة باستخدام خوارزمية Takai وJones في منطقة محفز النسخ. وعلى اعتبار أن بعض محفزات نسخ

الجينات لها أكثر من موقع لبدء عملية النسخ فإن CpGCluster قد يكون أفضل للكشف عن الجزيرة الموجودة في كل موقع بداية نسخ بدلاً من الكشف عن جزيرة واحدة كبيرة تغطي مواقع بدء النسخ جميعها وهو ما تفعله خوارزمية Takai وJones [11][12].

الدراسة العملية والنتائج:

1-تجميع بيانات جزر الـــCpG الموجودة على الصبغي 22 عند الإنسان:

كما ذكرنا سابقاً هَدَفَ هذا البحث إلى دراسة مجموعة خصائص لجزر الـــCpG الناتجة عن استخدام أداتي كشف مختلفتين. اختيرت newCpGreport وCpGCluster المتاحتان للاستخدام من خلال مواقعهما الالكترونية، علماً أن خوارزميتي الكشف الذكيتين [9][10] المذكورتين سابقاً غير متوافرتين كأدوات مفتوحة المصدر. وفيما يأتي لمحة عن الأداتين المختارتين توضح أسباب اختيارهما:

- newCpGReport: تعتمد مبدأ النافذة المنزلقة، وهذه الأداة هي المعتمدة للكشف عن جزر الـــCpG من قبل المختبر البيولوجي الجزيئي الأوروبي - معهد المعلوماتية الحيوية EMBL-EBI (European Molecular Biology Laboratory-European Bioinformatics Institute). الذي يعدُّ أحد أهم المراكز في مجال علم الجينات [4]، ويشمل عدداً كبيراً من قواعد البيانات البيولوجية التي يمكن الوصول إليها عبر الموقع الالكتروني للمعهد، ومن ضمنها قواعد بيانات جزر الـــCpG عند الإنسان المكتشفة باستخدام newCpGReport. وقد بُنِيَتْ هذه القاعدة بتطبيق newCpGReport على سلاسل من النكليوتيدات عند العتبات (200- 50%- 0.6) لكل من الطول، والمحتوى، ونسبة الملاحظ/المتوقع على الترتيب، ويمكن استخدام هذه الأداة من خلال الموقع الالكتروني الآتي:

http://emboss.bioinformatics.nl/cgi-bin/emboss/newcpgreport

- CpGCluster: وهي الأداة الشهيرة التي تعتمد مبدأ المسافة في الكشف عن جزر الـــCpG بدلاً من العتبات التقليدية. وتبدأ الجزر التي تكشفها هذه الأداة وتنتهي بزوج النكليوتيدات CG الأمر الذي يعدُّ أحد مميزات CpGCluster. كما بيّنت الدراسات أن عدد الجزر المكتشفة من قبل هذه الأداة أكبر من عدد الجزر المكتشفة باستخدام الأدوات المعتمدة على مبدأ النافذة المنزلقة، كما أن طول غالبية هذه الجزر أقصر [12]. ويمكن استخدام هذه الأداة من خلال الموقع الالكتروني الآتي الذي يتيح استخدام الأداة بحسب مطوريها:

http://bioinfo2.ugr.es/CpGcluster

اِسْتُخْدِمَتْ في هذا البحث الأداتان السابقتان للبحث عن جزر الـــCpG الموجودة على الصبغي 22 عند الإنسان. إِذْ اِسْتُخْدِمَتْ العتبات (200- 50%- 0.6) نفسها للأداة newCpGReport، وكان عدد الجزر المكتشفة باستخدامها 1648 جزيرة. وعدد الجزر المكتشفة باستخدام CpGCluster 2442 جزيرة. وهذه النتائج توضح بداية كل جزيرة ونهايتها وطولها ومحتوى C+G ونسبة الملاحظ/المتوقع. وتوضّح النتائج أن عدد الجزر التي اكتشفتها أداة CpGCluster أكبر من عدد الجزر التي اكتشفتها أداة newCpGReport. استُخدمت الجزر المكتشفة (1648 و 2442) جميعها ضمن المحاور الثلاثة للدراسة المذكورة عناوينها في فقرة المقدمة. وفيما يأتي محاور الدراسة.

2-دراسة طول جزر الـــCpG المكتشفة وعددها على الصبغي 22 عند الإنسان باستخدام الأداتين المذكورتين، وذلك عند قيم مختلفة لعتبات البارمترات التقليدية المعرفة للجزر:

دُرِسَ في هذا البحث تأثير تغيير العتبات لتصبح (500-200 و55%-0.65) على عدد الجزر المكتشفة باستخدام newCpGReport على الصبغي 22 عند الإنسان. وتبيّن أنه عند البحث عن الجزر التي تحقق العتبات السابقة ضمن الـــجزر المكتشفة بهذه الأداة البالغ عددها 1648، وُجدت 362 جزيرة تحقق ذلك. أي انخفض عدد الجزر بنسبة نحو78%(كما أن استخدام أداة newCpGReport للبحث عن الجزر التي تحقق العتبات الجديدة ضمن الصبغي 22 بشكل مباشر أدى إلى الكشف عن 291 جزيرة فقط أي انخفاض بنسبة نحو 82%). وبمزيد من الدراسة كُرِّرَ العمل السابق باستخدام العتبات (200-500 و55%-0.65)، إذْ تبيّن أنه عند البحث ضمن الـــ1648 جزيرة المكتشفة عن الجزر التي تحقق هذه العتبات كانت النتيجة أن 1419 جزيرة تحقق ذلك، أي انخفض عدد الجزر بنسبة نحو 14% فقط مقارنة مع 78% عندما كانت عتبة الطول 500 (والتطبيق المباشر لأداة newCpGReport أدى إلى انخفاض العدد بنحو36%). تبيّن هذه النتائج ونتائج الدراسات السابقة [11]و[12] التي درست خوارزمية Takai وJones، أن الخوارزميات المعتمدة على مبدأ النافذة المنزلقة جميعها تتشابه من حيث التأثير الكبير لعتبة الطول في عدد الجزر المكتشفة باستخدام هذه الأدوات. الجدول(1) يبيّن نسب التغير بعدد الجزر المكتشفة باستخدام newCpGReportعند تغيير قيم العتبات.

الجدول (1) نسب التغير بعدد الجزر المكتشفة على الصبغي 22 عند الإنسان باستخدام newCpGReport عند تغيير قيم العتبات

عتبة الطول	عتبة المحتوى	عتبة نسبة الملاحظ/المتوقع	نسبة الجزر التي تحقق العتبات من أصل 1648
200	%55	0.65	%86
500	%55	0.65	%22

أمَّا بالنسبة إلى أداة CpGCluster فهي كما ذُكِرَ آنفاً لا تستخدم عتبة الطول التقليدية المميزة لجزر الـــCpG، ومن ثَمَّ فإنه لا يوجد محدد لطول الجزر المكتشفة باستخدامها. هذا الأمر يبدو واضحاً بدراسة طول الجزر المكتشفة باستخدام هذه الأداة ضمن الصبغي 22. إذْ تبيّن أن عدد الجزر التي طولها أكبر أو يساوي 200 نكليوتيد هو 949 جزيرة من أصل 2442. وتبيّن دراسة قيم بارامترات المحتوى ونسبة الملاحظ/المتوقع بغض النظر عن الطول أن 98% من أصل 2442 جزيرة تحقق (50%-0.6) للبارامترين على الترتيب، و96% تحقق (55%-0.65). يبيّن الجدولان(2) و (3) النسب المئوية لعدد الجزر المكتشفة باستخدام CpGCluster التي تحقق قيماً محددةً للبارامترات التقليدية المعرفة لجزر CpG.

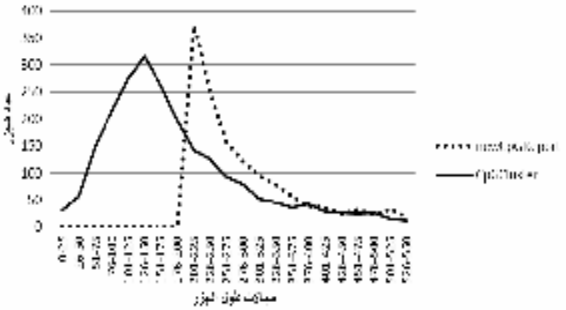
الجدول (2) النسب المئوية لعدد الجزر المكتشفة على الصبغي 22 عند الإنسان باستخدام CpGCluster التي تحقق قيماً محددةً لبارامتري المحتوى ونسبة الملاحظ/المتوقع بغض النظر عن الطول

عتبة المحتوى	عتبة نسبة الملاحظ/المتوقع	نسبة الجزر التي تحقق العتبات من أصل 2442
%50	0.6	%98
%55	0.65	%96

الجدول (3)النسب المئوية لعدد الجزر المكتشفة على الصبغي 22 عند الإنسان باستخدام CpGCluster التي تحقق قيماً محددة لبارامتر الطول بغض النظر عن قيم المحتوى ونسبة الملاحظ/المتوقع.

الطول	نسبة الجزر التي تحقق العتبات من أصل 2442
أكبر أو يساوي 200	%39
أصغر من 200	%61

الشكل (3) يبيّن توزع الجزر المكتشفة باستخدام newCpGReport وCpGCluster بحسب الطول.



الشكل (3) – توزع الجزر المكتشفة باستخدام newCpGReport وCpGCluster بحسب الطول

ونلاحظ من هذين المنحنيين البيانيين أن newCpGReport لا تستطيع الكشف عن الجزر الصغيرة التي يقل طولها عن 200 نكليوتيد التي تدعى CpG-islets. في حين تستطيع CpGCluster الكشف عن هذه الجزر القصيرة التي قد يصل طولها إلى حدود 8 نكليوتيدات وقد تكون جزراً فعالةً [12].

3-دراسة التقاطع بين الجزر المكتشفة باستخدام الأداتين:

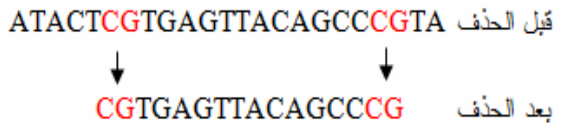
بُنِيَتْ في هذا البحث خوارزمية بسيطة باستخدام برنامج الماتلاب (Matlab) للبحث عن التقاطعات بين الجزر المكتشفة من قبل CpGCluster وnewCpGReport. حيث التقاطع هنا يعني أن تقع بداية جزيرة مكتشفة من قبل إحدى الأداتين أو نهايتها ضمن جزيرة مكتشفة من الأداة الثانية أو أن تكون الجزيرة محتواة بشكل كامل ضمن الجزيرة الثانية. وكانت نتيجة هذه الخوارزمية أن نحو 68% من الجزر المكتشفة باستخدام newCpGReport تتقاطع مع نحو 64% من الجزر المكتشفة من قبل CpGCluster. علماً أنه في بعض الحالات قد تتقاطع أكثر من جزيرة مكتشفة من قبل إحدى الأداتين مع جزيرة أو أكثر مكتشفة من قبل الأداة الثانية. وغالباً ما يحدث أن تتقاطع جزيرة واحدة اِكْتُشِفَتْ من قبل newCpGReport مع أكثر من جزيرة اِكْتُشِفَتْ من قبل CpGCluster. هذه النتيجة تؤكد النتيجة التي ذُكِرَتْ في الفقرة السابقة من حيث أن خوارزميات الكشف المعتمدة على مبدأ النافذة المنزلقة تتشابه من حيث الأداء.

بدراسة قيم المحتوى ونسبة الملاحظ/المتوقع للجزر المتقاطعة يتبيّن أنه في بعض الحالات تكون قيمة المحتوى للجزيرة المكتشفة باستخدام newCpGReport أعلى منها للجزيرة المتقاطعة معها والمكتشفة باستخدام CpGCluster ولكن قيمة نسبة الملاحظ/المتوقع للجزيرة الثانية أكبر منها للأولى، والعكس صحيح. هذه الملاحظة مع ما ذكرناه من تقاطع عدة جزر مكتشفة من قبل إحدى الأداتين مع جزيرة أو أكثر مكتشفة من قبل الأداة الثانية يجعل من الصعب اختيار الفُضْلَى بين هذه الجزر المتقاطعة.

4-دراسة تأثير جعل الجزر المكتشفة باستخدام newCpGReport تبدأ وتنتهي بـــCG:

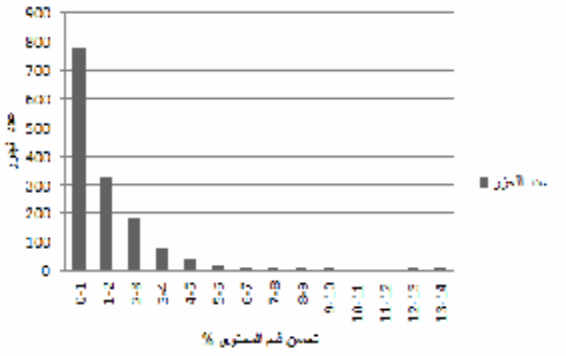
أحد عيوب خوارزميات الكشف التي تعتمد مبدأ النافذة المنزلقة ومن ضمنها newCpGReport أن الجزر المكتشفة لا تبدأ ولا تنتهي بـــCG بسبب تصميم الخوارزمية الذي يعتمد على مبدأ النافذة ذات الطول

المحدد، الأمر الذي تجاوزته خوارزمية CpGCluster. وقد قمنا في هذا البحث ببناء خوارزمية ماتلاب بسيطة تقوم بحذف النكليوتيدات جميعها من بداية كل جزيرة مكتشفة ونهايتها باستخدام newCpGReport حتى الوصول إلى أول زوج CG من الجهتين.



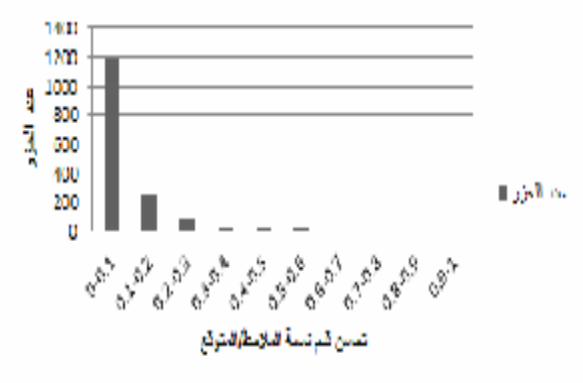
ثم تقوم الخوارزمية بحساب الطول الجديد ومحتوى C+G الجديد ونسبة الملاحظ/المتوقع الجديدة للجزر الجديدة، وتبيّن بدراسة البارامترات الجديدة مايأتي:

- ازدادت قيمة محتوى C+G لـــ1442جزيرة من أصل 1648 جزيرة. الشكل (4) يبيّن توزع التحسين الحاصل على قيمة المحتوى، إذْ إنَّ العدد الأكبر من الجزر التي ازدادت قيمة المحتوى لها كانت الزيادة فيها بنسبة تراوح بين (0-1)%. وأكبر قيمة للزيادة بالمحتوى كانت نحو 13% لجزيرة واحدة فقط.



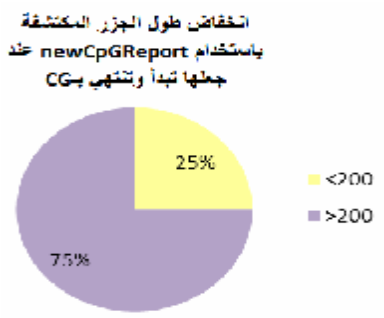
الشكل (4) – توزع التحسين الحاصل على قيمة المحتوى

- ازدادت قيمة نسبة الملاحظ/المتوقع لـــ1528جزيرة من أصل 1648 جزيرة. الشكل (5) يبيّن توزع التحسين الحاصل على قيمة نسبة الملاحظ/المتوقع، إذْ إنَّ العدد الأكبر من الجزر التي ازدادت قيمة الملاحظ/المتوقع لها كانت الزيادة فيها بنسبة تراوح بين (0-0.1). علماً أن أكبر قيمة لنسبة الملاحظ/المتوقع هي 2.16، وأكبر قيمة للزيادة بنسبة الملاحظ/المتوقع كانت نحو 0.55 لجزيرة واحدة فقط.



الشكل (5) – توزع التحسين الحاصل على قيمة نسبة الملاحظ/المتوقع

- نتيجة حذف النكليوتيدات فإن طول الجزر المكتشفة انخفض، وأصبح طول 420 جزيرة أقل من 200 نكليوتيد أو يساويه. الشكل (6) يبيّن نسبة الجزر التي انخفض طولها نتيجة جعل بداية الجزر ونهايتها تبدأ بـــCG.



الشكل (6) – نسبة الجزر التي ينخفض طولها نتيجة جعل بداية الجزر ونهايتها تبدأ بـــCG.

الخاتمة والعمل المستقبلي:

دُرِسَتْ في هذا البحث و حُلِّلَتْ مجموعة من خصائص جزر CpG المكتشفة باستخدام كل من خوارزميتي CpGCluster و newCpGReport للكشف عن هذه الجزر. وجرى بيان تأثير تغيير العتبات الثلاث للطول والمحتوى ونسبة الملاحظ/المتوقع على عدد الجزر المكتشفة. إذْ تبيّن أن تغير العدد بتأثير قيم العتبات المختلفة يكون متشابهاً بين الخوارزميات المعتمدة على مبدأ النافذة المنزلقة وأن بارامتر الطول له التأثير الأكبر. مستندين بهذه الملاحظة إلى نتائج بحثنا على أداة newCpGReport ونتائج الدراسة السابقة على خوارزمية Takai وJones [11] [12]. كما دُرِسَ في هذا البحث تأثير جعل الجزر المكتشفة باستخدام newCpGReport تبدأ وتنتهي بـــCG في خصائص هذه الجزر، الذي أدى إلى زيادة قيم المحتوى ونسبة الملاحظ/المتوقع لنسبة كبيرة من الجزر، ومن ثَمَّ تحسين نتائج الخوارزمية، ولكنه أدى إلى انخفاض طول نحو 25% من الجزر إلى ما دون 200 نكليوتيد.

ومن خلال البحث عن التقاطعات بين الجزر المكتشفة باستخدام الأداتين وجد أنه مع أن 61% من الجزر المكتشفة باستخدام خوارزمية CpGCluster ذات طول أقل من 200 نكليوتيد (كما تبيّن نتائج هذه الدراسة)، وهذا لا يتطابق مع التعريف الذي وضعه Gardiner-Garden وFrommer عام 1987، وهو أن هذه الجزر ذات طول لا يقل عن 200 نكليوتيد، إلا أن 64% من الجزر المكتشفة عن طريق خوارزميةCpGCluster تتقاطع مع الجزر المكتشفة باستخدام newCpGReport، التي تحقق طولاً لا يقل عن 200 نكليوتيد. وهذا يرجع إلى أنه في بعض الحالات قد تتقاطع أكثر من جزيرة مكتشفة من قبل إحدى الأداتين مع جزيرة أو أكثر مكتشفة من قبل الأداة الثانية. وبشكل عام فإن تقاطع الجزر بين الطريقتين هو نحو 60%. وهذا ما يثير نقطة مهمة، وهي وجود اختلاف في نتائج الخوارزميتين بحدود 40%. إن الدراستين السابقتين [11][12] اللتين قارنتا بين كلتا الخوارزميتين قد انحازت كلٌّ منهما خوارزمية معينة (كما ذُكِرَ سابقاً)، إلا أن هذه الدراسة تشير إلى أنه مع أن نسبة التقاطع كانت مرتفعة نسبياً إلا أن قيم بعض بارامترات الجزر المتقاطعة يجعل عملية تحديد الخوارزمية الفُضْلَى بينهما مسألة ذات أهداف متعددة ومتعارضة فيما بينها أحياناً (multi-objective optimization problem). ومن ثَمَّ تبيّن لنا من خلال هذه الدراسة أنه لا يمكن اعتماد أداة واحدة لتكون هي الأداة الفُضْلَى عن جزر الـــCpG بسبب وجود اختلاف واضح في نتائج كلتا الخوارزميتين (على عكس نتائج الدراستين السابقتين اللتين فضلتا خوارزمية على أخرى). وبناءً على ما سبق سنعمل مستقبلاً على بناء نظام كشف يكون قادراً على إعطاء أداء أفضل من الخوارزميات الموجودة حالياً مستفيدين من نتائج هذا البحث الذي مكنّنا من تكوين فهم أوسع لمشكلات الخوارزميات الحالية ومحاسنها.

المصطلحات:

Gene transcription regulation	**تنظيم نسخ سلسلة الـــDNA في الجينات**
Promoters	**محفزات نسخ الجينات**
Methylation	**المثيلة**
Cancerous tumours	**الأورام السرطانية**
Distance based algorithm	**خوارزميات الكشف التي تعتمد على المسافة**
Sliding window algorithm	**خوارزميات الكشف التي تعتمد النافذة المنزلقة**
C+G content	**محتوىC+G**
Observed/Expected ratio	**نسبة الملاحظ/ المتوقع**
Artificial intelligence	**ذكاء صنعي**
Machine learning	**تعليم الآلة**
Gene imprinting	**البصمة الوراثية**
Chromosome X inactivation	**تعطيل الصبغي X**
Transcription Factors	**عوامل النسخ**
Genetic Algorithm	**خوارزمية جينية**
Fuzzy System	**النظام العائم**

المراجع*

[1] Illingworth R.S., Bird A.P., (2009), "CpG islands – 'A rough guide'", FEBS letters, Vol. 583, p.p. 1713-1720.

[2] Hou P., Ji M., Liu Z., Shen J., Cheng L., He N., Lu Z., (2003), "A microarray to analyze methylation patterns of $p^{16Ink4a}$ gene 5'-CpG islands", Clinical Biochemistry, Vol. 36, p.p. 197-202.

[3] Bastian P.J., Yegnasubramanian S., Palapattu G.S., Rogers C.G., Lin X., De Marzoa A.M., Nelson W.G., (2004), "Molecular Biomarker in Prostate Cancer: The Role of CpG Island Hypermethylation", European Urology, Vol. 46, p.p. 698-708.

[4] Claverie J.M., Notredame C., (2007), "Bioinformatics for Dummies 2nd Edition" Wiley Publishing, Inc., USA.

[5] Wang Y., Leung F.C.C., (2004), "An evaluation of new criteria for CpG islands in the human genome as gene markers", Bioinformatics, Vol. 20(7), p.p. 1170-1177.

[6] Ponger L., Mouchiroud D., (2002), "CpGProd: Identifying CpG islands associated with transcription start sited in large genomic mammalian sequences", Bioinformatics, Vol. 18(4), p.p. 631-633.

[7] Hackenberg M., Previti C., Luque-Escamilla P.L., Carpena P., Martínez-Aroza J., Oliver J.L., (2006), "CpGcluster: a distance-based algorithm for CpG-island detection" BMC Bioinformatics, Vol.7, p.p. 446-459.

[8] Ye S., Asaithambi A., Liu Y., (2008), "CpGIF: an algorithm for the identification of CpG islands", Bioinformation, Vol.2(8), p.p. 335-338.

[9] Lan M., Xu Y., Li L., Wang F., Zuo Y., Chen Y., Tan C.L., Su J., (2009), "CpG-Discover: A Machine Learning Approach for CpG Islands Identification from Human DNA Sequence", Proceedings of international joint conference, p.p. 1702-1707.

[10] Chuang L., Chen Y., Yang C., (2009), "Designing of A Novel GA based on Fuzzy System for Prediction of CpG Islands in the Human Genome", Proceedings of FUZZ-IEEE international conference, p.p. 1009-1014.

[11] Han L., and Zhao Z., (2009),"CpG islands or CpG clusters: how to identify functional GC-rich regions in a genome?", BMC Bioinformatics, Vol.10, p.p. 65-71.

[12] Hackenberg M., Barturen G., Carpena P., Luque-Escamilla P., Previti C. and Oliver J., (2010), "Prediction of CpG-island function: CpG clustering vs. sliding-window methods", BMC Genomics, Vol.11, p.p. 327-340.

[13] Takai D., Jones P., (2002), "Comprehensive analysis of CpG islands in human chromosomes 21 and 22", Pubmed, Vol.99, p.p. 2740-2745.

Internet websites:

[14] Lopez R., (1999), "newCpGreport", emboss, http://emboss.bioinformatics.nl/cgi-bin/emboss/newcpgreport, (2011).

[15] Takai D., Jones P.A., (2003), "The CpG Island Searcher: A New WWW Resource", http://www.bioinfo.de/isb/2003030021/, (2011).

* تاريخ ورود البحث إلى مجلة جامعة دمشق2012/5/2